

B V RAJU COLLEGE

**VISHNUPUR
BHIMAVARAM**

STUDENT RESEARCH PUBLICATIONS

ACADAMIC YEAR 2022-23

INDEX

S.NO	REGD.NO	NAME OF THE STUDENT	PAPER TITLE	PAGE NO
1	2185351001	B LAKSHMI NARAYANAMMA	PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS	8-15
2	2185351002	A S S SAI MAHESH	AUTOMATIC DETECTION OF GENETIC DISEASES IN PEDIATRIC AGE USING PUPILLOMETRY	16-22
3	2185351003	A DEEPIKA	A SPAM TRANSFORMER MODEL FOR SMS SPAM DETECTION	23-29
4	2185351004	A KRISHNA PRIYA	DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS	30-36
5	2185351005	A SRAVANI	PLAGIARISM CHECKER USING NLTK	37-41
6	2185351006	GUNDUMOGULA ANUSHA	A STUDY OF BLOCKCHAIN TECHNOLOGY IN FARMER'S PORTAL	42-51
7	2185351007	A BHANUMATHI	DETECTION OF DEPRESSION - RELATED POSTS IN REDDIT SOCIAL MEDIA FORUM	52-58
8	2185351008	B SURESH KUMAR	AN OVERVIEW OF DEEP LEARNING IN MEDICAL IMAGING FOCUSING ON MRI	59-64
9	2185351009	B RAJESH	FISH DISEASE DETECTION USING IMAGE BASED MACHINE LEARNING TECHNIQUE IN AQUACULTURE	65-74
10	2185351010	B DURGA BHAVANI	ADAPTIVE HIERARCHICAL CYBER ATTACK DETECTION AND LOCALIZATION IN ACTIVE DISTRIBUTION SYSTEMS	75-81
11	2185351011	B SANJANA KEERTHI	ANALYSIS OF WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS	82-87
12	2185351012	CH SATYAVATHI HANUMAYAMMA	A STUDENT ATTENDANCE MANAGEMENT METHOD BASED ON CROWDSENSING IN CLASSROOM ENVIRONMENT	88-98
13	2185351013	CH DEVI NIKITHA	FARMING MADE EASY USING MACHINE LEARNING	99-107
14	2185351014	CH TEJVENKAT	CASHLESS SOCIETY MANAGING PRIVACY AND SECURITY IN THE TECHNOLOGICAL AGE	108-113
15	2185351015	CH LATHA NAGESWARI	CROP RECOMMENDATION USING RANDOM FOREST ML ALGORITHM	114-120
16	2185351016	CH VENKATA RAMANA	FLOOD FORECASTING USING MACHINE LEARNING	121-128
17	2185351017	CH N R L AISHWARYA	FACIAL EMOTION RECOGNITION SYSTEM THROUGH MACHINE LEARNING APPROACH	129-134
18	2185351018	CH VIJAYA PRIYA	A MACHINE LEARNING BASED CLASSIFICATION AND PREDICTION TECHNIQUE FOR DDOS ATTACKS	135-141
19	2185351019	CH SWATHI	WATERNET: A NETWORK FOR MONITORING AND ASSESSING WATER QUALITY FOR DRINKING AND IRRIGATION PURPOSES	142-149
20	2185351020	D KUSUMA PRABHA	COMPOSITE BEHAVIORAL MODELING FOR IDENTIFY THEFT DETECTION IN ONLINE SOCIAL NETWORKS	150-155
21	2185351021	D SRINATH	AN ARTIFICIAL INTELLIGENCE AND CLOUD BASED COLLABORATIVE PLATFORM FOR PLANT DISEASE IDENTIFICATION, TRACKING AND FORECASTING FOR FARMERS	156-166
22	2185351022	D L S PAVAN KUMAR	GROUND WATER LEVEL PREDICTION USING HYBRID ARTIFICIAL NEURAL NETWORK WITH GENETIC ALGORITHM	167-173
23	2185351023	D DIANA	FRAUD DETECTION IN ONLINE PRODUCT REVIEW SYSTEMS VIA HETEROGENEOUS GRAPH TRANSFORMER	174-180

24	2185351024	D VISHALI	PREGBOT : A SYSTEM BASED ON ML AND NIP FOR SUPPORTING WOMEN AND FAMILIES DURING PREGNANCY	181-187
25	2185351025	D SIVA SANKAR	HEARSMOKING: SMOKING DETECTION IN DRIVING ENVIRONMENT VIA ACOUSTIC SENSING ON SMARTPHONES	188-193
26	2185351026	G AJAY KUMAR	SLIDING WINDOW BLOCKCHAIN ARCHITECTURE FOR INTERNET OF THINGS	194-200
27	2185351027	G KAVYA	PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS	201-206
28	2185351028	G SOWMYA	SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK	207-210
29	2185351029	G JAYA DATTA SRI	BLOCK HUNTER FEDERATED LEARNING FOR CYBER THREAT HUNTING IN BLOCKCHAIN-BASED IIOT NETWORKS	211-215
30	2185351030	G JAMES MYDHILI	PREDICTING AND DEFINING B2B SALES SUCCESS WITH MACHINE LEARNING	216-225
31	2185351031	G PRADEEP	MACHINE LEARNING AND END-TO-END DEEP LEARNING FOR THE DETECTION OF CHRONIC HEART FAILURE FROM HEART SOUNDS	226-236
32	2185351032	G SAI	SCHOOL-ENTERPRISE COOPERATION ON PYTHON DATA ANALYSIS TEACHING	237-239
33	2185351033	G SRISHA	PRIVACY-PRESERVING SOCIAL MEDIA DATA PUBLISHING FOR PERSONALIZED RANKING-BASED RECOMMENDATION	240-245
34	2185351034	G BHARGAVI	SPAMMER DETECTION AND FAKE USER IDENTIFICATION ON SOCIAL NETWORKS	246-252
35	2185351035	G HARSHA NAGU	PREDICTING URBAN WATER QUALITY WITH UBIQUITOUS DATA - A DATA DRIVEN APPROACH	253-261
36	2185351036	G N S D DIVYA SRI	MACHINE LEARNING FOR FAST AND RELIABLE SOURCE-LOCATION ESTIMATION IN EARTHQUAKE EARLY WARNING	262-270
37	2185351037	G PAVANI SESA LAKSHMI	FADOHS: FRAMEWORK FOR DETECTION AND INTEGRATION OF UNSTRUCTURED DATA OF HATE SPEECH ON FACEBOOK USING SENTIMENT AND EMOTION ANALYSIS	271-279
38	2185351038	G CHARAN KUMAR	A SYSTEMATIC REVIEW OF PREDICTING ELECTIONS BASED ON SOCIAL MEDIA DATA	280-285
39	2185351039	G KARIMUNISA	A DEEP LEARNING APPROACH FOR ROBUST DETECTION OF BOTS IN TWITTER USING TRANSFORMERS	286-294
40	2185351040	HEMALATHA GUNTIPALLI	CAMPUS PLACEMENTS PREDICTION AND ANALYSIS USING MACHINE LEARNING ALGORITHM	295-300
41	2185351041	M HIMA BINDU	HEART DISEASE PREDICTION USING BIO INSPIRED ALGORITHMS	301-307
42	2185351042	I VINEETHA	EARLY DETECTION OF CANCER USING AI	308-318
43	2185351043	J SRIHARI	PERSONALIZED TRAVEL PLANNING SYSTEM	319-325
44	2185351044	J KIRAN KUMAR	FLIGHT DELAY PREDICTION BASED ON AVIATION BIG DATA AND MACHINE LEARNING	326-333
45	2185351045	J LAKSHMI THIRUPATHAMMA	AN EFFICIENT AND PRIVACY-PRESERVING BIOMETRIC IDENTIFICATION SCHEME IN CLOUD COMPUTING	334-340
46	2185351046	J SURESH	DETECTION OF STROKE DISEASE USING MACHINE LEARNING ALGORITHMS	341-349
47	2185351047	K KALYAN	PROTOTYPING MOBILE APP	350-357
48	2185351048	K VINEESHA	SCA SYBIL- BASED COLLUSION ATTACKS OF IIOT DATA POISONING IN FEDERATED LEARNING	358-365
49	2185351049	K NANI VIJATESH	FAKE IMAGE DETECTION USING MACHINE LEARNING	366-372

50	2185351050	K P V S RAVI TEJA	E-PILOTS: A SYSTEM TO PREDICT HARD LANDING DURING THE APPROACH PHASE OF COMMERCIAL FLIGHTS	373-379
51	2185351051	K S S VAMSI	HOTEL REVIEW ANALYSIS FOR THE PREDICTION OF BUSINESS USING DEEP LEARNING APPROACH	380-387
52	2185351052	K CHANDRA VAMSI	QOS RECOMMENDATION IN CLOUD SERVICES	388-394
53	2185351053	K N V SRI RAM	PREDICTION OF USED CAR PRICES USING ARTIFICIAL NEURAL NETWORKS AND MACHINE LEARNING	395-398
54	2185351054	K S S SRAVYA	A HYBRID DEEP LEARNING APPROACH FOR BOTTLENECK DETECTION IN IOT	399-402
55	2185351055	K PREM SAI	CREDIT CARD FRAUD DETECTION USING STATE-OF-THE-ART MACHINE LEARNING AND DEEP LEARNING ALGORITHMS	403-410
56	2185351056	K SRINIVAS VARMA	PREDICTING THE IMPACT OF DISRUPTIONS TO URBAN RAIL TRANSIT SYSTEMS	411-416
57	2185351057	K V V SOMESWAR	CONSTRUCTION SITE ACCIDENT ANALYSIS USING TEXT MINING AND NATURAL LANGUAGE PROCESSING TECHNIQUES	417-423
58	2185351058	K BHARGAVI	CREDIT CARD FRAUD DETECTION USING AUTO ENCODER AND DECODER	424-432
59	2185351059	K M V N S S PAVAN	STRESS DETECTION IN IT PROFESSIONALS BY IMAGE PROCESSING AND MACHINE LEARNING	433-440
60	2185351060	K N V GOWTHAM REDDY	NET SPAM DETECTION FRAMEWORK FOR REVIEWS IN ONLINE SOCIAL MEDIA	441-449
61	2185351061	K JAYASRI	PREDICTING DRUG-DRUG INTERACTIONS BASED ON INTEGRATED SIMILARITY AND SEMI-SUPERVISED LEARNING	450-457
62	2185351062	K VASUDHASRI	CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHM	458-465
63	2185351063	K SAI KIRAN	T-CREO: A TWITTER CREDIBILITY ANALYSIS FRAMEWORK	466-474
64	2185351064	K NAGA RAJU	CLOUDRAID: DETECTING DISTRIBUTED CONCURRENCY BUGS VIA LOG MINING AND ENHANCEMENT	475-480
65	2185351065	K MOUNIKA	INTELLIGENT AGENT BASED JOB SEARCH SYSTEM-DJANGO	481
66	2185351066	K SOUJANYA	GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES	482-486
67	2185351067	K DIVYA MEGHANA	CRYPTOCURRENCY PRICE ANALYSIS WITH ARTIFICIAL INTELLIGENCE	487-497
68	2185351068	K VISWASH	INTERNET FINANCIAL FRAUD DETECTION BASED ON A DISTRIBUTED BIG DATA APPROACH WITH NODE2VEC	498-506
69	2185351069	K VIJAYA RAMYA SRI	DETECTION OF MALICIOUS SOCIAL BOTS USING LEARNING AUTOMATA WITH URL FEATURES IN TWITTER NETWORK	507-512
70	2185351070	K LAKSHMI PRASANNA	MALWARE DETECTION A FRAMEWORK FOR REVERSE ENGINEERED ANDROID APPLICATIONS THROUGH MACHINE LEARNING ALGORITHMS	513-520
71	2185351071	K SAIRAM GUPTA	AN INTELLIGENT DATA-DRIVEN MODEL TO SECURE INTRAVEHICLE COMMUNICATIONS BASED ON MACHINE LEARNING	521-527
72	2185351072	L HARITHA	DRUG RECOMMENDATION SYSTEM BASED ON SENTIMENT ANALYSIS OF DRUG REVIEWS USING MACHINE LEARNING	528-533
73	2185351073	L JYOTHSNA	CHRONIC KIDNEY DISEASE STAGE IDENTIFICATION IN HIV INFECTED PATIENTS USING MACHINE LEARNING	534-539
74	2185351074	M PAVANI	AN EFFICIENT SPAM DETECTION TECHNIQUE FOR IOT DEVICES USING MACHINE LEARNING	540-548

75	2185351075	M VENKATESH	MULTIPLE DISEASE PRIDITION USING MACHINE LEARNING AND STREAMLIT	549-557
76	2185351076	M NARMADA BAI	BULLYNET: UNMASKING CYBERBULLIES ON SOCIAL NETWORKS	558-561
77	2185351077	M LAKSHMIAJITHA	TWITTER SENTIMENT ANALYSIS BASED ON ORDINAL REGRESSION	562-572
78	2185351078	M KARTHIK	ANALYZING AND DETECTING MONEY-LAUNDERING ACCOUNTS IN ONLINE SOCIAL NETWORKS	573-579
79	2185351079	M S V PRASAD	A MACHINE LEARNING MODEL FOR AVERAGE FUEL CONSUMPTION IN HEAVY VEHICLES	580-583
80	2185351080	M SIRISHA	A DEEP LEARNING-BASED APPROACH FOR INAPPROPRIATE CONTENT DETECTION AND CLASSIFICATION OF YOU TUBE VIDEOS	584-592
81	2185351081	M LOKESH VARMA	DEFENSIVE MODELING OF FAKE NEWS THROUGH ONLINE SOCIAL NETWORKS	593-599
82	2185351082	M NANDINI	PREDICTING STOCK MARKET TRENDS USING MACHINE LEARNINGAND DEEP LEARNING ALGORITHMS VIA CONTINOUS AND BINARY DATA; A COMPARATIVE ANALYSIS	600-605
83	2185351083	M SUSHMA	HATECLASSIFY: A SERVICE FRAMEWORK FOR HATE SPEECH IDENTIFICATION ON SOCIAL MEDIA	606-615
84	2185351084	N PHANINDHRA VARMA	STUDENTS PERFORMANCE PREDICTION IN ONLINE COURSES USING MACHINE LEARNING ALGORITHMS	616-625
85	2185351085	N SREE HIMAJA	HOSPITAL MANAGEMENT SYSTEM WITH CHATBOT	626-633
86	2185351086	N DEVIKA	COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING CRIME HOTSPOTS	634-642
87	2185351087	P LAKSHMI PRASANNA	FOUREYE: DEFENSIVE DECEPTION AGAINST ADVANCED PERSISTENT THREATS VIA HYPERGAME THEORY	643-650
88	2185351088	P GAYATHRI	ADAPTIVE DIFFUSION OF SENSITIVE INFORMATION IN ONLINE SOCIAL NETWORKS	651-660
89	2185351089	P.SANTOSH	EMPLOYEE PROMOTION SYSTEM	661-663
90	2185351090	P B V N PADMA	UNSUPERVISED DOMAIN ADOPTION FOR CRIME RISK PREDICTION ACROSS CITIES	664-669
91	2185351091	P JAGADEESH	EMOTION RECOGNITION BY TEXTUAL TWEETS CLASSIFICATION USING VOTING CLASSIFIER LR-SGD	670-673
92	2185351092	P JYOTHI KIRAN	FEATURE EXTRACTION FOR CLASSIFYING STUDENTS BASED ON THEIR ACADEMIC PERFORMANCE	674-680
93	2185351093	P SOWJANYA	SUICIDAL IDEATION DETECTION: A REVIEW OF MACHINE LEARNING METHODS AND APPLICATIONS	681-688
94	2185351094	P GOVARDHAN	LIVER DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES	689-695
95	2185351095	P GAYATHRI	AGRICULTURAL CROP RECOMMENDATIONS BASED ON PRODUCTIVITY AND SEASON	696-699
96	2185351096	P SRUTHI	TRUSTWORTHINESS ASSESSMENT OF USERS IN SOCIAL REVIEWING SYSTEMS	700-708
97	2185351097	P VAMSI KRISHNA	AUTOMATED DETECTING SPAMMERS IN SOCIAL MEDIA	709-718
98	2185351098	P SATYA HANUMA	PHISHING WEBSITE DETECTION USING LIGHT GBM AND SVM ALGORITHM	719-722
99	2185351099	P KEERTHI PRIYA	EMAIL SPAM DETECTION USING MACHINE LEARNING ALGORITHMS	723

100	2185351100	P ALEKYA DEVI	SOCIAL SPAMMER DETECTION VIA CONVEX NONNEGATIVE MATRIX FACTORIZATION	724-733
101	2185351101	P MEGHANA	APPLICATION OF MACHINE LEARNING IN THE FIELD OF MEDICAL CARE	734-740
102	2185351102	P DURGA SAI PRASAD	BUILDING SEARCH ENGINE USING MACHINE LEARNINGTECHNIQUE	741-746
103	2185351103	P V V S RAMA KRISHNA	MOVIE RECOMMENDATION SYSTEM USING SENTIMENT ANALYSIS FROM MICRO BLOGGING DATA	747-757
104	2185351104	R V SOMESWARA RAO	CROP RECOMMENDER SYSTEM USING MACHINE LEARNING APPROACH	758-764
105	2185351105	R HARINI	DEEP LEARNING BASED OBJECT DETECTION AND RECOGNITION FRAMEWORK FOR THE VISUALLY-IMPAIRED	765-774
106	2185351106	R SESHU	FEATURE EXTRACTION AND ANALYSIS OF NATURAL LANGUAGE PROCESSING FOR DEEP LEARNING ENGLISH LANGUAGE	775-787
107	2185351107	R VENKATA ARJUN NAIDU	MACHINE LEARNING FOR WEB VULNERABILITY DETECTION	788-800
108	2185351108	REJU BALAJI	DESIGNING CYBER INSURANCE POLICIES	801-806
109	2185351109	S CHENDANSREE	AGRICULTURAL LAND CLASSIFICATION BASED ON PHENOLOGICAL INFORMATION FROM DENSE SATELLITE IMAGE TIME SERIES	807-813
110	2185351110	S N V S L SUJITHA	CLASSIFYING FAKE NEWS ARTICLES USING NATURAL LANGUAGE PROCESSING TO IDENTIFY IN-ARTICLE ATTRIBUTION AS A SUPERVISED LEARNING ESTIMATOR	814-823
111	2185351111	S NAGA KOSHORE	A ROAD ACCIDENT PREDICTION MODEL USING DATA MINING TECHNIQUES	824-833
112	2185351112	sheik imran	META-HEURISTIC OPTIMIZATION ALGORITHMS BASED FEATURE SELECTION FOR CLINICAL BREAST CANCER DIAGNOSIS	834-848
113	2185351113	S ANNAPURNA	PREDICTION OF AIR POLLUTION USING MACHINE LEARNING ALGORITHM	849-853
114	2185351114	S N C SOUNDARYA VALLI	FINDING PSYCHOLOGICAL INSTABILITY USING MACHINE LEARNING	854-862
115	2185351115	T V V S SATYA PRASAD	IMPLEMENTATION OF FRUITS RECOGNITION CLASSIFIER USING CONVOLUTIONAL NEURAL NETWORK ALGORITHM FOR OBSERVATION OF ACCURACIES FOR VARIOUS HIDDEN LAYERS	863-870
116	2185351116	T RAVI KIRAN	PHISHING WEB SITES FEATURES CLASSIFICATION BASED ON EXTREME LEARNING MACHINE	871-876
117	2185351117	T JYOTHI	PREDICTION OF MODERNIZED LOAN APPROVAL SYSTEM BASED ON MACHINE LEARNING APPROACH	877-882
118	2185351118	V SANTHI	NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES	883-893
119	2185351119	V SUBHAKAR RAO	QR CODE BASED SMART ATTENDANCE SYSTEM	894-900
120	2185351120	V JOTHIKA	PREDICTION OF CARDIOVASCULAR DISEASE USING SUPERVISED MACHINE LEARNING	901-904
121	2185351121	V HEMA PAVAN KUMAR	DETECTION OF FAKE AND CLONE ACCOUNTS IN TWITTER USING CLASSIFICATION AND DISTANCE MEASURE ALGORITHMS	905-912
122	2185351122	V SAI PRASANNA SWARUPA	SOCIAL MEDIA AND MISLEADING INFORMATION IN A DEMOCRACY: A MECHANISM DESIGN APPROACH	913-921
123	2185351123	V N D GAYATHRI	DISEASE PREDICTION USING MACHINE LEARNING	922-924

124	2185351124	V CHARISHMA	AN EXPERIMENTAL STUDY FOR SOFTWARE QUALITY PREDICTION WITH MACHINE LEARNING METHODS	925-931
125	2185351125	V D P PRASAD	DEEP TEXTURE FEATURES FOR ROBUST FACE SPOOFING DETECTION	932-944
126	2185351126	Y N MAHA LAKSHMI	BOTNET: A SYSTEM FOR REAL TIME BOTNET COMMAND AND CONTROL TRAFFIC DETECTION	945-951
127	2185351127	Y KRISHNA MOULI	A RULE BASED APPROACH TO MITIGATE DDOS ATTACK IN IOT ENVIRONMENT	952-956
128	2185351128	Y VIKRAM	DETECT DUI IN CAR DETECTION SYSTEM FOR ALCOHOL AND BACS	957-965
129	2185351129	Y JYOTHSNA VINITHA	PREDECTIVE ANALYSIS OF BIG MART SALES USING MACHINE LEARNING ALGORITHMS	966-970

PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS

Bikkina lakshmi narayanamma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari
District, Andhra Pradesh, India, 534202.

Abstract— Currently, supermarket run-centres, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.

Keywords—Linear Regression, Polynomial Regression, Ridge Regression, Xgboost Regression

1. INTRODUCTION

Everyday competitiveness between various shopping centres as and as huge marts is becoming higher intense,

violent just because of the quick development of global malls also online shopping. Each market seeks to offer personalized and limited time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides methods for predicting or forecasting sales any kind of organization, extremely beneficial to overcome low – priced used for prediction. Always better prediction is helpful, both in developing and improving marketing strategies for the marketplace, which is also particularly helpful

1.1 RELEATED WORK

A great deal of work having been gotten really intended to date the territory of deals foreseeing. A concise audit of the important work in the field of big mart deals is depicted in this part.

Numerous other Measurable methodologies, for example, with regression, (ARIMA) Auto-Regressive Integrated Moving Average, (ARMA) Auto-Regressive Moving Average, have been utilized to develop a few deals forecast standards. Be that as it may, deals anticipating is a refined issue and is influenced by both outer and inside factors, and there are two significant detriments to the measurable technique as set out in A. S. Weigend et al. A mixture occasional quantum relapse approach and (ARIMA) Auto-Regressive Integrated Moving Average way to deal with every day food deals anticipating were recommended by N. S. Arun raj and furthermore found that the exhibition of the individual model was moderately lower than that of the crossover model.

E. Hadavandi utilized the incorporation of "Genetic Fuzzy Systems (FS)" and information gathering to conjecture the deals of the printed circuit board. In their paper, K-means bunching delivered K groups of all information records. At that point, all bunches were taken care of into autonomous with a data set tuning and rule-based extraction ability. Perceived work in the field of deals gauging was done by P.A. Castillo, Sales estimating of new distributed books was done in a publication market the

executives setting utilizing computational techniques. "Artificial neural organizations" are additionally utilized nearby income estimating. Fluffy Neural Networks have been created with the objective of improving prescient effectiveness, and the Radial "Base Function Neural Network (RBFN)" is required to have an incredible potential for anticipating deals.

2. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of

fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence

must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system require Here our project is basically meant for users within the company whose systems can be

interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the

beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

4.3.3 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is

approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the

input design is to make the data entry logical and free from errors. The errors in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

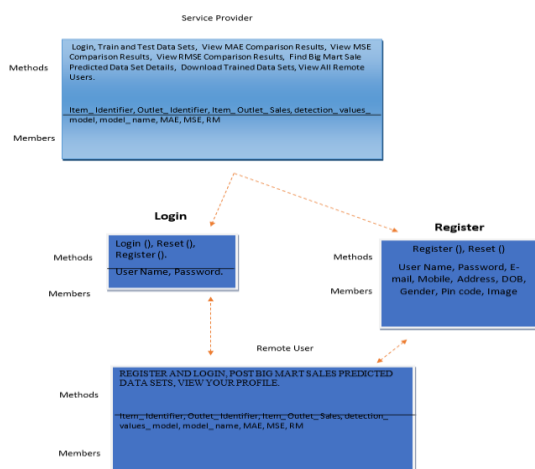
OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the

projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.

The application starts running when it is executed for the first time. The server has to be started and then the internet explorer in used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

➤ **Class Diagram :**



CONCLUSIONS

In this work, the effectiveness of various algorithms on the data on revenue and review of, best performance-algorithm, here propose a software to using regression approach for predicting the sales centered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Ridge regression, and Xgboost regression can be determined. So, we can conclude ridge and Xgboost regression gives the better prediction with respect to Accuracy, MAE and RMSE than the Linear and polynomial regression approaches. In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively. In future work we can also consider with the ARIMA model which shows the time series graph.

REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, “A comparative study of linear and nonlinear models for aggregate retails sales forecasting”, Int. Journal Production Economics, vol. 86, pp. 217- 231, 2003.

[2] Wang, Haoxiang. "Sustainable development and management in

consumer electronics using soft computation." *Journal of Soft Computing Paradigm (JSCP)* 1, no. 01 (2019): 56.- 2. Suma, V., and Shavige Malleshwara Hills.

"Data Mining based Prediction of D

[3] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." *Journal of Soft Computing Paradigm (JSCP)* 2, no. 02 (2020): 101-110

[4] Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", *Proc. of IEEE Conf. on Business Informatics (CBI)*, July 2017. [5]<https://halobi.com/blog/sales-forecasting-five-uses/>. [Accessed: Oct. 3, 2018]

[6] Zone-Ching Lin, Wen-Jang Wu, "Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone", *IEEE Trans. on Semiconductor Manufacturing*, vol. 12, no. 2, pp. 229 – 237, May 1999.

[7] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", *Int. Journal on Mathematical Theory and Modeling*, vol. 2, no. 2, pp. 14 – 23, 2012.

[8] C. Saunders, A. Gammerman and V. Vovk, "Ridge Regression Learning Algorithm in Dual Variables", *Proc. of Int.*

Conf. on Machine Learning, pp. 515 – 521, July 1998. *IEEE TRANSACTIONS ON INFORMATION THEORY*, VOL. 56, NO. 7, JULY 2010 3561.

[9] "Robust Regression and Lasso". Huan Xu, Constantine Caramanis, Member, IEEE, and Shie Mannor, Senior Member, IEEE. 2015 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration." An improved Adaboost algorithm based on uncertain functions". Shu Xinqing School of Automation Wuhan University of Technology. Wuhan, China Wang Pan School of the Automation Wuhan University of Technology Wuhan, China.

[10] Xinqing Shu, Pan Wang, "An Improved Adaboost Algorithm based on Uncertain Functions", *Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration*, Dec. 2015.

[11] A. S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.

[12] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily

food sales forecasting, Int. J. Production

Economics 170 (2015) 321-335P

AUTOMATIC DETECTION AND RECOGNITION FRAMEWORK FOR THE VISUALLY IMPAIRED

Addagarla Shanmukha Sri Sai Mahesh (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract Inherited retinal diseases cause severe visual deficits in children. They are classified in outer and inner retina diseases, and often cause blindness in childhood. The diagnosis for this type of illness is challenging, given the wide range of clinical and genetic causes (with over 200 causative genes). It is routinely based on a complex pattern of clinical tests, including invasive ones, not always appropriate for infants or young children. A different approach is thus needed, that exploits Chromatic Pupillometry, a technique increasingly used to assess outer and inner retina functions. This paper presents a novel Clinical Decision Support System (CDSS), based on Machine Learning using Chromatic Pupillometry in order to support diagnosis of Inherited retinal diseases in pediatric subjects. An approach that combines hardware and software is proposed: a dedicated medical equipment (pupillometer) is used with a purposely designed custom machine learning decision support system. Two distinct Support Vector Machines (SVMs), one for each eye, classify the features extracted from the pupillometric data. The designed CDSS has been used for diagnosis of Retinitis Pigmentosa in pediatric subjects. The results, obtained by combining the two SVMs in an ensemble model, show satisfactory performance of the system, that achieved 0.846 accuracy, 0.937 sensitivity and 0.786 specificity. This is the first study that applies machine learning to pupillometric data in order to diagnose a genetic disease in pediatric age.

INDEX TERMS Artificial intelligence, clinical decision support systems, machine learning, pupillometry, python, rare diseases, retinitis pigmentosa, retinopathy, support vector machine

1. INTRODUCTION

Inherited Retinal Diseases (IRDs) represent a significant cause of severe visual deficits in children [1]. They frequently are cause of blindness in childhood in Established Market Economies (1/3000 individuals). IRDs can be divided into diseases of the outer retina, namely photoreceptor degenerations (e.g., Leber Congenital Amaurosis, Retinitis Pigmentosa, Stargardt disease, Cone Dystrophy, Acromatopsia, Choroideremia, etc.), and diseases of the inner retina, mainly retinal ganglion cell degeneration

(e.g. congenital glaucoma, dominant optic atrophy, Leber hereditary optic neuropathy). Both conditions are characterized by extremely high genetic heterogeneity with over 200 causative genes identified to The associate editor coordinating the review of this manuscript and approving it for publication was Asad Waqar Malik . date, which represent a remarkable obstacle to a rapid and effective diagnosis (<https://sph.uth.edu/retnet/disease.htm>), also considering that the same gene could



cause different and heterogeneous clinical phenotypes.

2. LITERATURE OVERVIEW

A study of the state of the art was developed at the beginning of the activity. The search for previous articles in the literature was done on Scopus, IEEE Xplore and PubMed, using the following keywords: “clinical decision support system”, “eye diseases”, “rare eye diseases”, “CDSS”, “DSS”, “pupillometry”, “retinitis pigmentosa” and “machine learning”. No articles including all the above keywords were found. None of the found articles use both pupillometry and ML techniques. Most of the found articles refer to “clinical decision support system”, “machine learning” and “eye diseases”. The number of studies decreases when it deals with systems for “rare diseases”, “retinitis pigmentosa” and “pupillometry”. Among all the found articles, the seven resumed below were chosen based on regency and variety, so as to have different views of general approaches when ML interfaces with eye diseases. Brancati et al. [22] apply ML supervised techniques for detecting pigment signs on fundus images acquired with a digital retinal camera to study patients affected by RP. Gao et al. [23] apply the ML random forest algorithm on optical coherence tomography (OCT) images to support the diagnosis of choroideremia by detecting intact choriocapillaris. Four more articles apply similar supervised ML algorithms to common eye diseases such as age-related macular degenerations [24], [25] diabetic retinopathy [26] and glaucoma [27]. Gargeya et al. [28] bring a different

approach to support the diagnosis of diabetic retinopathy using deep learning.

3. REQUIREMENT ANALYSIS

The project involved analyzing the design of few applications so as to make the application more users friendly. To do so, it was really important to keep the navigations from one screen to the other well ordered and at the same time reducing the amount of typing the user needs to do. In order to make the application more accessible, the browser version had to be chosen so that it is compatible with most of the Browsers.

REQUIREMENT SPECIFICATION

Functional Requirements

- Graphical User interface with the User.

Software Requirements

For developing the application the following are the Software Requirements:

1. Python
2. Django
3. MySQL
4. MySQLclient
5. WampServer 2.4

Operating Systems supported

1. Windows 7
2. Windows XP
3. Windows 8

Technologies and Languages used to Develop

1. Python

Debugger and Emulator

- Any Browser (Particularly Chrome)



Hardware Requirements

For developing the application the following are the Hardware Requirements:

- Processor: Pentium IV or higher
- RAM: 256 MB
- Space on Hard Disk: minimum 512MB

4. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

5. ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

5.1 TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

5.2 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

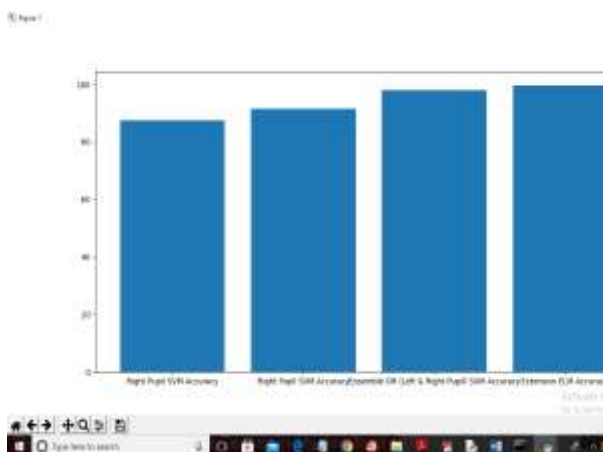
In this project as extension I am adding Extreme Machine Learning Algorithm which is an advance neural network based algorithm and the advantage of this algorithm is filtration of dataset to remove irrelevant columns/attributes and used only important attributes to build machine learning model and due to this filtration an efficient and accurate model will be generated and which can increase prediction accuracy. To run project follow projects screen shots and to run extension concept just you need to run all button and

then click on ‘Run Extension Extreme Learning Machine Algorithm’ button to apply ELM algorithm on dataset and to get below prediction accuracy.

Double click on ‘run.bat’ file to get below screen and then run all button and when u click on ‘Run Extension Extreme Learning Machine Algorithm’ button then you will get below extension results



In above screen on same dataset we got 99.57% accuracy and other algorithms got accuracy less than extension accuracy and now click on ‘Ac



Graph with Metrics’ button to get below screen

In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that Extension ELM accuracy is more than other algorithms

6 INPUT AND OUTPUT DESIGN INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of



the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each

output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.

7. CONCLUSION

This paper describes a new approach for supporting clinical decision for diagnosis of retinitis pigmentosa starting from analysis of pupil response to chromatic light stimuli in pediatric patients. The system was developed to clean artefacts, extract features and help the diagnosis of RP using a ML approach based on an ensemble model of two fine-tuned SVMs. Performances were evaluated with a leave-one-out cross-validation, also used to identify the best combination of internal parameters of the SVM, separately for both the left and right eyes. The class assigned to each eye were combined in the end with an OR-like approach so as to maximize the overall sensitivity of the



CDSS; the ensemble system achieved 84.6% accuracy, 93.7% sensitivity and 78.6% specificity. The small amount of data available for this work, calls for further tests with a larger data pool for validating the performance of the system. Future scope includes testing the same approach with different devices. A problem that came out with great evidence, at the signal acquisition stage, is the frequent presence of movement artifacts. This is due to the particular shape of the device, together with the young age of the enrolled patients. Devices with different frame, including also systems based on smartphones, are going to be investigated. Moreover, considering the duration of the whole acquisition protocol, the procedure would benefit of some systems to capture the attention of the young patient

REFERENCES

- [1] X.-F. Huang, F. Huang, K.-C. Wu, J. Wu, J. Chen, C.-P. Pang, F. Lu, J. Qu, and Z.-B. Jin, "Genotype-phenotype correlation and mutation spectrum in a large cohort of patients with inherited retinal dystrophy revealed by next-generation sequencing," *Genet. Med.*, vol. 17, no. 4, pp. 271–278, Apr. 2015.
- [2] R. Kardon, S. C. Anderson, T. G. Damarjian, E. M. Grace, E. Stone, and A. Kawasaki, "Chromatic pupil responses. Preferential activation of the melanopsin-mediated versus outer photoreceptor-mediated pupil light reflex," *Ophthalmology*, vol. 116, no. 8, pp. 1564–1573, 2009.
- [3] J. C. Park, A. L. Moura, A. S. Raza, D. W. Rhee, R. H. Kardon, and D. C. Hood, "Toward a clinical protocol for assessing rod, cone, and melanopsin contributions to the human pupil response," *Invest. Ophthalmol. Vis. Sci.*, vol. 52, no. 9, pp. 6624–6635, Aug. 2011.
- [4] A. Kawasaki, S. V. Crippa, R. Kardon, L. Leon, and C. Hamel, "Characterization of pupil responses to blue and red light stimuli in autosomal dominant retinitis pigmentosa due to NR2E3 mutation," *Investigative Ophthalmol. Vis. Sci.*, vol. 53, no. 9, pp. 5562–5569, 2012.
- [5] A. Kawasaki, F. L. Munier, L. Leon, and R. H. Kardon, "Pupillometric quantification of residual rod and cone activity in Leber congenital amaurosis," *Arch. Ophthalmol.*, vol. 130, no. 6, pp. 798–800, Jun. 2012.
- [6] A. Kawasaki, S. Collomb, L. Léon, and M. Münch, "Pupil responses derived from outer and inner retinal photoreception are normal in patients with hereditary optic neuropathy," *Exp. Eye Res.*, vol. 120, pp. 161–166, Mar. 2014.
- [7] P. Melillo, A. de Benedictis, E. Villani, M. C. Ferraro, E. Iadanza, M. Gherardelli, F. Testa, S. Banfi, P. Nucci, and F. Simonelli, "Toward a novel medical device based on chromatic pupillometry for screening and monitoring of inherited ocular disease: A pilot study," in *Proc. IFMBE*, vol. 68, 2019, pp. 387–390.
- [8] E. Iadanza, R. Fabbri, A. Luschi, F. Gavazzi, P. Melillo, F. Simonelli, and M. Gherardelli, "ORÁO: RESTful cloud-based ophthalmologic medical record for chromatic pupillometry," in *Proc. IFMBE*, vol. 73, 2020, pp. 713–720.
- [9] E. Iadanza, R. Fabbri, A. Luschi, P. Melillo, and F. Simonelli, "A



collaborative RESTful cloud-based tool for management of chromatic pupillometry in a clinical trial,” *Health Technol.*, pp. 1–14, Aug. 2019, doi: 10.1007/s12553-019-00362-z.

[10] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, “Supervised machine learning: A review of classification techniques,” *Emerg. Artif. Intell. Appl. Comput. Eng.*, vol. 160, pp. 3–24, Jun. 2007.

[11] J. A. Alzubi, “Optimal classifier ensemble design based on cooperative game theory,” *Res. J. Appl. Sci., Eng. Technol.*, vol. 11, no. 12, pp. 1336–1343, Jan. 2016.

[12] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: An overview,” *J. Phys., Conf. Ser.*, vol. 1142, Nov. 2018, Art. no. 012012.

A SPAM TRANSFORMER MODEL FOR SMS SPAM DETECTION

Akula Deepika (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- Currently, supermarket run-centres, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.

Keywords—Linear Regression, Polynomial Regression, Ridge Regression, Xgboost Regression

1. INTRODUCTION

THE Short Message Service (SMS) has been widely used as a communication tool over the past few decades as the popularity of mobile phone and mobile network grows. However, SMS users are also suffering from SMS spam. The SMS spam, also known as drunk message, refers to any irrelevant messages delivered using mobile networks [1]. There are several reasons that lead to the popularity of spam messages. Firstly, there is a large number of users who use mobile phones in the world, making the potential victims of the spam messages attack also high. Secondly, the cost of sending out spam messages is low, which could be good news to the spam attacker. Last but not least, the capability of the spam classifier on most mobile phones is relatively weak due to the shortage of computational resources, which limits them from identifying the spam message correctly and efficiently.

Machine learning is one of the most popular topics in the last few decades, and there are a great number of machine learning based classification applications in multiple research areas. Specifically, spam detection is a relatively mature research topic with several

established methods. However, most of the machine learning based classifiers were dependent on the handcrafted features extracted from the training data [2].

As a class of machine learning techniques, deep learning has been developing rapidly recently thanks to the surprising

growth of computational resources in the last few decades. Nowadays, deep learning based applications play a significant part in our society, making our lives much easier in many aspects. As one of the most effective and widely used deep learning architectures, Recurrent Neural Network (RNN), as well as its variants such as Long Short-Term Memory (LSTM), were applied to spam detection and proved to be extremely effective during the last few years.

The Transformer [3] is an attention-based sequence-to sequence model that was originally designated for translation task, and it achieved great success in English-German and English-French translation. Moreover, there are multiple improved Transformer-based models such as GPT-3 [4] and BERT [5] proposed recently to address different Natural Language Process (NLP) problems. The accomplishments of the Transformer and its successors have proved how powerful and promising they are. In this paper, we aim to explore whether it is possible to adapt the Transformer model to the SMS spam detection problem. Therefore, we propose a modified model based on the vanilla Transformer to identify SMS spam messages. Additionally, we analyze and compare the performance of SMS spam detection between traditional machine learning classifiers, an LSTM deep learning solution, and our proposed spam Transformer model.

2. EXISTING SYSTEM

In [12], Gupta *et al.* compared the performance of 8 different classifiers including SVM, NB, DT, LR, RF, AdaBoost, Neural Network, and CNN. The experimental tests on the SMS Spam Collection v.1 [13] dataset that was conducted by the authors shows that the CNN and Neural Network are better compared to other machine learning classifiers, and the CNN and Neural Network achieved an accuracy of 98.25% and 98.00%, respectively.

In [14], Jain *et al.* proposed a method to apply rule-based models on the SMS spam detection problem. The authors extracted 9 rules and implemented Decision Tree (DT), RIPPER [15], and PRISM [16] to identify the spam messages. According to the experimental results from

the authors, the RIPPER outperformed the PRISM and the DT, yielding a 99.01% True Negative Rate (TNR) and a 92.82% True

Positive Rate (TPR).

In [1], Roy *et al.* aimed to adapt the CNN and LSTM to the SMS spam messages detection problem. The authors evaluated the performance of CNN and LSTM by comparing them with Naïve Bayes (NB), Random Forest (RF), Gradient Boosting (GB) [17], Logistic Regression (LR), and Stochastic Gradient Descent (SGD) [18]. The experiments that were conducted by the authors showed that the CNN and

LSTM perform significantly better than the tested traditional machine learning approaches when it comes to SMS spam detection.

In [2], the authors proposed the Semantic Long Short-Term Memory (SLSTM), a variant of LSTM with an additional semantic layer. The authors employed the Word2vec [19], the WordNet [20], and the ConceptNet [21] as the semantic layer, and combined the semantic layer with the LSTM to train an SMS spam detection model. The experimental evaluation that was conducted by the authors claimed that

the SLSTM achieved an accuracy of 99% on the SMS Spam Collection v.1 dataset.

In [22], Ghourabi *et al.* proposed the CNN-LSTM model that consists of a CNN layer and an LSTM layer in order to identify SMS spam messages in English and Arabic. The authors evaluated the CNN-LSTM by comparing it with the CNN, LSTM, and 9 traditional machine learning solutions. The experimental tests that were conducted by the authors showed that the CNN-LSTM solution performed better than other approaches and yield an accuracy of 98.3% and an F1-Score of 0.914.

Disadvantages

- 1) .The system doesn't have Transformer-based models GPT-3 and BERT to measure an exact spam details.
- 2). There is no technique called SEQUENCE-TO-SEQUENCE MODELS aiming to find a mapping between two sequences for translation tasks.

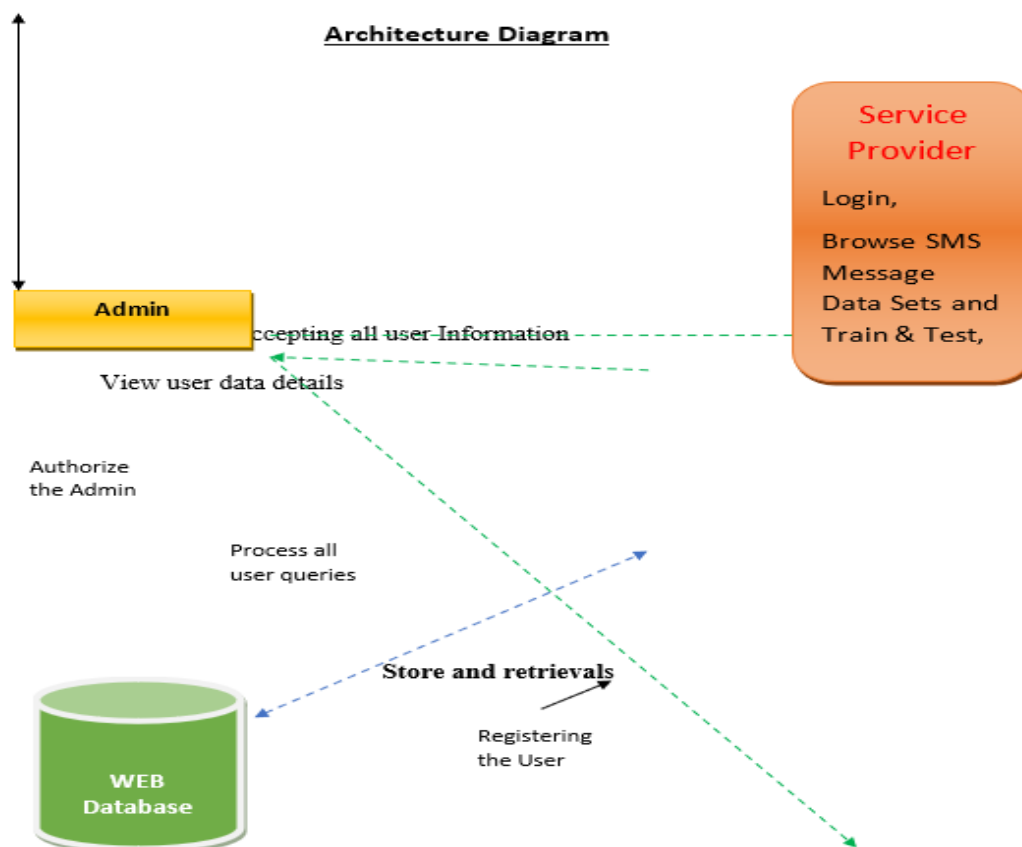
3.PROPOSED SYSTEM

The Transformer [3] is an attention-based sequence-to sequence model that was originally designated for translation task, and it achieved great success in English-German and English-French translation. Moreover, there are multiple improved Transformer-based models such as GPT-3 [4] and BERT [5] proposed recently to address different Natural Language Process (NLP) problems. The accomplishments of the Transformer and its successors have proved how powerful and promising they are. In this paper, we aim to explore whether it is possible to adapt the Transformer model to the SMS spam detection problem. Therefore, we propose a modified model based on the vanilla Transformer to identify SMS spam messages. Additionally, we analyze and compare the performance of SMS spam detection between traditional machine learning classifiers, an LSTM deep learning solution, and our proposed spam Transformer model.

Advantages

The system is more effective due to Long Short-Term Memory (LSTM).

The gives accurate results due to presence of HYPER-PARAMETERS TUNING.



4. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

Request Clarification

Feasibility Study

Request Approval

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires. Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched

5. CONCLUSION

In this paper, we proposed a modified Transformer model that aims to identify SMS spam. We evaluated our spam Transformer model by comparing it with several other SMS spam detection approaches on the SMS Spam Collection v.1 dataset and UtkMI's Twitter dataset. The experimental results show that, compared to Logistic Regression, Naïve Bayes, Random

Forests, Support Vector Machine, Long Short-Term Memory, and CNN-LSTM [22], our proposed spam Transformer model performs better on both datasets.

On the SMS Spam Collection v.1 dataset, our spam Transformer has a better performance in terms of accuracy, recall, and F1-Score compared to other classifiers. Specially, our modified spam Transformer approach accomplished an exceeding result on F1-Score.

Additionally, on the UtkMI's Twitter dataset, the results from our modified spam Transformer model demonstrate its improved performance on all four aspects in comparison to other alternative approaches mentioned in this paper. Concretely, our spam Transformer does exceptionally well on recall, which contributes to a distinct F1-Score.

6. REFERENCES

- [1] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS spam," *Future Gener. Comput. Syst.*, vol. 102, pp. 524_533, Jan. 2020.
- [2] G. Jain, M. Sharma, and B. Agarwal, "Optimizing semantic LSTM for spam detection," *Int. J. Inf. Technol.*, vol. 11, no. 2, pp. 239_250, Jun. 2019.
- [3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999_6009.
- [4] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, *arXiv:2005.14165*. [Online]. Available: <http://arxiv.org/abs/2005.14165>
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171_4186.
- [6] G. Sonowal and K. S. Kuppusamy, "SmiDCA: An anti-Smishing model with machine learning approach," *Comput. J.*, vol. 61, no. 8, pp. 1143_1157, Aug. 2018.
- [7] J. W. Joo, S. Y. Moon, S. Singh, and J. H. Park, "S-detector: An enhanced security model for detecting Smishing attack for mobile computing," *Telecommun. Syst.*, vol. 66, no. 1, pp. 29_38, Sep. 2017.

- [8] S. Mishra and D. Soni, "Smishing detector: A security model to detect Smishing through SMS content analysis and URL behavior analysis," *Future Gener. Comput. Syst.*, vol. 108, pp. 803_815, Jul. 2020.
- [9] C. Li, L. Hou, B. Y. Sharma, H. Li, C. Chen, Y. Li, X. Zhao, H. Huang, Z. Cai, and H. Chen, "Developing a new intelligent system for the diagnosis of tuberculous pleural effusion," *Comput. Methods Programs Biomed.*, vol. 153, pp. 211_225, Jan. 2018.
- [10] T. K. Ho, "Random decision forests," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, 1995, pp. 278_282.

DIABETES DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

Allam Krishna Priya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West
Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract—This paper deals with the prediction of Diabetes Disease by performing an analysis of five supervised machine learning algorithms, i.e. K-Nearest Neighbours, Naïve Baye, Decision Tree Classifier, Random Forest and Support Vector Machine. Further, by incorporating all the present risk factors of the dataset, we have observed a stable accuracy after classifying and performing cross-validation. We managed to achieve a stable and highest accuracy of 76% with KNN classifier and remaining all other classifiers also give a stable accuracy of above 70%. We analyzed why specific Machine Learning classifiers do not yield stable and good accuracy by visualizing the training and testing accuracy and examining model overfitting and model underfitting. The main goal of this paper is to find the most optimal results in terms of accuracy and computational time for Diabetes disease prediction.

1. INTRODUCTION

In this day and age, one of the most notorious diseases to have taken the world by storm is Diabetes, which is a disease which causes an increase in blood glucose levels as a result of the absence or low levels of insulin. Due to the many criterion to be taken into consideration for an individual to harbour this disease, it's detection and prediction might be tedious or sometimes inconclusive. Nevertheless, it isn't impossible to detect it, even at an early stage. Federation-IDF). 79% of the adult population were living in the countries with the low and middle-income groups. It is estimated that by the year 2045 approx. 700 million people will have diabetes (IDF).

Diabetes is increasing day by day in the world because of environmental, genetic factors. The numbers are rising rapidly due to several factors which includes unhealthy foods, physical inactivity and many more. Diabetes is a hormonal disorder in which the inability of the body to produce insulin causes the metabolism of sugar in the body to be abnormal, thereby, raising the

blood glucose levels in the body of a particular individual. Intense hunger, thirst and frequent urination are some of the observable characteristics. Certain risk factors such as age, BMI, Glucose Levels, Blood Pressure, etc., play an important role to the contribution of the disease.

we can see that the number of cases is rising every year and there is not slowing down in the active cases. It is a very crucial thing to worry as diabetes has become one of the most dangerous and fastest diseases to take the lives of many individuals around the globe.

Machine Learning is very popular these days as it is used everywhere, where a large amount of data is present, and we need some knowledge from it. Generally, we can categorise the Machine Learning algorithms in two types but not limited to-

- Unsupervised Learning: In unsupervised learning, the information is not labelled and also not trained. Here, we just put the data in action to find some patterns if possible.

- Supervised Learning: In supervised learning, we train the model based on the labels attached to the information and based on that we classify or test the new data with labels.

With the rise of Machine Learning and its relative algorithms, it has come to light that the significant problems and hindrances in its detection faced earlier, can now be eased with much simplicity, yet, giving a detailed and accurate outcome. As of the modern-day, it is comprehended that Machine Learning has become even more effective and helpful in collaboration with the domain of Medicine. Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual. Such early tries can lead to the inhibition of disease as well as obstruction of permitting the disease to reach a critical degree. The work which will be described in this paper is to perform the diabetes disease prediction using machine learning algorithms for early care of an **individual**.

2. EXISTING SYSTEM

In [2], they have used the WEKA tool for data analytics for diabetes disease prediction on Big Data of healthcare. They used the publicly available dataset from UCI and applied different machine learning classifiers on it. The classifiers which they incorporated are Naive Bayes, Support Vector Machine, Random Forest and Simple CART.

Their approach starts with accessing the dataset, preprocess it in Weka tool and then did the 70:30 train and test split for applying different machine algorithms. They did not go with the cross-validation step as it is imperative to get the optimal and accurate results as well.

The authors in [3], also used the publicly available dataset named as Pima Indians Diabetes Database for performing their experiment. Their framework of performing the prediction starts with the dataset selection and then with data pre-processing. Once the data was preprocessed, they applied three classification algorithms, i.e. naive Bayes, SVM and Decision tree. As they incorporated different evaluation metrics, they did compare the different performance measure and comparatively analyzed the accuracy. The highest accuracy achieved with their experiment was 76.30%. Like [2] they have also not practised Cross-validation.

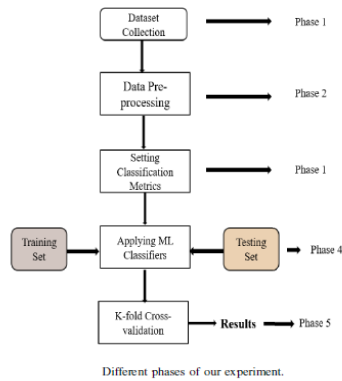
In [4], the authors proposed the neural network-based diabetes disease prediction on Indians Pima Diabetes Dataset. They have used several hidden layers to find patterns in the data, and with the help of those patterns, they predicted the outcome. They name their proposed algorithms as ADAP, which is a custom neural network with multiple partitions and with the set of association weights and units. They managed to achieve a crossover point for sensitivity, and specificity at 0.76 and are trying to precise their result in future.

Disadvantages

- 1). There are no techniques and models for analyzing large scale datasets in the existing system.
- 2). There is no facility for diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results.

3. PROPOSED SYSTEM

To perform our experiment, we have used a publicly available dataset named as Pima Indians Diabetes Database This dataset includes a various diagnostic measure of diabetes disease. The dataset was originally from the National Institute of Diabetes and Digestive and Kidney Diseases. All the recorded instances are of the patients whose age are above 21 years old. Our proposed model exists of 5 phases which are shown in the proposed system by following Figure.



Advantages

- The system more effective due to fitting datasets for different ML Models by Applying Machine Learning Algorithms.
- The Early determination of a disease can be made possible through machine learning by studying the characteristics of an individual in the proposed system.

4. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires. Here our project is basically meant for users within the company whose systems can be

interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility: Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic

Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

4.3 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

5. CONCLUSIONS

One of the significant impediments with the progression of technology and medicine is the early detection of a disease, which is in this case, diabetes. However, in this study, systematic efforts were made into designing a model which is accurate enough in determining the onset of the disease. With the experiments conducted on the Pima Indians Diabetes Database, we have readily predicted this disease. Moreover, the results achieved proved the adequacy of the system, with an accuracy of 76% using the K-Nearest Neighbours classifiers. With this being said, it is hopeful that we can implement this model into a system to predict other deadly diseases as well. There can be room for further improvement for the automation of the analysis of diabetes or any other disease in the future. In future, we will try to create a diabetes dataset in collaboration with a hospital or a medical institute and will try to achieve better results. We will be incorporating more Machine Learning and Deep learning models for achieving better results as well.

6. REFERENCES

[1] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Research and Clinical Practice*, vol. 157, p. 107843, 2019.

- [2] A. Mir and S. N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, pp. 1–6.
- [3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Computer Science*, vol. 132, pp. 1578 – 1585, 2018, international Conference on Computational Intelligence and Data Science. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1877050918308548>
- [4] J. Smith, J. Everhart, W. Dickson, W. Knowler, and R. Johannes, "Using the adap learning algorithm to forcast the onset of diabetes mellitus," *Proceedings - Annual Symposium on Computer Applications in Medical Care*, vol. 10, 11 1988.
- [5] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, 2018, pp. 1 4.
- [6] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, St´efan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61.
- [7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R’io, M. Wiebe, P. Peterson, P. G’erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: <https://doi.org/10.1038/s41586-020-2649-2>
- [8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and ´ Edouard Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, p. 28252830, 2011.



PLAGIARISM CHECKER USING - NLTK

Allu Sravani (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Plagiarism is when someone takes another author's works, thoughts, ideas, etc. without proper referencing and claim it as his/her own works. Plagiarism detection is the process to find the plagiarism within a work or documents. With the advance of modern technology, it makes it easier for people to search for information and plagiarize the work of others. Although the effort and ideas for an image-based plagiarism detection has been increasing over the years, flaws are still present in the current systems. This paper proposes a new system that can cover those flaws. It consists of three stages: the pre-processing, feature extraction and comparison stage. The results showed in an ascending order of similarity index and true and false. However, the accuracy is 100% in case of unedited images and varied in other operations such as flipped, rotated, greyscales and cropped

Keywords :Image Plagiarism, Image Retrieval, Feature extraction.

1. INTRODUCTION

Plagiarism basically means the wrongful stealing of an author's work, thoughts, ideas, etc. and claiming it as your own original work. Plagiarism is considered as deceit and a breach of ethics. In academics, students that are caught with plagiarism are exposed to various levels of penalties and punishment and may even lead to expulsion. Plagiarism in itself cannot be considered as a crime but as copyright violation. In the academics and other industries that are sensitive to copyright infringement, plagiarism is grave misconduct in integrity. The law cannot and usually will not punish plagiarism, but it is up to the institution on how to handle it once it happens [1]. Plagiarism detection is usually split into two which are text-based plagiarism detection and image-based plagiarism detection. For text-based plagiarism detection there are currently five techniques that are used most

often in different fields. These techniques are Fingerprinting, String Matching, Bag of Words, Citation Analysis and Stylometry. String Matching is mostly used in computer science where it compares the documents word for word. Bag of words represents the documents in one or two vectors for comparison. Citation analysis is mainly used in scientific texts because it only compares the citation and reference of the documents. Stylometry checks the author's unique writing style for detection of author's ownership [2]. For image-based plagiarism detection, there are no commonly used techniques like the text-based plagiarism detection, but they usually share the same processes and steps. When we say plagiarism checking or detection we usually mean checking only the text in the file or document for plagiarism. Most of the times when you check your documents or files for plagiarism through a plagiarism



checker software they will check for images and then discard them. This is one of the fatal flaws that the current system is facing. In the field of research, images and flowchart can carry vital piece of information that can easily be plagiarized if the flaw in the system is there.

2. PLAGIARISM

From the Oxford Dictionary, plagiarism means the act of taking someone else's work, ideas, thought, etc. and claiming it as your own work. The word plagiarism comes from the Latin word Plagiarius which means kidnapper, plunderer or seducer. The word Plagium which means

kidnapper is derived from the word Plaga which means to capture or trap. Modern days the word plagiarism means to plagiarize. The process of checking a work or documents for plagiarism is called Plagiarism Detection or Checker [3]. The history of plagiarism first began from religious

texts where most of them were authorless, so it is copied extensively and merged into later works. At the mid-1600, it is very common for there to be accusation of plagiarism for every creative field [3]. In the year 1709, the first copyright law was passed but it has more to do with protecting the publisher's right than the author's, but another law was passed soon after that to protect author's right. James Boswell, who is also known as the biographer for Samuel Johnson, was a lawyer that opposed how long the copyright of the author lasted which at that time ended up to 21 years [4]. In the beginning of the 19th century, the laws for copyright is pretty

like what we have today. The only difference is the issue of enforcing those laws across the borders. Most European country sign an agreement to prevent book piracy except for America which signs it at the year 1891 [4].

3. PLAGIARISM DETECTION TECHNIQUE

A. Text-Based Plagiarism Detection Technique

The main technique used for text-based plagiarism detection is, Fingerprinting, String Matching, Bag of Words, Citation Analysis and Stylometry. The most used technique is the Fingerprinting technique where the system will select a set of multiple substrings from the documents and the sets signifies the fingerprints which is made up of the elements called minutae. The plagiarism checking is done by taking the fingerprints of the documents and comparing them to a String matching is mainly used in the computer science field where the system will compare word for word on each document. This system detect plagiarism in a pair with the original documents and with the collection of

references. Although plentiful methods have been proposed but this system is still computationally expensive making it unfit for large number of documents. For Bag of Word

technique, it used a vector space retrieval representation where the documents are represented with one or two vectors to be used for similarity comparison. The system can use the regular cosine similarity measure or further advance similarity comparison technique. Citation analysis more widely



used for checking plagiarism in scientific text because it is the only technique that does not use textual analysis but examines the citation and references of the documents to recognize comparable pattern. This technique is still considerably new, so it is not ready for commercial use yet [2]. The last technique, Stylometry detects plagiarism by checking the writer's unique writing style so it is more widely used for checking the original owner attribute. The system compares the stylometric models for different text segments that are stylistically different from others.

B. Image-Based Plagiarism Detection Technique

Image-based plagiarism detection is less used compared to text-based plagiarism detection, so it does not have a widely popular technique that is used everywhere. Instead, they are still researching for a good method to detect plagiarism in images. Here we will discuss a few of the methods that were proposed every year. Method by Popescu and Farid [5] proposed a system by using the Principal Part Analysis (PCA). This system divides the numerous tiny sized blocks into vectors which are then organized lexicographically before matching them. The main drawback for this system is that if the image quality is too low then the accuracy will fall as well. Mahdian and Saic [6] apply a blur movement variant to signify the image region so that the images will not be degraded from blurring and noises. This system begins by tilting the image with selected size blocks and defining it with blur invariants. The drawback from this method is that the computation time for this system

is very long. An experiment conducted by Wang et al [7] used a copy-move plagiarism detection system by applying the victimization Hu moments to cut down the computation time of the system. This system divides the image into numerous sized blocks and then applying the Hu moments on the block they computed the Eigen value. A more enhanced method has been proposed by Zimba and Xingming [8] for the copy-move detection system. The system starts by converting the image into a gray scale image and then applying DWT. The image is then divided into overlapping blocks and then PCA is done to each block. This method cuts down the computation time of the system compared to the PCA method by reducing the size of the image. Bravo-Solorio and Nandi [9] proposed a system to detect reflection, scaling and rotation of an images. The drawback is that their methods produced a lot of matches which needed to be further improved. A system that had been proposed by Sridevi et al [10] used the copy-move detection system in parallel which makes the system unable to use methods that have a requirement of long computation time. The only disadvantage with this system is that it cannot detect colored images.

4. RELATED WORK

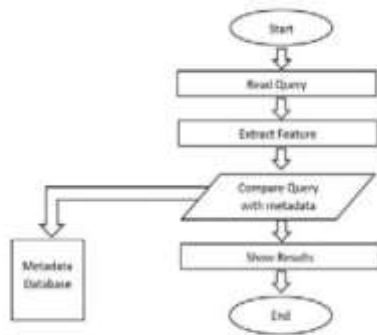


Figure 1: Flowchart of shape-based flowchart detection [11]

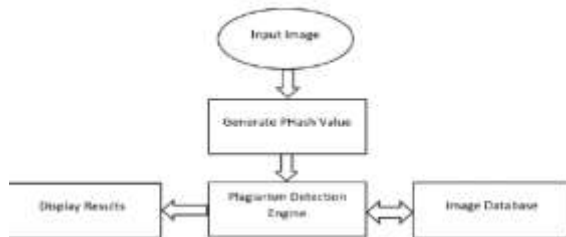


Figure 3: Flowchart of Perceptual Hash [13]

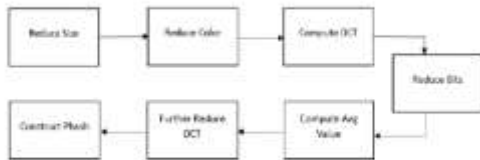


Figure 4: Workflow of Phash Generation [13]

5. CONCLUSION

Plagiarism detection is a well-known phenomenon in the academic arena. Copying other people is considered as a serious offence that needs to be checked. In this paper, an enhanced system to detect the Plagiarism of Images was proposed. The feature extracted from images and saving it to

the databases is color using the RGB and HSV color space, texture using Tamura texture and then shape using canny edge algorithm. The results showed in an

ascending order of similarity index and true and false. However, the accuracy is 100% in case of unedited images and varied in other operations such as flipped, rotated, greyscales and cropped.

6. REFERENCES

- [1]. Green, Stuart P, Plagiarism, Norms, and the Limits of Theft Law: Some Observations on the Use of Criminal Sanctions in Enforcing Intellectual Property Rights, *Hastings Law Journal*, 2002, 54 (1)
- [2]. How Plagiarism Detection Works. (2016, May 18). Retrieved April 17, 2018, from <https://www.plagiarismtoday.com/2016/05/03/plagiarismdetection-works/>
- [3]. Vinod K.R.*, Sandhya.S, Sathish Kumar D, Harani A, David Banji and Otilia JF Banji, Plagiarism – History, Prevention and Detection, *journal for drugs and medicines*, 2011, 3(1), 1-4
- [4]. Eshgh, A. (n.d.). Copyright Timeline: A History of Copyright in the United States. Retrieved April 17, 2018, from <http://www.arl.org/focus-areas/copyright-ip/2486-copyright-timeline>
- [5]. AC Popescu and H Farid, Exposing Digital Forgeries by Detecting Duplicated Image Regions, *Dept Computer Science, Dartmouth College, Hanover*, 2004, 515.
- [6]. B Mahdian and S Saic, Detection of Copy–Move Plagiarism using a Method based on Blur Moment Invariants, *Forensic Science International*, 2007, 171(2), 180-189.
- [7]. JW Wang, GJ Liu, Z Zhang, Y Dai and Z Wang, Fast and robust forensics for image region-duplication Plagiarism, *Acta*



Automatica Sinica, 2009, 35(12), 1488-1495.

[8]. M Zimba, and S Xingming, DWT-PCA (EVD) Based Copy-move Image Plagiarism Detection, International Journal of Digital Content Technology and its Applications, 2011, 5(1), 251-258.

[9]. S Bravo Solorio and AK Nandi, Automated Detection and Localisation of Duplicated Regions Affected by Reflection, Rotation and Scaling, Image Forensics Signal Processing, 2011, 91(8), 1759-1770.

[10]. M Sridevi, C Mala and S Sandeep, Copy-Move Image Plagiarism Detection, Journal of Computer Science and Information Technology, 2012, 52, 19-29.

[11]. Senosy Arrish, Fadhil Noer Afif, Ahmadu Maidorawa and Naomie Salim, Shape-Based Plagiarism Detection for Flowchart Figures in Texts, International Journal of Computer Science & Information Technology (IJCSIT), 2014, 6(1).

[12]. Prajakta Ovhal, Detecting Plagiarism in Images, International Conference on Information Processing (ICIP), 2015.

[13]. Vipul Bajaj, Sanket Keluskar, Ravi Jaisawal and Prof. Rupali Sawant, Plagiarism Detection of Images, International Journal of Innovative and Emerging Research in Engineering, 2015, 2(2).

[14]. Siddharth Srivastava, Prerana Mukherjee and Brejesh Lall, imPlag: Detecting Image Plagiarism Using Hierarchical Near Duplicate Retrieval, IEEE INDICON, 2015.

[15]. Jithin S Kuruvila, Midhun Lal V L, Rejin Roy, Tomin Baby, Sangeetha Jamal, Sherly K K*,

[16]. Flowchart Plagiarism Detection System: An Image Processing Approach, International Conference on Advances in Computing & Communications, 2017.

A STUDY OF BLOCK CHAIN TECHNOLOGY IN FORMERS PORTAL

Anusha Gundumogula (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Blockchain is a method in which a confirmation of a transaction is kept by means of a cryptocurrency. The record is maintained transversely, linking several computers in a peer to peer network. Contracts, transactions, and the records of them define the economic system of a country. They set boundaries and provide security to the assets. Considering the features of blockchain such as immutability and maintaining the footage of transaction details, this paper highlights the usage of blockchain technology with farmer's portal that keep the footage of selling and buying information of crops. The proposed solution uses the python as a programming language in integration with the blockchain system that will benefit the farmers or vendors and individuals by preserving the contract of trade. An interface for the farmers is designed using a python programming language in addition with blockchain technology, which is used to store the information related to seller, buyer, selling and buying an item and total value transacted.

1.INTRODUCTION

Blockchain an open, disseminated and decentralized ledger that evidences transactions involving two parties capably in a confirmable and stable way (Iansiti, Lakhani 2017). In the above given definition, open means the blockchain is accessible to one and all, disseminated means that there is no single party control and decentralized means there is no central third party available, capable means it is fast and more scalable than the conventional technologies, confirmable means that everyone can check the validity of the information and stable means that the data is nearly immutable that is it is nearly impossible to change or tamper the data or information. They

verify and validate the identities and chronological events. They guide every action, transactions that have taken place among individuals, communities, organizations and nations as well. In this era of digitization, the way maintained and regulated these type of data must be changed, it must be highly secure and the blockchain is the solution to this.

In the era of information and communication technology, a farmer's portal has always been helpful for farmers in many ways, providing ease of use and convenience of information to the farmers [1]. The Government of India has also taken many initiatives for the same. Few examples of such portals are Krishijagran.com, farmer.gov.in, agricoop.nic.in and agriwatch.com etc. Apart from these some E-commerce websites are also available; fert.nic.in and enam.gov.in etc. The sectors currently using blockchain are shown in Fig.1. Using blockchain technology in the field can make available decentralized computation and information sharing platform that enables multiple authoritative domains, which do not trust each other, to cooperate, coordinate and collaborate in a rational decisionmaking process, a reliable information recording system can be made that can contribute for the development in the

agriculture sector. Since blockchain works like a public ledger, so it can be utilized to ensure many different aspects

such as [3]:

- **Protocols for Commitment:** Ensure that every valid transaction from the clients are committed and included in the blockchain within a finite time.
- **Consensus:** Ensure that the local copies are consistent and updated.
- **Security:** The data needs to be tamper -proof. Note that the client may act maliciously or can be compromised.
- **Privacy and Authenticity:** The data or transactions belong to various clients; privacy and authenticity need to be ensured.

Cryptography is a foremost part of the functioning of blockchain technology [4]. Public key encryption is the root of blockchain wallets and transaction, cryptography hash functions endow

with the trait of immutability and merle trees systematize transactions while enabling blockchain to be more competent.

Ensuring the above aspects numerous work has been carried out in the field of blockchain. The presented portal is a contribution over them. It can help to maintain a secure platform for farmers, where they can trade with the customers electronically. The main objective of this study is to record the secure transactions between a seller and a buyer that ensures a contract between the two. This can help farmers to get a legitimate price for their commodity. The system also facilitates a single place to record the whole trade transaction.

The availability and accessibility of information are the crucial points in taking the optimal decision at right time. Nowadays, advancement of ICT make possible to retrieve almost any information from the global repository (internet). The information in internet is primarily maintained in English. So, a large number of people are deprived from the benefit of internet due to technical and English language illiteracy. This scenario is very bad in developing country like India where nearly 76 % are English illiterate ¹ . Moreover, a large percentage of the English literate people are also unable to find their exact need form the large database of internet due to lack of proficient knowledge in English. Indian farmers belong to such type of people who are not much sound in both technical as well as in English.

So, they are unable to access required information on the farming life cycle, seed selection, pesticides, market price etc. from the internet. As a consequence, they are not capable to take optimal decisions at different stages of farming life cycle, which make huge impact on the farmer's revenue. As a result suicide rate has been increased rapidly among the Indian farming community. According to the reports, those pathetic incidents are mainly happened due to the frustration that they are unable to pay their debts. These types of situations create huge impact on the agriculture sector. Consequently, the focus of new generation is shifted from farming sector which will be threatening the near future in India. Our preliminary studies reveal that farmers require information at the right stage of the farming life cycle to take the right decisions [1]. However, farmers are unable to get this information from internet due to English language and technical illiteracy. Recently, some webpages like –Wikipedia, Indian Railway web page, etc. provide facility of internet access in many users' language other than English by supporting UTF-8 encoding³ . However, it is observed that information is not so useful to the people who

are having poor knowledge on internet and web browsing [2]. Moreover, this type of attempt is meaningless for the illiterate people. A large number of people from the Indian farmer community are unable to read/write even their own mother tongue. So, it is obvious that text based interface, instead of supporting farmer's own language, are not able to provide the required information. The above mentioned scenario states that there is a requirement of alternative interaction technique(s). By considering this fact, Plauché et al. proposed a speech-driven agricultural query system for Tamil Nadu state of India [3]. However this work does not able to address the scenario of total India. Patel et al. designed an interactive voice application for small-scale farmers in Gujarat, India [4]. However, it does not provide a feature to search for specified content in the forum. There, user needs to answer the questions sequentially starting from the most recent question. User does not have the option to skip any question. Moreover, there is no guarantee of giving accurate answer, as the questions are answered by other users. Furthermore, this work is also confined to a particular area of India. In some recent efforts, expert system based text animation has been proposed for diagnosis of most common diseases occurring in Indian mango [5]. This work also uses picture based system alongwith the text query for easier understanding of the disease symptoms. Though, it is a good initiative for Indian farmer, but limited to a particular fruit. Another notable work was mobile based multimedia social networking platform – GappaGoshti for information and advice exchange, proposed by Lobo et al. [6]. Ramamritham et al. [7] design an online multilingual, multimedia based forum for common man of India. However, those forums and social networking platforms provide limited number of information as compared to the internet. Moreover, quality of information may not be up to the mark, so illiterate people are unable to get any information from there. To overcome the limitation of illiteracy, Samanta et al. [2] proposed and multimodal interface for the Indian common man. However, the iconic module of this work is not related to the agricultural domain. Other works [8, 9] also highlight the need of a systematic approach which is required to provide the precise information to farming oppurmmunity. Moreover, not only providing of the information to farmer, it is also essential to identify that how the farmers are motivated toward accessing the information [10]. All the aforementioned observations motivate us to conduct in depth research toward making an interface for Indian farmer community, which will be more useable, systematical, and needful for them irrespective of language and technical proficiency. Here, we propose an iconic interface integrated with a text to speech (TTS) engine to access the

agricultural information from the internet's global repository for Indian farmer community. Further, we also integrate a local repository with the interface to access urgent information without connecting the internet.

2. LITERATURE SURVEY

2.1 Krishi-Bharat i: an interface for Indian farmer

AUTHORS: Ghosh, Soumalya, A. B. Garg, Sayan Sarcar, PSV S. Sridhar, Ojasvi Maleyvar, and Raveesh Kapoor

Rapid growth in the field of ICT helps in basic aspects of mankind like- agriculture, education, healthcare etc. However, the moderate technical growth of ICT applications is confined to the community of a limited number of people, who live in digital pockets. The illiterate people like – farmer, shopkeeper etc. are unable to take the advantages of the ICT revolution. According to the UNESCO report, population of such people in the globe is 64% who are unable to use the technology either language or technical barrier. Moreover the percentage (76%) must be increased in the context of developing countries. The essential agriculture information is very useful to a farmer for taking effective decision thus we proposed to develop an iconic interface which is integrated with speech based interaction in Indian languages. The proposed interface is critically evaluated with the farmer from different states of India. The evaluation results proved the effectiveness of the proposed interface.

2.2 Krishi Ville—Android based solution for Indian agriculture

AUTHORS: Singhal, Manav, Kshit ij Verma, and Anupam Shukla

Information and Communication Technology (ICT) in agriculture is an emerging field focusing on the enhancement of agricultural and rural development in India. It involves innovative applications using ICT in the rural domain. The advancement of ICT can be utilized for providing accurate and timely relevant information and services to the farmers, thereby facilitating an environment for remunerative agriculture. This paper describes a mobile based application for farmers which would help them in their farming activities. We propose an android based mobile application - Krishi Ville which would take care of the updates of the

different agricultural commodities, weather forecast updates, agricultural news updates. The application has been designed taking indian farming in consideration.

2.3 Blockchain based provenance for agricultural products:

A distributed platform with duplicated and shared bookkeeping

AUTHORS : Hua, Jing, Xiujuan Wang, Mengzhen Kang, Haoyu Wang, and Fei-

The provenance (tracing) system of agricultural products is important for ensuring food safety. However, the stakeholders (growers, farmers, sellers etc.) are numerous and physically dispersed, making it difficult to manage data and information with a centralized approach. As a result, the production procedure remains non-transparent and trust is hard to build. In this paper, we propose an agricultural provenance system based on techniques of blockchain, which is featured by decentralization, collective maintenance, consensus trust and reliable data, in order to solve the trust crisis in product supply chain. Recorded information includes the management operations (fertilizing, irrigation, etc.) with certain data structure. Applying blockchain techniques to the provenance of agricultural product not only widens the application domain of blockchain, but also supports building a reliable community among different stakeholders around agriculture production.

2.4 Bitcoin and beyond: A technical survey on decentralized digital currencies

AUTHORS : Tschorsch, Florian, and Björn Scheuermann

Besides attracting a billion dollar economy, Bitcoin revolutionized the field of digital currencies and influenced many adjacent areas. This also induced significant scientific interest. In this survey, we unroll and structure the manifold results and research directions. We start by introducing the Bitcoin protocol and its building blocks. From there we continue to explore the design space by discussing existing contributions and results. In the process, we deduce the fundamental structures and insights at the core of the Bitcoin protocol and its applications. As we show and discuss, many key ideas are likewise applicable in various other fields, so that their impact reaches far beyond Bitcoin itself.

2.5 Towards using ICT to enhance flow of information to aid farmer sustainability in Sri Lanka

AUTHORS: L. N. De Silva, J. S. Goonetillake, G. N. Wikramanayake, and A.

Ginige

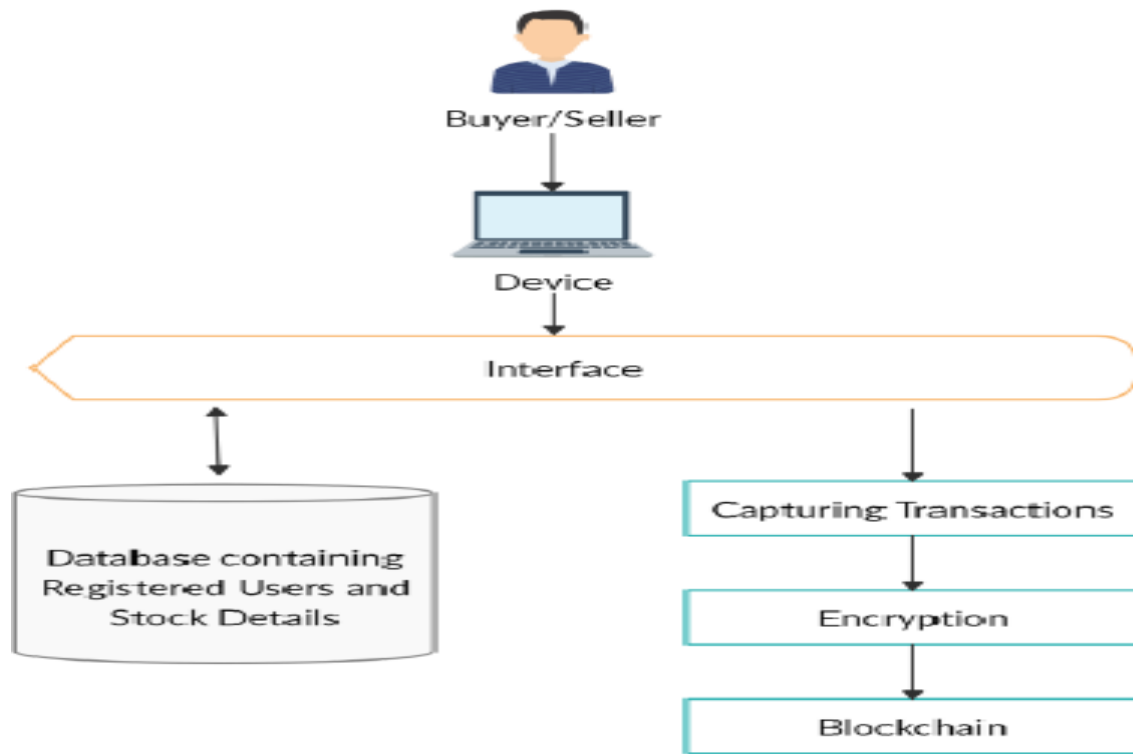
Farmers need information at all stages of the farming life cycle to make optimal decisions. The required information includes not only prior knowledge but also real time (dynamic) information such as market prices and current production levels. Some valuable information needed by the farmers is produced by government organizations and is available in different locations in different formats. Although farmer is the most important stakeholder in agriculture, there has not been much effort to provide the essential information to farmers on a real time basis. This lack of information is creating many difficulties for farmers as they are not being able to make the correct decisions relating to their farming activities. Through field studies we have identified information required by farmers at various stages of the farming cycle and official sources where this information is available. Next we developed an information flow model that connects various information sources to farmers' information needs. Based on these findings we are now developing a mobile phone based information system to deliver the required information to farmers in real time.

3. Further Enhancement

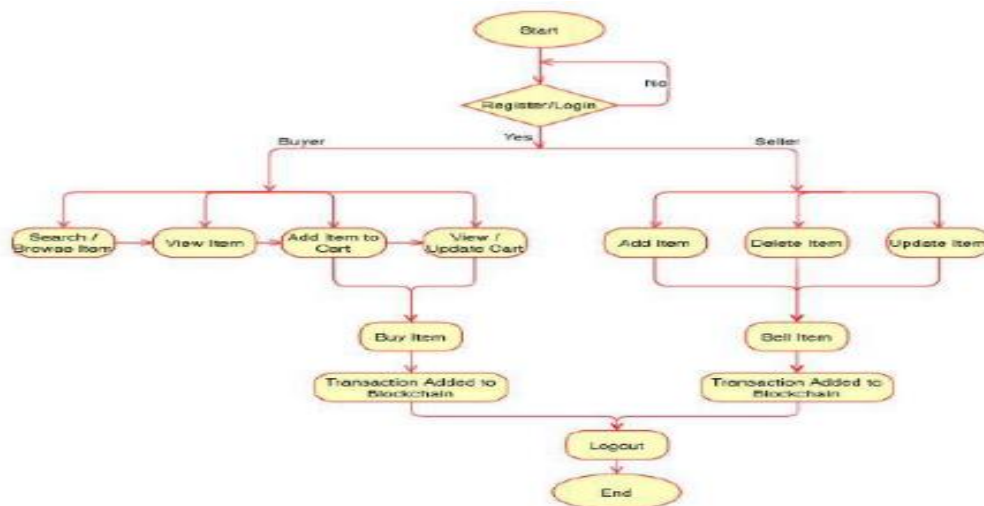
Akin kind of portal can be implemented by the government and its confederate bureaus to ensure amelioration in the field of farming and commerce of crops which will improve the prominence of the nation's farmers. This application can be more refined with increasing integration of blockchain in a spectrum of areas and constellating it into a single paramount portal for farmers. This can be done by putting farmer's crop details to the blockchain, buyer's data to the blockchain and adding more features and services to the single portal and bringing all possible facilities for farmers of the nation under sui generis awning. Information integrity and precision issues can be solved using open, protected and trusted systems presumptuous; the infrastructure dispensation and footage connections are protected and suitably provided. The blockchain

technology did not promise the information reliability in the footage. Thus realization in blockchain faces several boundaries that might require a vital authority or protected footage of confirmation.

4. SYSTEM ARCHITECTURE:



5. DATAFLOW DIAGRAM:



6. CONCLUSION

Blockchain Technology in the field of agriculture can bring a revolutionary enhancement in the area of maintaining farmers data securely, ensuring the quality of seed, monitoring of moisture content in the soil, data of crop yield and lastly demand and sale price of crops. In this work, a blockchain-based portal is proposed to deal with the issue of demand and sale price of crops which in result ensure crop security to farmers as well as to get fair price of the crop. For this, a portal is proposed on which a farmer can register and sell his crops, recording a transaction on a blockchain at a point when buyers commit to buy a farmer's crop. This transaction is capable of recording crop details, the price at which it is committed to buying and quantity of crop purchased. This immutable nature of blockchain technology will fortify farmers to get a legitimate price of crop and reduce the cost of operation for selling and buying crops when compared to traditional methods.

7. REFERENCES

- [1] Lakhani, Karim R., and M. Iansiti. "The truth about blockchain." *Harvard Business Review* 95 (2017): 118-127.
- [2] Hileman, Garrick, and Michel Rauch. "2017 global blockchain benchmarking study." Available at SSRN 3040224 (2017).
- [3] Mohanta, Bhabendu K., Debasish Jena, Soumyashree S. Panda, and Srichandan Sobhanayak. "Blockchain Technology: A Survey on Applications and Security Privacy Challenges." *Internet of Things* (2019): 100107.
- [4] Yadav, Vinay Surendra, and A. R. Singh. "A Systematic Literature Review of Blockchain Technology in Agriculture."
- [5] Ghosh, Soumalya, A. B. Garg, Sayan Sarcar, PSV S. Sridhar, Ojasvi Maleyvar, and Raveesh Kapoor. "Krishi-Bharati: an interface for Indian farmer." In *Proceedings of the 2014 IEEE Students' Technology Symposium*, pp. 259-263. IEEE, 2014.

[6] Singhal, Manav, Kshitij Verma, and Anupam Shukla. "Krishi Ville— Android based solution for Indian agriculture." In 2011 Fifth IEEE international conference on advanced telecommunication systems and networks (ANTS), pp. 1-5. IEEE, 2011.

[7] Potts, Jason. "Blockchain in Agriculture." Available at SSRN 3397786 (2019).

[8] Hua, Jing, Xiujuan Wang, Mengzhen Kang, Haoyu Wang, and Fei-Yue Wang. "Blockchain based provenance for agricultural products: A distributed platform with duplicated and shared bookkeeping." In 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 97-101. IEEE, 2018.

[9] Zhu, Xingxiong, and Dong Wang. "Research on Blockchain Application for E-Commerce, Finance and Energy." In IOP Conference Series: Earth and Environmental Science, vol. 252, no. 4, p. 042126. IOP Publishing, 2019.

DETECTION OF DEPRESSION-RELATED POSTS IN REDDIT SOCIAL MEDIA FORUM

Athava Bhanu Mathi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West
Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

ABSTRACT :- Depression is viewed as the largest contributor to global disability and a major as on for suicide. It has an impact on the language usage reflected in the written text. The key objective of our study is to examine Reddit users' posts to detect any factors that may reveal the depression attitudes of relevant online users. For such purpose, we employ the Natural Language Processing (NLP) techniques and machine learning approaches to train the data and evaluate the efficiency of our proposed method. We identify a lexicon of terms that are more common among depressed accounts. The results show that our proposed method can significantly improve performance accuracy. The best single feature is bigram with the Support Vector Machine (SVM) classifier to detect depression with 80% accuracy and 0.80 F1 scores. The strength and effectiveness of the combined features (LIWC+LDA+bigram) are most successfully demonstrated with the Multilayer Perceptron (MLP) classifier resulting in the top performance for depression detection reaching 91% accuracy and 0.93 F1 scores. According to our study, better performance improvement can be achieved by proper feature selections and their multiple feature combinations.

1. INTRODUCTION

Depression as a common mental health disorder has long been defined as a single disease with a set of diagnostic criteria. It often co-occurs with anxiety or other psychological and physical disorders; and has an impact on feelings and behavior of the affected individuals [1]. According to the WHO study, there are 322 million people estimated to suffer from depression, equivalent to 4.4% of the global population. Nearly half of the in-risk individuals live in the South-East Asia (27%) and Western Pacific region (27%) including China and India. In many countries depression is still under-diagnosed and left without any adequate treatment which can lead into a

serious self-perception and at its worst, to suicide [2]. In addition, the social stigma surrounding depression prevents many affected individuals from seeking an appropriate professional assistance.

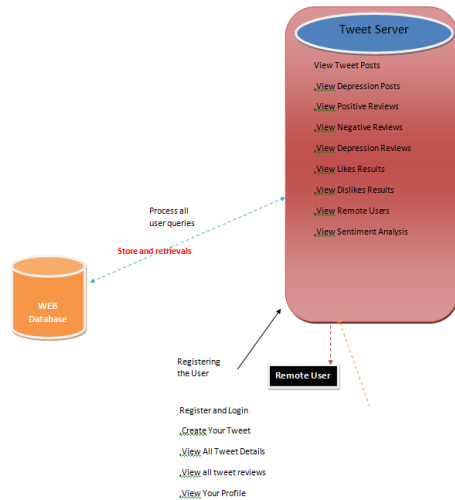
As a result, they turn to less formal resources such as social media. With the development of Internet usage, people have started to share their experiences and challenges with mental health disorders through online forums, micro-blogs or tweets. Their online activities inspired many researchers to introduce new forms of potential health care solutions and methods for early depression detection systems. Using different Natural Language Processing (NLP) techniques and text classification approaches, they tried to succeed in a higher performance improvement. Some studies use single set features, such as bag of words (BOW) [3], [4], N-grams [5], LIWC [6] or LDA [7], [8] to identify depression in their posts. Some other papers compare the performance of individual features with various machine learning classifiers [9] [12]. Recent studies examine the power of single features and their combinations such as N-grams+LIWC [13] or BOW+LDA and TF-IDF+LDA [14] to improve the accuracy results. They experiment with a smarter text pre-processing, and introduce different substitute words depending on the nature of the original string. For instance, Tyshchenko et al. [14] suggested categorizing the stop words and adding LIWC-like word categories as an extra feature to an already designed method (BOW+TFIDF+LIWC). In addition, he applied multiple feature combinations to increase the performance using Convolutional Neural Networks (CNN) which consist of neurons with learnable weights and differ in terms of their layers. CNNs are very similar to simple feed-forward neural networks and state of the art method in the text and sentence classification tasks. A meta-analysis by Guntuku et al. [15] summarizes several iterations of depression detection tasks in computational linguistics. Another interesting review for mental health support and intervention in social media is written by Calvo et al. [16] who reviewed the taxonomy of data sources, NLP techniques and computational methods to detect various mental health applications. Even with this significant progress, challenges still remain. This paper aims to search for a solution to a performance increase through a proper features selection and their multiple feature combinations. First, we choose the most beneficial linguistic features applied for depression identification to characterize the content of the posts. Second, we analyze the correlation significance, hidden topics and word frequency extracted from the text. Regarding the correlation, we focus on the LIWC dictionary and its three feature

types (linguistic dimensions, psychological processes and personal concerns). For the topic examination, we choose the LDA method as one of the successful features. For the word frequency, we use unigrams and bigrams by leveraging the vectors based on TF-IDF scheme. Finally, we set five text classifying techniques and conduct their execution using the extracted data to detect depression. We compare the performance results based on three single feature sets and their multiple feature combinations. In our experiment, we use data collected from the Reddit social media platform. It was chosen as the

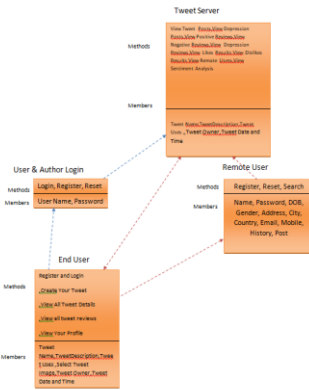
data source as it allows longer posts. Targeting technical approaches towards detection tasks, our paper follows the lines of Calvo et al. research [17].

Our study has four specific contributions: first, to examine the relationship between depression and user’s language usage; second, to design three LIWC features for our specific research problem; third, to evaluate the power of N-grams probabilities, LIWC and LDA as single features for performance accuracy; fourth, to show the predictive power of both single and combined features with proposed classification approaches to achieve a higher performance in depression identification tasks.

Architecture Diagram



2. CLASS DIAGRAM



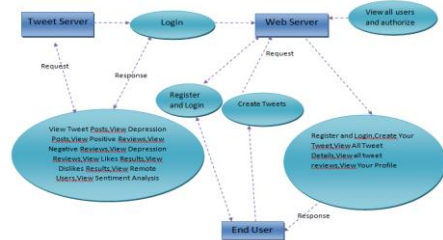
The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed.

In the diagram, classes are represented with boxes which contain three parts

- The upper part holds the name of the class
- The middle part contains the attributes of the class
- The bottom part gives the methods or operations the class can take or undertake

In the design of a system, a number of classes are identified and grouped together in a class diagram which helps to determine the static relations between those objects. With detailed modeling, the classes of the conceptual design are often split into a number of subclasses.

3. DATAFLOW DIAGRAM



3.1 PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSIONS

In this paper, we tried to identify the presence of depression in Reddit social media; and searched for affective performance increase solutions of depression detection. We characterized a closer connection between depression and a language usage by applying NLP and text classification techniques. We identified a lexicon of words more common among the depressed accounts. According to our findings, the language predictors of depression contained the words related to preoccupation with themselves, feelings of sadness, anxiety, anger, hostility or suicidal thoughts, with a greater emphasis on the present and future. To measure the signs of depression, we examined the performance of both single feature and combined feature sets using various text classifying methods. Our results show that a higher predictive performance is hidden in proper features selection and their multiple feature combinations. The strength and effectiveness of combined features are demonstrated with the MLP classifier reaching 91% accuracy and 0.93 F1 score achieving the highest performance degree for detecting the presence of depression in Reddit social media conducted in our study. Additionally, the best feature among the single feature sets is bigram; with SVM classifier it can detect depression with 80% accuracy and 0.79 F1 score. Considering LIWC and LDA features, LIWC outperformed topic models generated by LDA. Although our experiment shows that the performances of applied methodologies are reasonably good, the absolute values of the metrics indicate that this is a challenging task and worthy of further exploration. We believe this experiment could further underline the infrastructure for new mechanisms applied in different areas of healthcare to estimate depression and related variables. It can be beneficial for the individuals suffering from mental health disorders to be more proactive towards their fast recovery. In our future work, we will try to examine the relationship between the users' personality [65] and their depression-related behavior reflected in social media.

7. REFERENCES

- [1] W. H. Organization, “Depression and other common mental disorders: Global health estimates. geneva: World health organization; 2017. licence: Cc by-nc-sa 3.0 igo.” <http://www.who.int/en/news-room/fact-sheets/detail/depression>, 2017.
- [2] M. Friedrich, “Depression is the Leading Cause of Disability Around the World Depression Leading Cause of Disability Globally Global Health,” *JAMA*, vol. 317, no. 15, pp. 1517–1517, 2017.
- [3] M. Nadeem, “Identifying depression on twitter,” *CoRR*, vol. abs/1607.07384, 2016.
- [4] S. Paul, S. K. Jandhyala, and T. Basu, “Early detection of signs of anorexia and depression over social media using effective machine learning frameworks,” in *CLEF*, 2018.
- [5] A. Benton, M. Mitchell, and D. Hovy, “Multi-task learning for mental health using social media text,” *CoRR*, vol. abs/1712.03538, 2017.
- [6] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, “From adhd to sad: Analyzing the language of mental health on twitter through selfreported diagnoses,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 1–10.
- [7] D. Maupomé and M.-J. Meurs, “Using topic extraction on social media content for the early detection of depression,” in *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.
- [8] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, “Beyond lda: Exploring supervised topic modeling for depression-related language in twitter,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 99–107.
- [9] D. Preotiuc-Pietro, J. C. Eichstaedt, G. J. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. H. Ungar, “The role of personality, age, and gender in tweeting about mental illness,” in *CLPsych@HLT-NAACL*, 2015.
- [10] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, “Affective and content analysis of online depression communities,” *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

DEEP LEARNING APPLICATIONS IN MEDICAL IMAGE ANALYSIS-BRAIN TUMOR

Basivireddy Suresh Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- The tremendous success of machine learning algorithms at image recognition tasks in recent years intersects with a time of dramatically increased use of electronic medical records and diagnostic imaging. This review introduces the machine learning algorithms as applied to medical image analysis, focusing on convolutional neural networks, and emphasizing clinical aspects of the field. The advantage of machine learning in an era of medical big data is that significant hierarchical relationships within the data can be discovered algorithmically without laborious hand-crafting of features. We cover key research areas and applications of medical image classification, localization, detection, segmentation, and registration. We conclude by discussing research obstacles, emerging trends, and possible future directions.

1. INTRODUCTION

Machine learning algorithms have the potential to be invested deeply in all fields of medicine, from drug discovery to clinical decision making, significantly altering the way medicine is practiced. The success of machine learning algorithms at computer vision tasks in recent years comes at an opportune time when medical records are increasingly digitalized. The use of electronic health records (EHR) quadrupled from 11.8% to 39.6% amongst office-based physicians in the US from 2007 to 2012 [1]. Medical images are an integral part of a patient's EHR and are currently analyzed by human radiologists, who are limited by speed, fatigue, and experience. It takes years and great financial cost to train a qualified radiologist, and some health-care systems outsource radiology reporting to lower-cost countries such as India via tele-radiology. A delayed or erroneous diagnosis causes harm to the patient. Therefore, it is ideal for medical image analysis to be carried out by

an automated, accurate and efficient machine learning algorithm.

2. LITERATURE SURVEY

- Trends in electronic health record system use among office-based physicians: United states, 2007-2012
The National Ambulatory Medical Care Survey (NAMCS) is based on a national probability sample of nonfederal office-based physicians who see patients in an office setting. Prior to 2008, data on physician characteristics were collected through in-person interviews with physicians. To increase the sample for analyzing physician adoption of EHR systems, starting in 2008, NAMCS physician interview data were supplemented with data from an EHR mail survey. This report presents estimates from the 2007 in-person interviews, combined 2008-2010 data from both the in-person interviews and the EHR mail



surveys, and 2011-2012 data from the EHR mail surveys. Sample data were weighted to produce national estimates of office-based physician characteristics and their practices.

- Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems

Context: Use of diagnostic imaging has increased significantly within fee-for-service models of care. Little is known about patterns of imaging among members of integrated health care systems.

Objective: To estimate trends in imaging utilization and associated radiation exposure among members of integrated health care systems. Design, setting, and participants: Retrospective analysis of electronic records of members of 6 large integrated health systems from different regions of the United States. Review of medical records allowed direct estimation of radiation exposure from selected tests. Between 1 million and 2 million member-patients were included each year from 1996 to 2010. Main outcome measure: Advanced diagnostic imaging rates and cumulative annual radiation exposure from medical imaging. Results: During the 15-year study period, enrollees underwent a total of 30.9 million imaging examinations (25.8 million person-years), reflecting 1.18 tests (95% CI, 1.17-1.19) per person per year, of which 35% were for

advanced diagnostic imaging (computed tomography [CT], magnetic resonance imaging [MRI], nuclear medicine, and ultrasound). Use of advanced diagnostic imaging increased from 1996 to 2010; CT examinations increased from 52 per 1000 enrollees in 1996 to 149 per 1000 in 2010, 7.8% annual increase (95% CI, 5.8%-9.8%); MRI use increased from 17 to 65 per 1000 enrollees, 10% annual growth (95% CI, 3.3%-16.5%); and ultrasound rates increased from 134 to 230 per 1000 enrollees, 3.9% annual growth (95% CI, 3.0%-4.9%). Although nuclear medicine use decreased from 32 to 21 per 1000 enrollees, 3% annual decline (95% CI, 7.7% decline to 1.3% increase), PET imaging rates increased after 2004 from 0.24 to 3.6 per 1000 enrollees, 57% annual growth. Although imaging use increased within all health systems, the adoption of different modalities for anatomic area assessment varied. Increased use of CT between 1996 and 2010 resulted in increased radiation exposure for enrollees, with a doubling in the mean per capita effective dose (1.2 mSv vs 2.3 mSv) and the proportion of enrollees who received high (>20-50 mSv) exposure (1.2% vs 2.5%) and very high (>50 mSv) annual radiation exposure (0.6% vs 1.4%). By 2010, 6.8% of enrollees who underwent imaging received high annual radiation exposure (>20-50 mSv) and



3.9% received very high annual exposure (>50 mSv).

- A survey on deep learning in medical image analysis.

Deep learning algorithms, in particular convolutional networks, have rapidly become a methodology of choice for analyzing medical images. This paper reviews the major deep learning concepts pertinent to medical image analysis and summarizes over 300 contributions to the field, most of which appeared in the last year. We survey the use of deep learning for image classification, object detection, segmentation, registration, and other tasks. Concise overviews are provided of studies per application area: neuro, retinal, pulmonary, digital pathology, breast, cardiac, abdominal, musculoskeletal. We end with a summary of the current state-of-the-art, a critical discussion of open challenges and directions for future research.

- A logical calculus of the ideas immanent in nervous activity
Because of the “all-or-none” character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it

describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

- The perceptron: A probabilistic model for information storage and organization in the brain

To answer the questions of how information about the physical world is sensed, in what form is information remembered, and how does information retained in memory influence recognition and behavior, a theory is developed for a hypothetical nervous system called a perceptron. The theory serves as a bridge between biophysics and psychology. It is possible to predict learning curves from neurological variables and vice versa. The quantitative statistical approach is fruitful in the understanding of the organization of cognitive systems. 18 references. (PsycINFO Database Record (c) 2016 APA, all rights reserved.

- Receptive elds, binocular interaction and functional architecture in the cat's visual cortex

What chiefly distinguishes cerebral cortex from other parts of the central nervous system is the great diversity of its cell types and inter-connexions. It would be astonishing if such a structure did not profoundly modify the response patterns of fibres coming into it. In the cat's visual cortex, the receptive field arrangements of



single cells suggest that there is indeed a degree of complexity far exceeding anything yet seen at lower levels in the visual system. In a previous paper we described receptive fields of single cortical cells, observing responses to spots of light shone on one or both retinas (Hubel & Wiesel, 1959). In the present work this method is used to examine receptive fields of a more complex type (Part I) and to make additional observations on binocular interaction (Part II). This approach is necessary in order to understand the behaviour of individual cells, but it fails to deal with the problem of the relationship of one cell to its neighbours. In the past, the technique of recording evoked slow waves has been used with great success in studies of functional anatomy. It was employed by Talbot & Marshall (1941) and by Thompson, Woolsey & Talbot (1950) for mapping out the visual cortex in the rabbit, cat, and monkey. Daniel & Whittetide (1959) have recently extended this work in the primate. Most of our present knowledge of retinotopic projections, binocular overlap, and the second visual area is based on these investigations. Yet the method of evoked potentials is valuable mainly for detecting behaviour common to large populations of neighbouring cells; it cannot differentiate functionally between areas of cortex smaller than about 1 mm². To overcome this difficulty a method has in recent years been developed for studying cells separately or in small groups during long micro-electrode penetrations through nervous tissue. Responses are correlated with cell location by reconstructing the electrode tracks from histological material. These techniques have been applied to CAT VISUAL CORTEX

107 the somatic sensory cortex of the cat and monkey in a remarkable series of studies by Mountcastle (1957) and Powell & Mountcastle (1959). Their results show that the approach is a powerful one, capable of revealing systems of organization not hinted at by the known morphology. In Part III of the present paper we use this method in studying the functional architecture of the visual cortex.

- Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition

A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by “learning without a teacher”, and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname “neocognitron”. After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The network consists of an input layer (photoreceptor array) followed by a cascade connection of a number of modular structures, each of which is composed of two layers of cells connected in a cascade. The first layer of each module consists of “S-cells”, which show characteristics similar to simple cells or lower order hypercomplex cells, and the second layer consists of “C-cells” similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. The network has an ability of unsupervised learning: We do not need

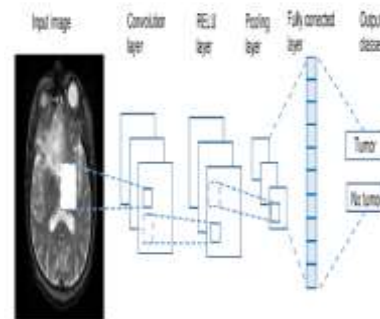
any “teacher” during the process of self-organization, and it is only needed to present a set of stimulus patterns repeatedly to the input layer of the network. The network has been simulated on a digital computer. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cell of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at all. Neither is it affected by a small change in shape nor in size of the stimulus pattern.

3. EXISTING SYSTEM

There is a myriad of imaging modalities, and the frequency of their use is increasing. Smith-Bindman *et al.* [2] looked at imaging use from 1996 to 2010 across six large integrated healthcare systems in the United States, involving 30.9 million imaging examinations. The authors found that over the study period, CT, MRI and PET usage increased 7.8%, 10% and 57% respectively. The symbolic AI paradigm of the 1970s led to the development of rule-based, expert systems. One early implementation in medicine was the MYCIN system by Shortliffe [3], which suggested different regimes of antibiotic therapies for patients. Parallel to these developments, AI algorithms moved from heuristics-based techniques to manual, handcrafted feature extraction techniques. and then to supervised learning techniques. Unsupervised machine learning methods are also being researched, but the majority of the algorithms from

2015-2017 in the published literature have employed supervised learning methods,

4. SYSTEM ARCHITECTURE



5. CONCLUSION

A recurring theme in machine learning is the limit imposed by the lack of labelled datasets, which hampers training and task performance. Conversely, it is acknowledged that more data improves performance, as Sun *et al.* shows using an internal Google dataset of 300 million images. In general computer vision tasks, attempts have been made to circumvent limited data by using smaller filters on deeper layers, with novel CNN architecture combinations, or hyperparameter optimization

6. REFERENCES

- [1] C.-J. Hsiao, E. Hing, and J. Ashman, “Trends in electronic health record system use among office-based physicians: United states, 2007-2012,” *Nat. Health Stat. Rep.*, vol. 75, pp. 118, May 2014.
- [2] R. Smith-Bindman et al., “Use of diagnostic imaging studies and associated radiation exposure for patients enrolled in large integrated health care systems, 1996-2010,” *JAMA*, vol. 307, no. 22, pp. 2400-2409, 2012.



[3] E. H. Shortliffe, Computer-Based
Medical Consultations: MYCIN, vol. 2.
New York, NY, USA: Elsevier, 1976.

FISH DISEASE DETECTION USING IMAGE BASED MACHINE LEARNING TECHNIQUE IN AQUACULTURE

Bodducharla Rajesh (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract :- Fish diseases in aquaculture constitute a significant hazard to nutriment security. Identification of infected fishes in aquaculture remains challenging to find out at the early stage due to the dearth of necessary infrastructure. The identification of infected fish timely is an obligatory step to thwart from spreading disease. In this work, we want to find out the salmon fish disease in aquaculture, as salmon aquaculture is the fastest-growing food production system globally, accounting for 70 percent (2.5 million tons) of the market. In the alliance of flawless image processing and machine learning mechanism, we identify the infected fishes caused by the various pathogen. This work divides into two portions. In the rudimentary portion, image pre-processing and segmentation have been applied to reduce noise and exaggerate the image, respectively. In the second portion, we extract the involved features to classify the diseases with the help of the Support Vector Machine (SVM) algorithm of machine learning with a kernel function. The processed images of the first portion have passed through this (SVM) model. Then we harmonize a comprehensive experiment with the proposed combination of techniques on the salmon fish image dataset used to examine the fish disease. We have conveyed this work on a novel dataset compromising with and without image augmentation. The results have bought a judgment of our applied SVM performs notably with 91.42 and 94.12 percent of accuracy, respectively, with and without augmentation.

1. INTRODUCTION

The word aquaculture is related to firming, including breeding, raising, and harvesting fishes, aquatic plants, crustaceans, mollusks, and aquatic organisms. It involves the cultivation of both freshwater and saltwater creatures under a controlled condition and is used to produce food and commercial products as shown in Figure 1. There are mainly two types of aquaculture. The first

one is Mariculture which is the farming of marine organisms for food and other products such as pharmaceuticals, food additives, jewelry (e.g., cultured pearls), nutraceuticals, and cosmetics. Marine organisms are farmed either in the natural marine environment or in the land- or sea-based enclosures, such as cages, ponds, or raceways. Seaweeds, mollusks, shrimps, marine fish, and a wide range of other minor species such as sea cucumbers and sea horses are among the wide range of organisms presently farmed around the world's coastlines. It contributes to sustainable food production and the economic development of local communities. However, sometimes at a large scale of marine farming become a threat to marine and coastal environments like degradation of natural habitats, nutrients, and waste discharge, accidental release of alien organisms, the transmission of diseases to wild stocks, and displacement of local and indigenous communities.

The second one is Fish farming which is the cultivation of fish for commercial purposes in human-made tanks and other enclosures. Usually, some common types of fish like catfish, tilapia, salmon, carp, cod, and trout are farmed in these enclosures. Nowadays, the fish-farming industry has grown to meet the demand for fish products [34]. This form of aquaculture is widespread for a long time as it is said to produce a cheap source of protein.

Global aquaculture is one of the quickest growing food productions, accounting for almost 53% of all fish and invertebrate production and 97% of the total seaweed manufacture as of 2020. Estimated global production of farmed salmon stepped up by 7 percent in 2019, to just over 2.6 million tonnes of the market [12]. Global aquaculture of salmon has a threat of various diseases that can devastate the conventional production of salmon.

Diseases have a dangerous impact on fishes in both the natural environment and in aquaculture. Diseases are globally admitted as one of the most severe warnings to the economic success of aquaculture. Diseases of fishes are provoked by a spacious range of contagious organisms such as bacteria, viruses, protozoan, and metazoan parasites. Bacteria are accountable for the preponderance of the contagious diseases in confined fish [26]. Infectious diseases create one in every foremost vital threat to victorious aquaculture. The massive numbers of fishes gathered in a tiny region gives an ecosystem favorable for development and quickly spreads contagious diseases. In this jam-packed situation, a comparatively fabricated environment, fishes are stressed and also respond to disease. Furthermore, the water ecosystem and insufficient water

flow make it easier for the spread of pathogens in gathered populations [29]. Detection of disease with the cooperation of some image processing can help to extract good features.

Image segmentation becomes indispensable for various research fields like computer vision, artificial intelligence, etc. The k means segmentation is a popular image processing technique that mainly partitions different regions in an image without loss of information. In [18], authors applied k means segmentation for authentication of images. Another application of k means segmentation shown at [11] where they use this technique to recognize handwritten Hindi characters.

One of the most popular supervised machine learning techniques, support vector machine (SVM), has brought convenient solutions for many classification problems in various fields. It is a powerful classification tool that brings out quality predictions for unlabeled data. In [19] Authors built an SVM model based on three kernel functions to differentiate dengue human infected blood sera and healthy sera. For image classification, another SVM architecture has been proposed in [3] where they emulate the architecture by combining convolutional neural network (CNN) with SVM. SVM provides remarkable accuracy in many contexts.

2. Literature Survey

2.1 Extended cubic b-spline interpolation method applied to linear two-point boundary value problem.

Second order linear two-point boundary value problems were solved using extended cubic B-spline interpolation method. Extended cubic B-spline is an extension of cubic B-spline consisting of one shape parameter, called λ . The resulting approximated analytical solution for the problems would be a function of λ . Optimization of λ was carried out to find the best value of λ that generates the closest fit to the differential equations in the problems. This method approximated the solutions for the problems much more accurately compared to finite difference, finite element, finite volume and cubic B-spline interpolation methods.

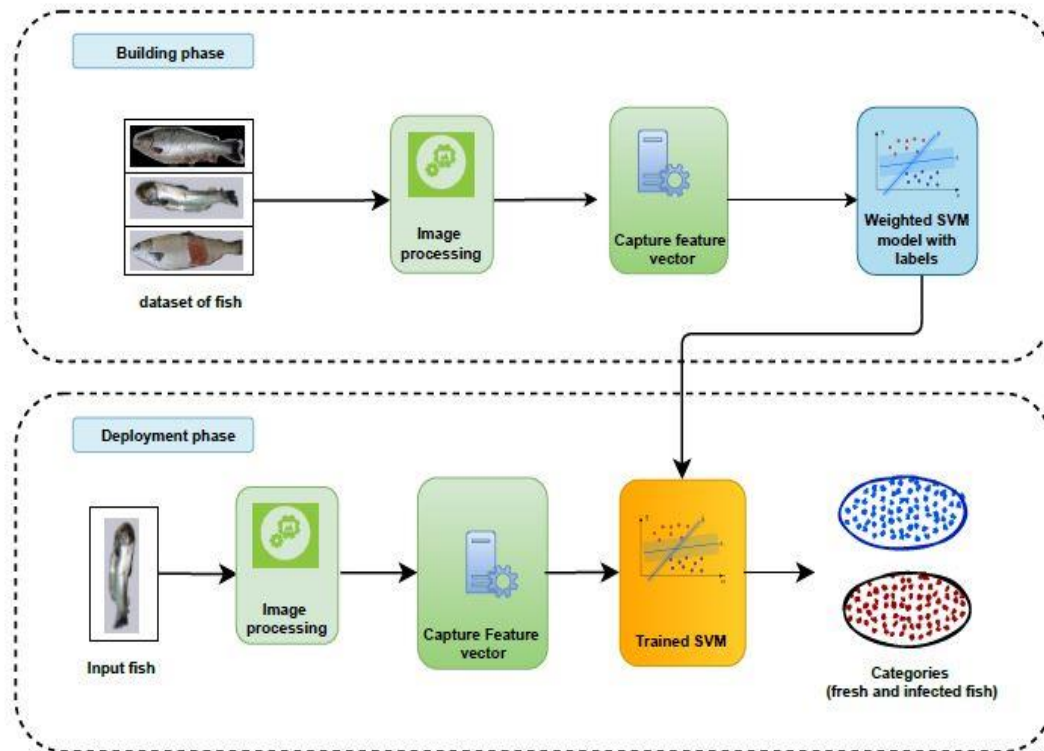
Median computation-based integrated color interpolation and color space conversion methodology from 8-bit bayer pattern rgb color space to 24-bit cie xyz color space

What is disclosed is an integrated color interpolation and color space conversion technique and apparatus. A raw image that is arranged in a Bayer pattern where each pixel has only one of the color components needed to form a full color resolution pixel may be converted using this technique directly to a XYZ space image without any intermediate conversion or interpolation steps. Specifically, in one instance, an 8-bit Bayer pattern raw image may be converted directly to a 24-bit XYZ space in a single pass approach. A method comprising: providing an integrated color interpolation and color space conversion technique, said technique including determination of a missing color component for a pixel location by determining a median of adjacent pixels associated with the same color of said missing color component and integrating the operation of color interpolation and color space conversion into a single operation; and applying said technique to pixels of a raw image, said raw image pixels without full color resolution, said technique generating therefrom a color.

An architecture combining convolutional neural network (cnn) and support vector machine (svm) for image classification. Convolutional neural networks (CNNs) are similar to "ordinary" neural networks in the sense that they are made up of hidden layers consisting of neurons with "learnable" parameters. These neurons receive inputs, performs a dot product, and then follows it with a non-linearity. The whole network expresses the mapping between raw image pixels and their class scores. Conventionally, the Softmax function is the classifier used at the last layer of this network. However, there have been studies (Alalshekmubarak and Smith, 2013; Agarap, 2017; Tang, 2013) conducted to challenge this norm. The cited studies introduce the usage of linear support vector machine (SVM) in an artificial neural network architecture. This project is yet another take on the subject, and is inspired by (Tang, 2013). Empirical data has shown that the CNN-SVM model was able to achieve a test accuracy of ~99.04% using the MNIST dataset (LeCun, Cortes, and Burges, 2010). On the other hand, the CNN-Softmax was able to achieve a test accuracy of ~99.23% using the same dataset. Both models were also tested on the recently-published Fashion-MNIST dataset (Xiao, Rasul, and Vollgraf, 2017), which is suppose to be a more difficult image classification dataset than MNIST (Zalandoresearch, 2017). This proved to be the case as CNN-SVM reached a test accuracy of ~90.72%, while the CNN-Softmax reached

a test accuracy of ~91.86%. The said results may be improved if data preprocessing techniques were employed on the datasets, and if the base CNN model was a relatively more sophisticated than the one used in this study.

3. SYSTEM ARCHITECTURE



4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

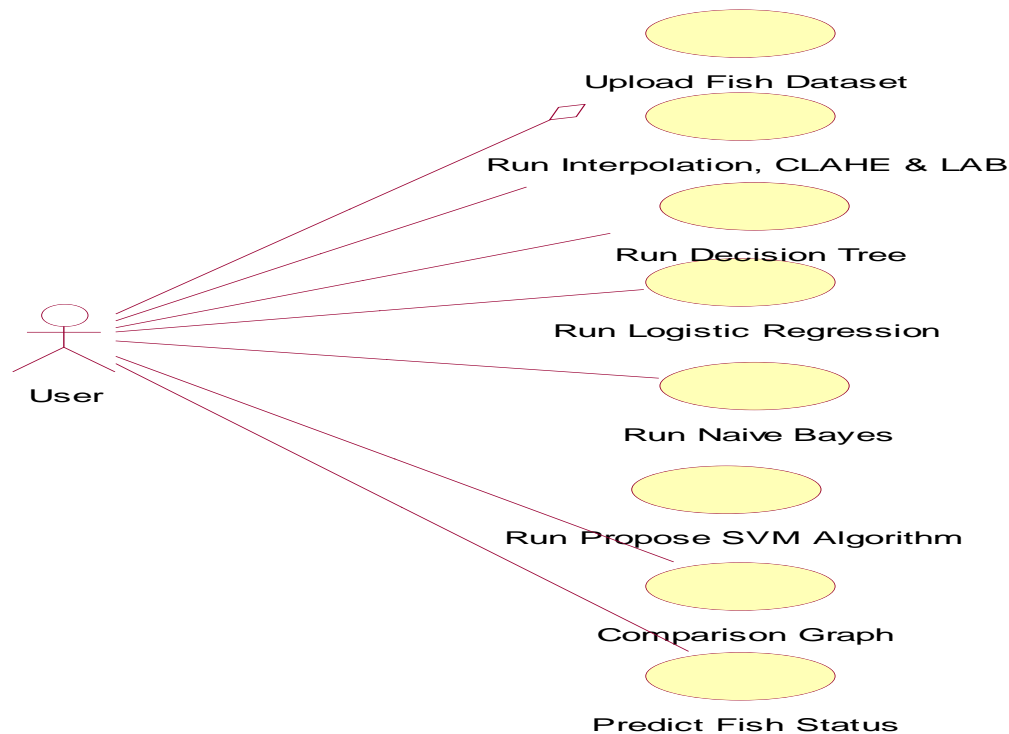
- ◆ ECONOMICAL FEASIBILITY

- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

5. GOALS

1. The Primary goals in the design of the UML are as follows:
2. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
3. Provide extendibility and specialization mechanisms to extend the core concepts.
4. Be independent of particular programming languages and development process.
5. Provide a formal basis for understanding the modeling language.
6. Encourage the growth of OO tools market.
7. Support higher level development concepts such as collaborations, frameworks, patterns and components.
8. Integrate best practices.

6. USE CASE DIAGRAM



7. IMPLEMENTATION

7.1 MODULES:

1) Upload Fish Dataset:

using this module we will upload dataset to application

2) Run Interpolation, CLAHE & LAB:

using this module we will read all images and then apply interpolation, CLAHE and LAB to process all images and then normalize images and then split dataset into train and test

3) Run Decision Tree:

processed train images will be input to decision tree to trained a model and this model will be applied on TEST images to calculate prediction accuracy and other metrics

4) Run Logistic Regression: processed train images will be input to logistic regression to trained a model and this model will be applied on TEST images to calculate prediction accuracy and other metrics

5) Run Naive Bayes:

processed train images will be input to naïve bayes to trained a model and this model will be applied on TEST images to calculate prediction accuracy and other metrics

6) Run Propose SVM Algorithm:

processed train images will be input to SVM algorithm to trained a model and this model will be applied on TEST images to calculate prediction accuracy and other metrics

7) Comparison Graph:

using this module we will plot accuracy and other metric graphs

8) Predict Fish Status:

using this module we will upload test image and then SVM algorithm will predict whether image contains fresh or infected fish

8. CONCLUSION

We introduce a significant machine learning-based classification model (SVM) to identify infected fishes in this research work. The real-world without augmented dataset (163 infected and 68 fresh) and augmented dataset (785 infected and 320 fresh) are used to train our model is new and novel. We mainly classify fishes into two individual classes: fresh fish and another is infected fish. We appraise our model with various metrics and show the classified outcome with visual interaction from those classification results. Besides developing our classifier, we applied updated image processing techniques like k-means segmentation, cubic spline interpolation, and adaptive histogram equalization for transforming our input image more adaptable to our classifier. We also compare our model results with three classification models and observe that our proposed classifier is the best solution in this case.

This work contributes to bringing out a superior automated fish detection system than the existed systems based on image processing or lower accuracy. We not only depend on the modern image processing technique but also adjoin demandable supervised learning techniques. We prosperously develop the classifier that predicts infected fish with the best accuracy rate than other systems for our real-world novel dataset.

In the future, we stratagem to utilize various Convolutional Neural Networks (CNN) architecture for identifying fish disease more precisely and meticulously. Moreover, we will focus on the implementation of a real-life IoT device using the proposed system. Doing so can be a specific solution for the farmers in aquaculture to identify infected salmon fishes and take proper steps before facing any unexpected loss in their farming. We will work with different fish datasets to make our system more usable in other sectors of aquaculture. We will also concentrate on increasing our existing dataset as salmon fish is one of the demanding elements worldwide.

9. REFERENCES

[1] A. A. M. Abd Hamid, N. and A. Izani. Extended cubic b-spline interpolation method applied to linear two-point boundary value problem. World Academy of Science, 62, 2010.

- [2] T. Acharya. Median computation-based integrated color interpolation and color space conversion methodology from 8-bit bayer pattern rgb color space to 24-bit cie xyz color space, 2002. US Patent 6,366,692.
- [3] A. F. Agarap. An architecture combining convolutional neural network (cnn) and support vector machine (svm) for image classification. arXiv preprint arXiv:1712.03541, 2017.
- [4] A. Ben-Hur and J. Weston. A user's guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer, 2010.
- [5] S. Bianco, F. Gasparini, A. Russo, and R. Schettini. A new method for rgb to xyz transformation based on pattern search optimization. *IEEE Transactions on Consumer Electronics*, 53(3):1020–1028, 2007.
- [6] E. Bisong. Google colab. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, pages 59–64. Springer, 2019.
- [7] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.
- [8] S. A. Burney and H. Tariq. K-means cluster analysis for image segmentation. *International Journal of Computer Applications*, 96(4), 2014.
- [9] M. A. Chandra and S. Bedi. Survey on svm and their application in image classification. *International Journal of Information Technology*, pages 1–11, 2018.
- [10] L. de Oliveira Martins, G. B. Junior, A. C. Silva, A. C. de Paiva, and M. Gattass. Detection of masses in digital mammograms using kmeans and support vector machine. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 8(2):39–50, 2009.
- [11] A. Gaur and S. Yadav. Handwritten hindi character recognition using k-means clustering and svm. In *2015 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services*, pages 65–70. IEEE, 2015.
- [12] M. grapple with Atlantic salmon price rollercoaster. Global aquaculture alliance. <http://www.fao.org/familyfarming/detail/en/c/1263890/>, 2020.

- [13] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.
- [14] H. Hel-Or. Color conversion algorithms. http://cs.haifa.ac.il/hagit/courses/ist/Lectures/Demos/ColorApplet2/t_convert.html. [Last accessed: 2 Nov, 2020].
- [15] M. S. Hitam, E. A. Awalludin, W. N. J. H. W. Yussof, and Z. Bachok. Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In 2013 International conference on computer applications technology (ICCAT), pages 1–5. IEEE, 2013.

ADAPTIVE HIERARCHICAL CYBER ATTACK DETECTION AND LOCALIZATION IN ACTIVE DISTRIBUTION SYSTEMS

Botta Durga Bhavani (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Development of a cyber security strategy for the active distribution systems is challenging due to the inclusion of distributed renewable energy generations. This paper proposes an adaptive hierarchical cyber attack detection and localization framework for distributed active distribution systems via analyzing electrical waveforms. Cyber attack detection is based on a sequential deep learning model, via which even minor cyber attacks can be identified. The two-stage cyber attack localization algorithm first estimates the cyber attack sub-region, and then localize the specified cyber attack within the estimated subregion. We propose a modified spectral clustering-based network partitioning method for the hierarchical cyber attack ‘coarse’ localization. Next, to further narrow down the cyber attack location, a normalized impact score based on waveform statistical metrics is proposed to obtain a ‘fine’ cyber attack location by characterizing different waveform properties. Finally, compared with classical and state-of-art methods, a comprehensive quantitative evaluation with two case studies shows promising estimation results of the proposed framework.

3. INTRODUCTION

CYBER attack localization is important to protect smart distribution grids, but also a challenging task because of the inherent distributed energy resources (DER) and topology complexities [1], [2]. Raw electrical waveforms, signals of electrical networks, together with those in cyber networks provide great potentials in cyber attack detection [3]. For example, devices in power networks must leave clues of their operational status and health (including faults or attacks) information in the raw electrical waveform signals: a cyber-device in fault or under attack will cause unusual energy consumption pattern in power networks [4]; a power electronics or electric machine in fault or under attack may cause unusual harmonics or energy profile in electrical

networks [5]. By analyzing the electrical waveform signals and their root cause, waveform analytics can present utilities with a complete picture of the health and status of their system, both during outages and normal operating conditions. It could also provide a variety of operational benefits to system operators, asset management personnel, and repair crew. Electronic sensors placed on power grids and distribution systems can either measure the electricity properties, such as phasor measurement unit (PMU) sensors [6], [7] or directly record the raw electrical waveform using waveform measurement unit (WMU) [8]– [12], depending on the needed fidelity of monitoring applications. Thanks to developed network connectivity, the streaming monitoring data flow can be obtained and analyzed online and in real-time

The network of the waveform sensors form an Internet of Things (IoT) system [4], [14], where the waveform sensors are viewed as networked IoT sensing devices. Therefore, we can potentially use the information embedded in electrical signals to enable security monitoring, diagnosis, and prognosis in the power networks. The possibility may be well beyond what we can imagine now. It broadly applies to many cyber-physical systems (CPS) and applications, such as power distribution networks, multi-stage manufacturing systems, electric vehicles, and so on [15]–[17]. Cyber attacks towards connected IOT devices trigger anomalies in system statistics, energy consumption, as well as electrical waveforms [4], [14], [18], [19]. Thus, recorded waveform which carries high fidelity current and voltage information should be adequate for cyber attack characterization. Furthermore, the transmission of the high-frequency waveform data is feasible in practice [20]–[22].

Data-driven methods have been widely adopted for event localization in power electronics networks and active distribution systems. Rule-based data-driven analytics [23], signal property-based approach [24], and neural networks (NN) based algorithms, such as autoencoders [25], convolutional neural network (CNN) [26], have been developed. However, NN based algorithms typically require a large amount of training data to capture the sophisticated features, which cannot be fully simulated or acquired from real applications. Thus, combining the rule-based signal processing methods and machine learning methods could lead to a solution tackling the challenging problem using an affordable data size.

There have been numerous works targeting the event and cyber attack localization problem [1], [2], [27]. Dynamic data analytics based localization is always a major branch for the distribution networks [1], DC microgrid [2], islanded microgrid [27]. This paper proposes a new adaptive hierarchical framework for efficient and accurate cyber attack detection and localization by taking advantage of the electrical waveforms (Fig. 1). The proposed approach has a hierarchical architecture that divides the whole network into sub-groups and then locates the cyber attack within one local cluster. Based on a modified unsupervised clustering and an deep learning based anomaly detection method, cyber attacks in the active distribution systems can be adaptively detected and located. The performance of the proposed approach has been tested by multiple cyber attack scenarios in two representative case studies. Our contributions are summarized as follows:

We propose an adaptive hierarchical cyber attack detection and localization framework for active distribution systems with DERs using the electrical waveform;
High fidelity models of DER and cyber attacks are built to analyze the impacts of cyber attacks towards the distribution networks;
Extensive experiments are utilized to evaluate the proposed approach performances with quantitative analytics;

2. EXISTING SYSTEM

Cyber and physical attacks threaten the security of distribution power grids. The emerging renewable energy sources such as photovoltaics (PVs) introduce new potential vulnerabilities. Based on the electric waveform data measured by waveform sensors in the distribution power networks, in this article, an existing system develops a novel high-dimensional data-driven cyber physical attack detection and identification (HCADI) approach.

First, we analyze the cyber and physical attack impacts (including cyber attacks on the solar inverter causing unusual harmonics) on electric waveforms in the distribution power grids. Then, we construct a high-dimensional streaming data feature matrix based on signal analysis of multiple sensors in the network.

Next, we propose a novel mechanism including leverage score-based attack detection and binary matrix factorization-based attack diagnosis. By leveraging the data structure and binary coding, our HCADI approach does not need the training stage for both detection and the root cause

diagnosis, which is needed for machine learning/deep learning-based methods. To the best of our knowledge, it is the first attempt to use raw electrical waveform data to detect and identify the power electronics cyber/physical attacks in distribution power grids with PVs.

Disadvantages

- ❖ The system is not implemented Network Partition based on Modified Spectral Clustering.
- ❖ The system is not implemented Cyber Attack Localization within Sub-regions.

3. PROPOSED SYSTEM

The system proposes an adaptive hierarchical cyber attack detection and localization framework for active distribution systems with DERs using the electrical waveform;

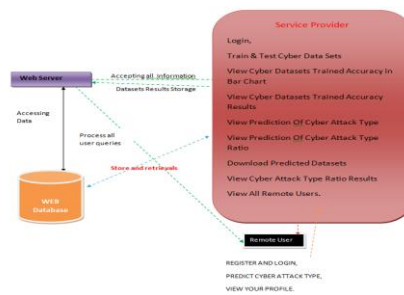
High fidelity models of DER and cyber attacks are built to analyze the impacts of cyber attacks towards the distribution networks;

Extensive experiments are utilized to evaluate the proposed approach performances with quantitative analytics.

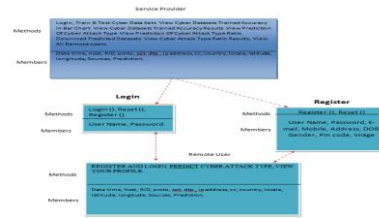
Advantages

- In the proposed system, the cyber attack can be detected based on the deviation of the monitoring metrics from steady-state, which, in our study, is an anomaly detection problem.
- To efficiently locate the cyber attacks, the system proposes to first partition the active distribution systems into several subregions.

4. ARCHITECTURE



5. CLASS DIAGRAM



6. SYSTEM STUDY

6.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

7. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

8. CONCLUSION

In this paper, we proposed an adaptive hierarchical cyber attack localization approach for active distribution systems. Electric waveform signals obtained by WMU sensors are used to capture the abnormal features, which would be otherwise ignored. To improve the efficiency, we propose a modified spectral clustering method to first partition the whole large network into smaller ‘coarse’ sub-regions. Next, the accurate ‘fine’ cyber attack location can be determined by calculating and analyzing Impact Score of each sensor in the potential sub-region. Furthermore, we compare our method with other methods in each step in cyber attack detection, sub-graph clustering, and localization, respectively. The results from two representative distribution grids show that our method shows promising performances.

9. REFERENCES

- [1] I. Džafić, R. A. Jabr, S. Henselmeyer, and T. Đonlagić, “Fault location in distribution networks through graph marking,” *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 1345–1353, 2016.
- [2] R. Bhargav, B. R. Bhalja, and C. P. Gupta, “Novel fault detection and localization algorithm for low voltage dc microgrid,” *IEEE Transactions on Industrial Informatics*, 2019.

- [3] G. Wu, G. Wang, J. Sun, and J. Chen, “Optimal partial feedback attacks in cyber-physical power systems,” *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3919–3926, 2020.
- [4] F. Li, Y. Shi, A. Shinde, J. Ye, and W.-Z. Song, “Enhanced cyberphysical security in internet of things through energy auditing,” *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5224–5231, 2019.
- [5] A. J. Wilson, D. R. Reising, R. W. Hay, R. C. Johnson, A. A. Karrar, and T. D. Loveless, “Automated identification of electrical disturbance waveforms within an operational smart power grid,” *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4380–4389, 2020.
- [6] P. Dutta, A. Esmailian, and M. Kezunovic, “Transmission-line fault analysis using synchronized sampling,” *IEEE transactions on power delivery*, vol. 29, no. 2, pp. 942–950, 2014.
- [7] I. Sadeghkhan, M. E. H. Golshan, A. Mehrizi-Sani, J. M. Guerrero, and A. Ketabi, “Transient monitoring function-based fault detection for inverter-interfaced microgrids,” *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2097–2107, 2016.
- [8] A. F. Bastos, S. Santoso, W. Freitas, and W. Xu, “SynchrowaveformB measurement units and applications,” in *2019 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 2019, pp. 1–5.
- [9] Schweitzer Engineering Laboratories, Pullman, WA, USA., “SEL-T400L Time Domain Line Protection,” <https://selinc.com/products/T400L/>, Last Access: July 31, 2020.
- [10] Candura instruments, Oakville, ON, Canada., “iPSR intelligent Power System Recorder,” <https://www.candura.com/products/ipsr.html>, Last Access: July 31, 2020.



ANALYSIS OF WOMEN SAFETY IN INDIAN CITIES USING MACHINE LEARNING ON TWEETS

Bunga Sanjana Keerthi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West
Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

ABSTRACT: - Women and girls have been experiencing a lot of violence and harassment in public places in various cities starting from stalking and leading to abuse harassment or abuse assault. This research paper basically focuses on the role of social media in promoting the safety of women in Indian cities with special reference to the role of social media websites and applications including Twitter platform Facebook and Instagram. This paper also focuses on how a sense of responsibility on part of Indian society can be developed the common Indian people so that we should focus on the safety of women surrounding them. Tweets on Twitter which usually contains images and text and also written messages and quotes which focus on the safety of women in Indian cities can be used to read a message amongst the Indian Youth Culture and educate people to take strict action and punish those who harass the women. Twitter and other Twitter handles which include hash tag messages that are widely spread across the whole globe sir as a platform for women to express their views about how they feel while we go out for work or travel in a public transport and what is the state of their mind when they are surrounded by unknown men and whether these women feel safe or not?

1. INTRODUCTION

There are certain types of harassment and Violence that are very aggressive including staring and passing comments and these unacceptable practices are usually seen as a normal part of the urban life. There have been several studies that have been conducted in cities across India and women report similar type of sexual harassment and passing off comments by other unknown people. The study that was conducted across most popular Metropolitan cities of India including Delhi, Mumbai and Pune, it was shown that 60 % of the women feel unsafe while going out to work or while travelling in public transport. Women have the right to the city which means that they can go freely

whenever they want whether it be too an Educational Institute, or any other place women want to go. But women feel that they are unsafe in places like malls, shopping malls on their way to their job location because of the several unknown Eyes body shaming and harassing these women point Safety or lack of concrete consequences in the life of women is

the main reason of harassment of girls. There are instances when the harassment of girls was done by their neighbours while they were on the way to school or there was a lack of safety that created a sense of fear in the minds of small girls who throughout their lifetime suffer due to that one instance that happened in their lives where they were



forced to do something unacceptable or was sexually harassed by one of their own neighbor or any other unknown person. Safest cities approach women safety from a perspective of women rights to the affect the city without fear of violence or sexual harassment. Rather than imposing restrictions on women that society usually imposes it is the duty of society to imprecise the need of protection of women and also recognizes that women and girls also have a right same as men have to be safe in the City. Analysis of twitter texts collection also includes the name of people and name of women who stand up against sexual harassment and unethical behaviour of men in Indian cities which make them uncomfortable to walk freely. The data set that was obtained through Twitter about the status of women safety in Indian society was for the processed through machine learning algorithms for the purpose of smoothening the data by removing zero values and using Laplace and porter's theory is to developer method of analyzation of data and remove retweet and redundant data from the data set that is obtained so that a clear and original view of safety status of women in Indian society is obtained.

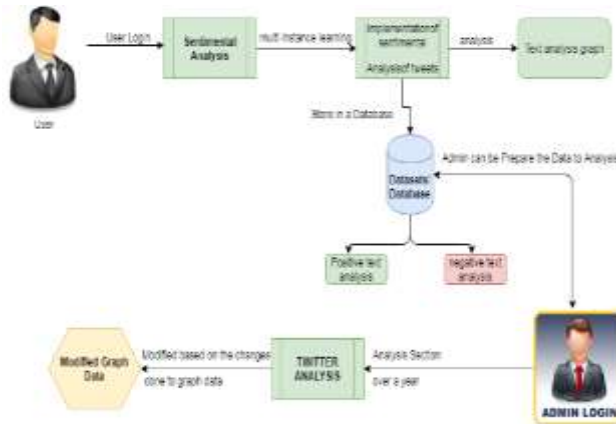
2. LITERATURE REVIEW

People often express their views freely on social media about what they feel about the Indian society and the politicians that claim that Indian cities are safe for women [1]. On social media websites people can freely Express their view point and women can share their experiences where they have faced sexual harassment or where we would have fight back against the sexual

harassment that was imposed on them[2] . The tweets about safety of women and stories of standing up against sexual harassment further motivates other women data on the same social media website or application like Twitter. Other women share these messages and tweets which further motivates other 5 men or 10 women to stand up and raise a voice against people who have made Indian cities and unsafe place for the women. In the recent years a large number of

people have been attracted towards social media platforms like Facebook, Twitter and Instagram point and most of the people are using it to express their emotions and also their opinions about what they think about the Indian cities and Indian society. There are several method of sentiment that can be categorized like machine learning hybrid and lexicon-based learning. [5] Also there are another categorization Janta presented with categories of statistical, knowledge-based and age wise differentiation approaches. It is a common practice to extract the information from the data that is available on social networking through procedures of data extraction, data analysis and data interpretation methods. The accuracy of the Twitter analysis and prediction can be obtained by the use of behavioural analysis on the basis of social networks.

3. ARCHITECTURE



4. FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4.1 ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only

the customized products had to be purchased.

4.2 TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

4.3 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system



meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

5.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

5.2 Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

5.3 Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

6. SYSTEM TEST

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process



descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

7. CONCLUSION

Throughout the research paper we have discussed about various machine learning algorithms that can help us to organize and analyze the huge amount of Twitter data obtained including millions of tweets and text messages shared every day. These machine learning algorithms are very effective and useful when it comes to analyzing of large amount of data including the SPC algorithm and linear algebraic Factor Model approaches which help to further categorize the data into meaningful groups. Support vector machines is yet another form of machine learning algorithm that is very popular in extracting Useful information from the Twitter and get an idea about the status of women safety in Indian cities.

8. REFERENCES



- [1] Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. Mckeown. "Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams." Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009.
- [2] Barbosa, Luciano, and Junlan Feng. "Robust sentiment detection on twitter from biased and noisy data." Proceedings of the 23rd international conference on computational linguistics: posters. Association for Computational Linguistics, 2010.
- [3] Bermingham, Adam, and Alan F. Smeaton. "Classifying sentiment in microblogs: is brevity an advantage?." Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010.
- [4] Gamon, Michael. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.
- [5] Kim, Soo-Min, and Eduard Hovy. "Determining the sentiment of opinions." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.
- [6] Klein, Dan, and Christopher D. Manning. "Accurate unlexicalized parsing." Proceedings of the 41st Annual Meeting on Association for Computational Linguistics- Volume 1. Association for Computational Linguistics, 2003..
- [7] Charniak, Eugene, and Mark Johnson. "Coarse-to-fine n-best parsing and MaxEnt discriminative reranking." Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, 2005.
- [8] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 0975-8887.
- [9] Sahayak, V., Shete, V., & Pathan, A. (2015). Sentiment analysis on twitter data. International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(1), 178-183.
- [10] Mangain, N., Mehta, E., Mittal, A., & Bhatt, G. (2016, March). Sentiment analysis of top colleges in India using Twitter data. In Computational Techniques in Information and Communication Technologies (ICCTICT), 2016 International Conference on (pp. 525-530). IEEE.

A STUDENT ATTENDANCE MANAGEMENT METHOD BASED ON CROWDSENSING IN CLASSROOM ENVIRONMENT

Chalnati Satyavathi Hanumayamma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT - In smart cities, intelligent learning environment is an important application scenario, and class attendance checking is an important measure to urge students to attend on time and ensure the quality of learning. Aiming at the existing problems in class attendance checking, such as low efficiency and easy to cheat, this paper proposes a student attendance management method named AMMoC (Attendance Management Method based on Crowdsensing). AMMoC includes two phases, i.e., the initialization phase and the authentication phase. In the initialization phase, a teacher sends an attendance checking request to the server. After receiving the request, the server sends a request to tell students to submit their location information, and then forms the student location map once the server receives all the response from students. In the authentication phase, the server verifies the truth of the location information by sending requests to several students to count the number of students. The authentication phase includes two modules, i.e., the task assignment module and the attendance verification module. In the task assignment module, AMMoC first finds the optimized sequence of subregions and verifiers by using the Monte Carlo algorithm, and then requires the verifiers to count the number of students in the subregion. Finally, the statistics results will be verified in the attendance verification module. Experiment comparisons and analyses show that AMMoC has the advantages of good anti-cheating performance, fast speed, and little disturbance to class, and is suitable for attendance checking applications in classroom environment.

1. INTRODUCTION

With the popularity of mobile devices, how to build a mobile learning interactive environment has become an important problem during the construction of smart cities. Mobile learning is increasingly becoming an indispensable learning paradigm in modern education systems. Applying the mobile computing technology to the classroom environment (i.e., mobile

education) can solve many problems in traditional class learning systems, e.g., laborious class management, non-timely feedback in teaching effect, and poor communication between teachers and students. Nowadays, mobile education has become one of the hotspots in the modern education field. Class attendance ratio is one crucial indicator for evaluating the quality of a course. Lukkarinen et al. used clustering and regression analysis to study the relationship between college students' class attendance ratio and academic performance [1]. They found that it is positively correlated between attendance ratio and scores, and the high attendance ratio of students will improve the effect of teaching. Besides, absence from class will affect not only the individual scores but also the learning atmosphere of a class [2]. Here fore, attendance has always been an important part of school management. The existing class attendance checking is usually carried out in manual mode, and it can be divided into two forms, i.e., the one without teacher supervision and the one with teacher supervision. During the class attendance checking without teacher supervision, students pass a check-in form in the classroom to complete the attendance checking, but the delivery of the check-in form will not only interfere in the class order, but also cause a certain degree of fake attendance checking [3]. During the class attendance checking with teacher supervision, teachers (or teaching assistants) confirm the attendance of students by roll-calling one by one. This kind of roll-calling method is very inefficient. When the number of students is large, the roll-calling process will take up a lot of teaching time [4]. By analyzing the manual attendance checking process, we found that it is because students need to complete the attendance checking tasks one by one, so that they cannot perform the attendance checking at the same time. Therefore, parallelizing the attendance checking process is the key to improving attendance checking efficiency. A feasible solution is to deploy several RFID (Radio Frequency Identification) readers in a classroom, and students complete their attendance checking by placing their RFID card on the RIFD readers [5]. Although this scheme can greatly improve the attendance checking efficiency, its shortcomings are also obvious. First, the cost of deploying RFID readers in classrooms is very high. Second, RFID readers cannot verify the identity of the cardholder, so we are still not sure whether an imposter is signing in. With the popularization of mobile smart devices, some new techniques are emerging to solve the above-mentioned problems. For example, attendance checking-related applications can be developed on mobile devices, and students only need to complete attendance checking on the applications [6]. This solution greatly reduces the cost of system deployment,

but it still cannot judge whether an imposter is signing in. Students can bring the mobile phones of others into the classroom to complete the attendance checking for them. Aiming at this problem, some scholars proposed that biometric technology could be applied to the attendance checking system, such as fingerprint recognition, facial recognition and voiceprint recognition [7-9]. Facial recognition and voiceprint recognition are more suitable for class attendance checking systems because these biological characteristics can be collected through mobile devices, which can reduce the costs. Although biometric authentication solves the problem of fake attendance checking, it may expose students' privacy and endanger their property safety

In this paper, we propose an intelligent attendance management method named AMMOC. AMMOC need neither deploy additional hardware devices in the classroom, nor collect the biological characteristics of students. AMMOC only needs to install two Android applications on mobile devices of teachers and students respectively, and uses mutual verification between students to complete attendance checking. AMMOC divides the classroom into several sub regions, and assigns students to verify the student number of sub regions. The verification process is classified into a series of crowd sensing tasks [11]. At the beginning of attendance checking, students submit their location information to AMMOC within a time limit. After AMMOC obtains the location information of students, it uses an algorithm based on intelligent search, selects several students to complete the crowd sensing tasks which require to submit the number of students of a specific sub region, etc. AMMOC will analyze the truth of the initial location information based on the results of the crowd sensing tasks submitted by the students. The main contributions of this paper are as follows.

- (1) This paper presents a student attendance management method that combines the active reporting and sampling check of students' location information, which has the advantages of high real-time performance and low disturbance.
- (2) This paper proposes a method which evaluates the value of sub regions based on the remaining number of students, which can accurately select the optimize sub regions for attendance verification.
- (3) This paper proposes a sub region generation method based on certain randomness, which can fully explore the possible sub regions space, and improve the anti-cheating performance of the attendance checking. The rest of this paper is organized as follows. Section 2 introduces the related work in the field of intelligent attendance systems. Section 3 gives the architecture and

functions of AMMOC. Section 4 details the implementation of AMMOC. Section 5 conducts the experiments and analyzes the experiment results. Finally, Section 6 summarizes the work of this paper.

2. EXISTING SYSTEM

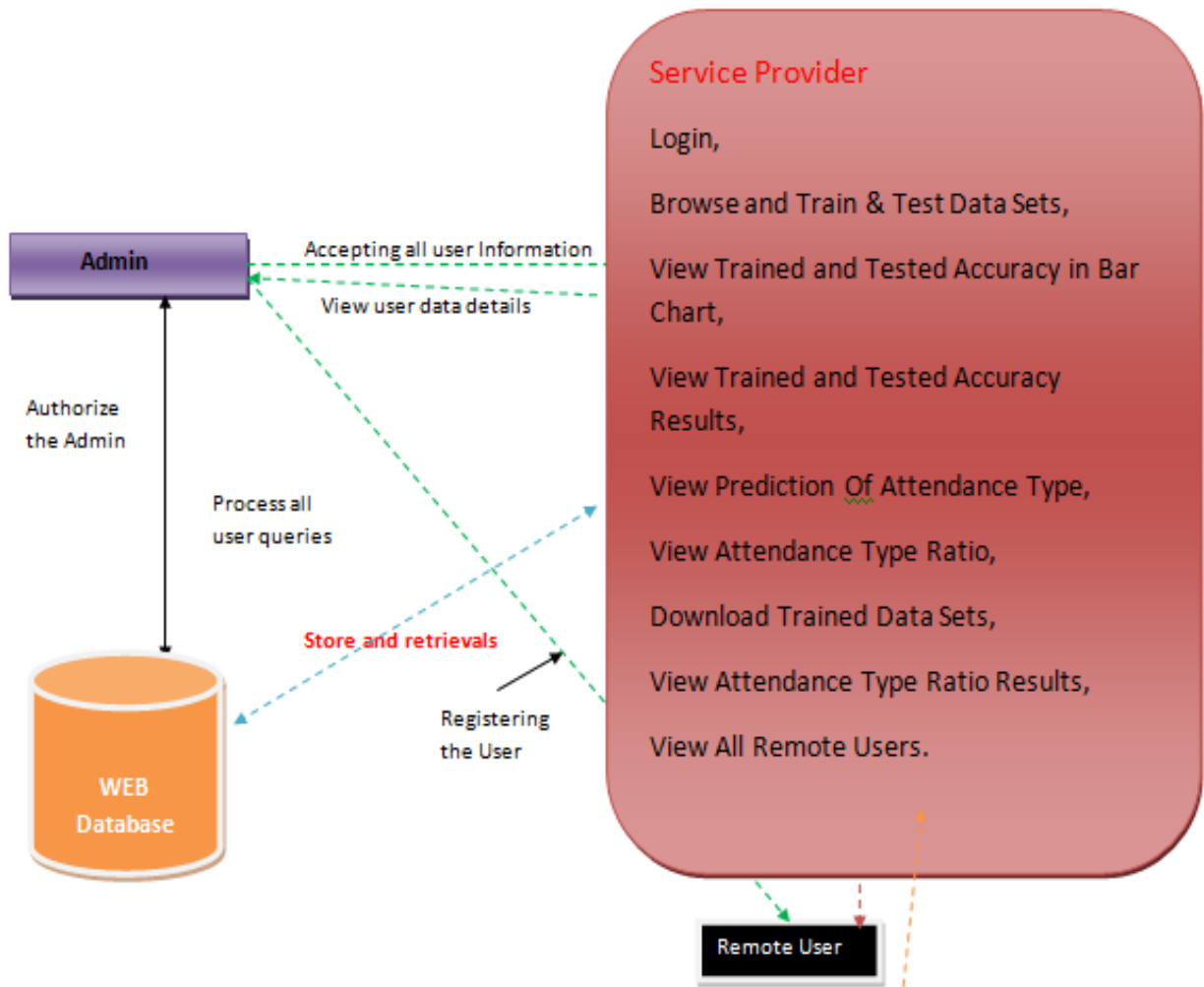
The ID-based attendance checking system usually uses RFID and NFC (Near Field Communication) technology. Rjeib et al. proposed a RFID-based attendance management and information service system named AMS [13]. In AMS, each student's identity information and class schedule are bounded to the RFID tag of the student ID card. All attendance records and student information are stored in the database and displayed on a web application.

Ahmad et al. designed an NFC-based attendance checking system named TouchIn [14]. TouchIn includes two main units, the reader unit and the web server unit. Students can use mobile devices or student ID cards with NFC tags to touch the NFC reader to complete the attendance checking. Jacob et al. integrated the one-time password (OTP) technology into the ID-based attendance checking system [15]. Once the NFC reader detects that a student has entered the classroom, the server will randomly generate a unique one-time password for each student, and send it to the student's mobile device. After receiving the information, the student needs to submit the password through the pre-installed application on the mobile device to complete the attendance checking.

The biometrics-based attendance checking systems usually identify students by fingerprint recognition, face recognition and other biometric technologies. Muchtar et al. developed an attendance checking system based on fingerprint recognition [20]. By using Arduino and Raspberry Pi to manage the fingerprint data centrally, each user can be identified on different fingerprint sensors, which improves the efficiency of the attendance management. Arsenovic et al. proposed a face recognition attendance checking system named FaceTime based on deep learning [21]. Students first submit the identity information of their ID cards, and then FaceTime will call the webcam to collect and recognize their faces. Yang et al. proposed an intelligent attendance checking system based on voiceprint recognition and real-time location positioning [22], and developed a corresponding mobile device application. During attendance checking, the application turns on the device's microphone, and students complete the attendance checking by reading a paragraph of text. They tested this application in an undergraduate computer science

course with about 120 students. On condition that the application meets the required attendance checking accuracy, the attendance checking time can be limited to 5 minutes.

3. ARCHITECTURE

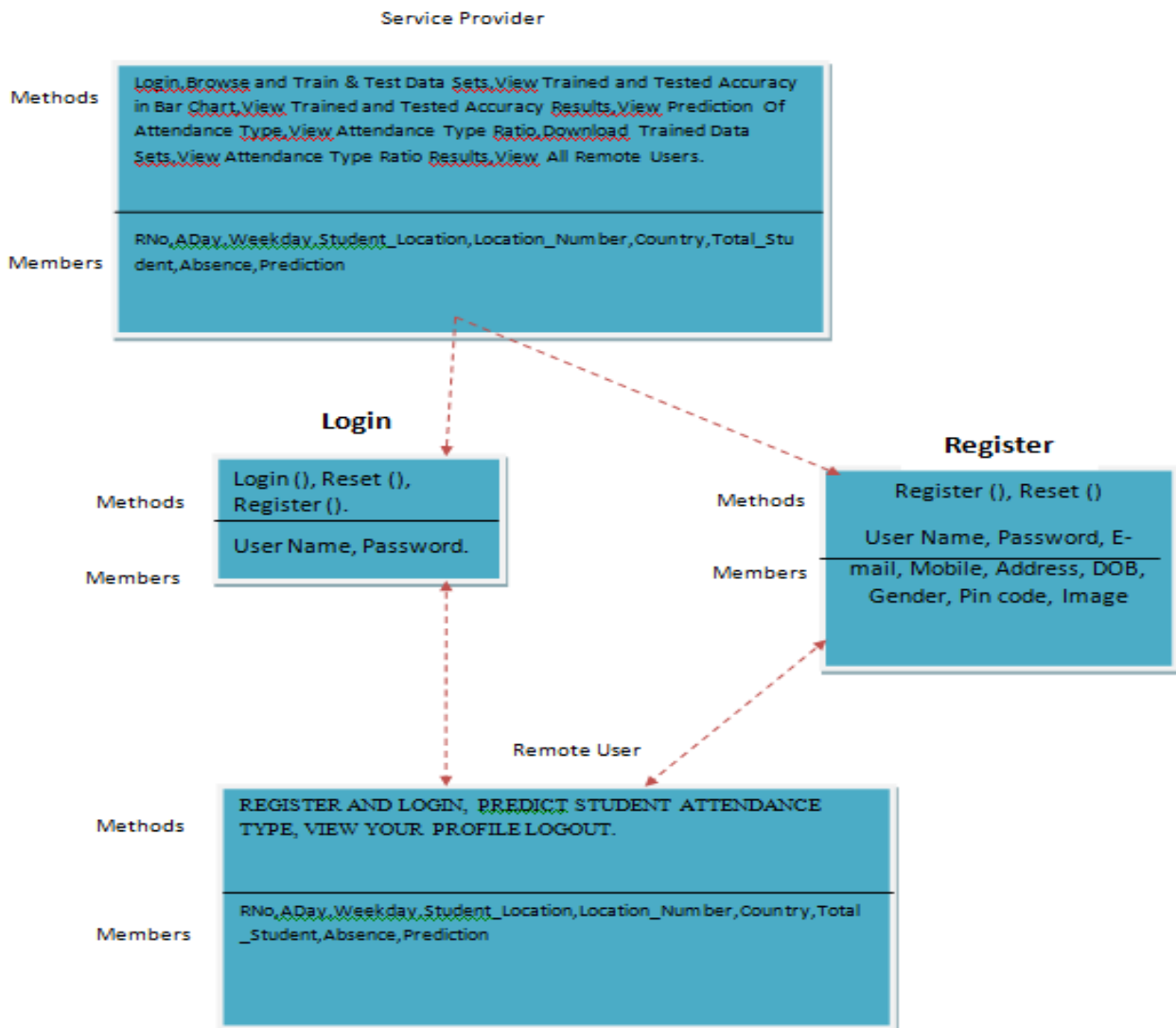


Algorithm Used

Naive Bayes, SVM, Logistic Regression, Decision Tree Classifier, KNeighborsClassifier

REGISTER AND LOGIN,
 PREDICT STUDENT ATTENDANCE TYPE,
 VIEW YOUR PROFILE LOGOUT.

4. CLASS DIAGRAM



5. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**

- **Feasibility Study**
- **Request Approval**

5.1 REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

6. FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The

Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

7.SYSTEM STUDY

7.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

8.SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that theSoftware system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

8.1 TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

9.CONCLUSION

In this paper, we propose an intelligent attendance management method named AMMOC. AMMOC consists of the initialization phase and the authentication phase. In the initialization phase, each student will submit his location information from the student side. In the authentication phase, AMMOC first optimizes the assignment of crowd sensing tasks, and then the MCTS algorithm selects several students to perform student verification. AMMOC will analyze the truth of the submitted locations based on the student number of sub regions submitted by the verifiers. The experiment results show that the AMMOC has the advantages of short attendance checking time and high accuracy. Therefore, it is suitable for AMMOC to perform attendance checking in a classroom environment. In the future work, we plan to shift the attendance checking scene into the virtual one in order to extend the on-site classroom attendance checking to the attendance checking in the online learning environment. We also

hope to achieve continuous non-disturbance attendance checking in order to be suitable for the applications of multiple learning scenarios.

10. REFERENCES

- [1] A. Lukkarinen, P. Koivukangas, and T. Seppälä, "Relationship between class attendance and student performance," *Procedia-Social and Behavioral Sciences*, vol. 228, no. 16, pp. 341-47, Jun. 2016.
- [2] V. Kassarning, A. Bjerre-Nielsen, E. Mones, S. Lehmann, and D. D. Lassen, "Class attendance, peer similarity, and academic performance in a large field study," *PloS one*, vol. 12, no. 11, pp. e0187078, Nov. 2017.
- [3] N. K. Balcoh, M. H. Yousaf, W. Ahmad, and M. I. Baig, "Algorithm for efficient attendance management: Face recognition based approach," *International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 146, Jul. 2012.
- [4] S. C. Kohalli, R. Kulkarni, M. Salimath, M. Hegde, and R. Hongal, "Smart Wireless Attendance System," *International Journal of Computer Sciences and Engineering*, vol. 4, no. 10, pp. 131-137, Sept. 2016.
- [5] M. M. Islam, M. K. Hasan, M. M. Billah, and M. M. Yddin, "Development of smartphone-based student attendance system," in *proceedings of 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, Dhaka, Bangladesh, 2017, pp. 230-233.
- [6] S. M. Čisar, P. Printer, V. Vojnić, V. Tumbas, and P. Čisar, "Smartphone application for tracking students' class attendance," in *proceedings of 2016 IEEE 14th international symposium on intelligent systems and informatics (SISY)*, Subotica, Serbia, 2016, pp. 227-232.
- [7] R. Cappelli, M. Ferrara, and D. Maltoni, "Minutia cylinder-code: A new representation and matching technique for fingerprint recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2128-2141, Mar. 2010.
- [8] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 2106-2112, Jul. 2010.
- [9] A. Boles and P. Rad, "Voice biometrics: Deep learning-based voiceprint authentication system," in *proceedings of 2017 12th System of Systems Engineering Conference (SoSE)*, Waikoloa, HI, USA, 2017, pp. 1-6.

[10] E. Harinda and E. Ntagwirumugara, "Security & privacy implications in the placement of biometric-based ID card for Rwanda Universities," *Journal of Information Security*, vol. 6, no. 02, pp. 93, Jan. 2015.

FARMING MADE EASY USING MACHINE LEARNING

Chalagalla Devi Nikitha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Agriculture is the primary mainstay of the economy in our country. In recent years because of uncertain trends in climate and other fluctuations in the price trends, the price of the crop has varied to a larger level. Farmers remain oblivious of these uncertainties, which spoils the crops and causes massive loss. They are unaware of the crop type which would benefit them most. Due to their limited knowledge of different crop diseases and their specific remedies, crops get damaged. This system is handy, easy-to-use. It provides accurate results in predicting the price of the crop. This framework utilizes Machine Learning's Decision Tree Regression Algorithm to predict crop price. The attributes considered for prediction are rainfall, wholesale price index, month, and year. Consequently, the system gives an advance forecast to the farmers' which grows the speed of profit to them and consequently the country's economy. This system also incorporates other modules like weather forecast, crop recommendation, fertilizer recommendation, and shop, chat portal, and guide are also implemented.

1. INTRODUCTION

India being a rural nation, its economy transcendentally relies upon agricultural yield development and unified agroindustry items. It is currently quickly advancing towards a specialized turn of events. India now is rapidly progressing towards technical development. Smart farming is changing the face of agriculture in India. Technology can provide a solution to most challenges farmers face. It can help them predict weather more accurately, decrease waste, boost output and increase their profit margins. In the status quo, the farmers and the consumers find it difficult in the real world to determine the accurate prices of crops without having prior knowledge of the fluctuating trend prices or weather conditions. Accordingly, innovation will end up being helpful to agriculture. The paper aims to predict crop prices in advance. This work is based on finding proper regional datasets that help us in achieving high accuracy and better performance. Our

system, Agro-Genius, is using Machine Learning to build the Price Predicting Model. In the past few years, a lot of fluctuation in the prices of the crop has been seen. This has increased the rate of crop damage produced each year. The main aim of this prediction system is to ensure that the farmers get a better idea about their yield and deal with the value risk. Weather is also highly unpredictable these days. It also affects the crop production. The proposed system will also forecast the weather helping the farmer make correct decisions regarding field ploughing, field harvesting etc. Similarly, fertilizers play an important role. Fertilizers load the soil with the required nutrients that the crops eliminate from the soil. Crop yields and production will be fundamentally decreased if fertilizers are not used. That is the reason fertilizers are utilized to enhance the soil's supplement stocks with minerals that can be immediately assimilated and utilized by crops. Our system will provide fertilizer consumption based on different crops and provide a portal to buy the fertilizers and seeds from the user's location. They can even get the exact location along with the address of the fertilizer and seed shop. The provided fertilizers will get more profit to the farmers on the growing system suggested crop. It will also show the best suited crop based on cultivation date and month and location details, thereby maximizing the yield.

It will provide multilingual and region specific guide books for the farmers. Any farmer who is new to this field and who wishes to gain information from his ancestors but having the same methods documented will be highly beneficial. We have also provided maps for the farmers to gain knowledge. Our system will provide two different types of maps for the farmer to gain the knowledge about how the land and where they should start their farming. Irrigation maps show the irrigated-non irrigated area over the country. Agriculture land view map will provide an overview of agricultural land present in various states of India and help farmers to analyze the non Agricultural land which can further be improved. Maps make the farmers easy to understand they have to just hover on the state they are thinking of starting their farming and they will get the information about that state and they can decide whether they should change the place or should start farming. If the farmers are new in this field it is the best thing for them as the most important thing in farming is to firstly choose the land and place of farming. Moving in the same direction, our system will incorporate a chat application which helps in information sharing. Often farmers have certain queries which cannot be solved due to their limited knowledge, hence we are building a platform where information can be exchanged. Language can pose as a barrier

to the users. Since the majority of non- English speaking farm workers in India are native Hindispeakers, we anticipate that once these resources are developed they might be translated to other languages as well. Hence, to make the website user friendly, we have provided language translation. Farmers should know about their location, date of cultivation of their crop. Our system is a web application, which is developed based on machine learning concepts. The proposed system applies machine learning and prediction algorithms like Naive Bayes, Decision Trees and K-Nearest Neighbour to identify the most accurate model and then process it. This in turn will help predict the price of the crop.

2 .EXISTING SYSTEM

We have used Python for basic programming in all modules. Flask is used for hosting. Socket Programming is used for a chat application. Chart.js is used for visualizing the maps. JavaScript is used for validation purposes. For Weather Forecast and fertilizer shop location, we have used APIs. Using the self-made dataset and concept of linear regression in machine learning we have implemented a Crop recommendation model so that a farmer can learn about the best suited crop for a particular region. In Fertilizer Recommendation we have used a dataset for predicting which fertilizer should be used for the disease present on crops. Socket programming is used for farmers interaction using provided chat application [3]. Google API is used for providing a multilingual website for ease to read.

3. LITERATURE SURVEY

The following papers focused on predicting crop price using Machine Learning and providing results. In April 2019, the exploration targets foreseeing both the cost and benefit of the given harvest before planting. The preparing datasets so acquired give enough bits of knowledge to foresee the suitable cost and request in the business sectors[1]. The authors have predicted the most profitable crops and its expected price during harvesting time according to the location, by predicting different historical raw datasets using different machine learning algorithms. The work shown by Nishiba [2] is the expected utilization of data mining procedures in foreseeing the harvest yield dependent on the input parameters average rainfall and area of the field. The easy-to-use website page created for anticipating crop yield can be utilized by any client by giving the normal precipitation and region of that place. Different Data Mining techniques are

applied to different datasets. This paper can also include certain modules [11] which can help farmers to make certain decisions based on the harvested area or current trends in the market. The system can be extended by visualizing the crop details in a map with details, which will help farmers to view the nearby district cultivation details. Proposed system can be enhanced by providing a graphical visualization of predicted prices for better understanding. This system is proposed to provide help to the farmers for expecting the best amount for their crops and for predicting the best price for the crops. This also helps the farmers to check previous prices of different commodities. The system can predict crops using Random forest, Polynomial Regression and Decision Tree algorithms. The best crop and its required fertilizers make the farmer more confident about the crop and its yield and also our system will do marketing work [4] by estimating total value of the crop based on current market price. The idea of the system can be extended by adding some extra features to the system like providing a nearby shop location portal for purchasing seeds and fertilizers. These papers aim at predicting the price and forecast through web application and it runs on efficient machine learning algorithms like using an Autoregressive Integrated Moving Average (ARIMA) model, Traditional ARIMA [6], Support Vector Regression Algorithm [8], and technologies having a general easy to use interface to the clients. The training datasets [7] acquired give sufficient bits of knowledge to foreseeing the appropriate price [10] and request in the markets. The results are displayed as web applications in order that poor farmers can access easily. Models can be improved by integrating this with other departments like horticulture, sericulture, and others towards the agricultural development of our country. Different agriculture departments have various problems in the current time. Incorporating them will not only increase the scope but also help the farmers new to this part of the spectrum. Their work may be expanded by building a framework for suggesting agriculture produce and dispersion for farmers. Utilizing this framework, We ought to get the same accuracy indeed when an information autonomous framework is utilized. Further, can be enhanced by making an android application for the same.

4. PROPOSED SYSTEM

In this paper author is using various machine learning algorithms such as Random Forest, Decision Tree and KNN to predict crop prices. All this algorithms get train on Crop Prices

dataset which contains crop details weather details such as Rainfall and below screen showing dataset details with crop name, market name with prices and Rainfall

5. SYSTEM STUDY

5.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**
- ◆ **SYSTEM DESIGN**

6. UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group. The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML. The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems. The UML represents a collection of best engineering practices that have proven successful in the modeling

of large and complex systems. The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

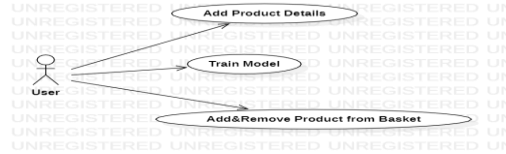
7. GOALS

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

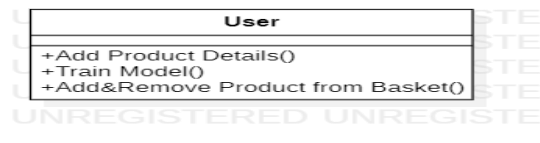
8. USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



8. CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



9. CONCLUSION

This project is undertaken using machine learning and evaluates the performance by using KNN, Naive Bayes, and Decision Tree algorithms. In our proposed model among all the three algorithm Decision Tree gives the better yield prediction as compared to other algorithms. As most extreme sorts of harvests will be secured under this system, farmers may become more acquainted with the yield which may never have been developed. The work exhibited the expected utilization of machine learning methods in foreseeing the harvest cost dependent on the given attributes. The created web application is easy to understand and the testing accuracy is over 90%.

10. REFERENCES

[1] Rachana, Rashmi, Shravani, Shruthi, Seema Kousar, Crop Price Forecasting System Using Supervised Machine Learning Algorithms, International Research Journal of Engineering and Technology (IRJET), Apr 2019

- [2] Nishiba Kabeer, Dr.Loganathan.D, Cowsalya.T, Prediction of Crop Yield Using Data Mining, International Journal of Computer Science and Network, June 2019
- [3] J. Vijayalakshmi, K. PandiMeena, Agriculture TalkBot Using AI, International Journal of Recent Technology and Engineering (IJRTE), July 2019
- [4] Gamage, A., & Kasthurirathna, D. Agro-Genius: Crop Prediction Using Machine Learning, International Journal of Innovative Science and Research Technology, Volume 4, Issue 10, October – 2019
- [5] Vohra Aman, Nitin Pandey, and S. K. Khatri. "Decision Making Support System for Prediction of Prices in Agricultural Commodity." 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE, 2019.
- [6] Nguyen, Huy Vuong, et al. "A smart system for short-term price prediction using time series models." Computers & Electrical Engineering 76 (2019)
- [7] Sangeeta, Shruthi G, Design And Implementation Of Crop Yield Prediction Model In Agriculture, International Journal Of Scientific & Technology Research Volume 8, Issue 01, January 2020
- [8] Rohith R, Vishnu R, Kishore A, Deeban Chakkarawarthy, Crop Price Prediction and Forecasting System using Supervised Machine Learning Algorithms, International Journal of Advanced Research in Computer and Communication Engineering, March 2020
- [9] Naveen Kumar P R, Manikanta K B, Venkatesh B Y, Naveen Kumar R, Amith Mali Patil, Journal of Xi'an University of Architecture & Technology, 2020.
- [10] Kumar, Y. Jeevan Nagendra, et al. "Supervised Machine learning Approach for Crop Yield Prediction in the Agriculture Sector." 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020.
- [11] Pandit Samuel, B.Sahithi , T.Saheli , D.Ramanika , N.Anil Kumar, Crop Price Prediction System using Machine learning Algorithms, Quest Journals Journal of Software Engineering and Simulation, 2020

[12] Rubhi gupta, Review on weather prediction using machine learning, International Journal of Engineering Development and Research



CASHLESS SOCIETY MANAGING PRIVACY AND SECURITY IN THE TECHNOLOGICAL AGE

Cheera Tej venkat (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- A cashless society is an economic state which handles financial transactions not in the form of traditional mediums of currency, such as cash or coins, but by transferring digital data (usually by electronic means, such as credit cards and mobile data) between participating parties. Participants of a cashless society must figure out a way to protect their transaction data, acknowledging the risks of organizations collecting mass amounts of said data, which result in a reduction of personal privacy. Balancing individual privacy with data security is vital in the information age, especially considering the increasing risk of data breaches and exploitation. In order to increase privacy in a cashless society, a few courses of action can be combined to produce a lasting and desirable result for users: A new kind of banking service that assigns randomized numbers to credit cards, the use of blockchain to monitor all transactions from individuals, and a campaign to educate and inform key stakeholders about security and privacy risks to provide the necessary tools and background knowledge to safeguard their own information before interaction with a foreign entity or other third parties (i.e. cybersecurity departments, IT technicians, etc). Blockchain and card number randomization are both susceptible to zero-day errors, bugs, and varied levels of social acceptance. This preliminary research draws on a systems analysis of cashless systems to identify and analyze a set of social and technical solutions to support a robust cashless system that protects users' privacy and maintains the security of the system. The information found and analyzed will be beneficial by exposing weak points in current methods of data integrity and security. Learning about current and future methods of managing privacy and data security in the technological age would be helpful in creating preventative countermeasures. This study provides critical steps to prevent the loss of personal privacy in a cashless system.

1. INTRODUCTION

Systems exist in a constant state of change, and their components must be updated in order to increase, or maintain, the ability to effectively accomplish a task and fulfill a purpose. The currency system is a complex one and requires a thorough analysis of its components, in order to operate at an acceptable level. A cashless system is an

economic state where all transactions are performed without physical means of currency, such as coins or paper bills. For a cashless system, privacy is a crucial component in need of evaluation. Increasing privacy is and will continue to be a necessary undertaking in a cashless society. A majority of users are unaware of what kind of data is being collected about them

and how that data is being used. We thought the whole paper has realized the need for improving privacy, and we propose to do so with a three pronged solution. First, promoting proper education about data collection and privacy will help people realize the need for increased privacy. Second, a randomized credit card system will help prevent unwanted parties from collecting sensitive and personal information about people. Third, block chain will prove to be a powerful authentication tool. Security will be drastically improved through the introduction of these three approaches.

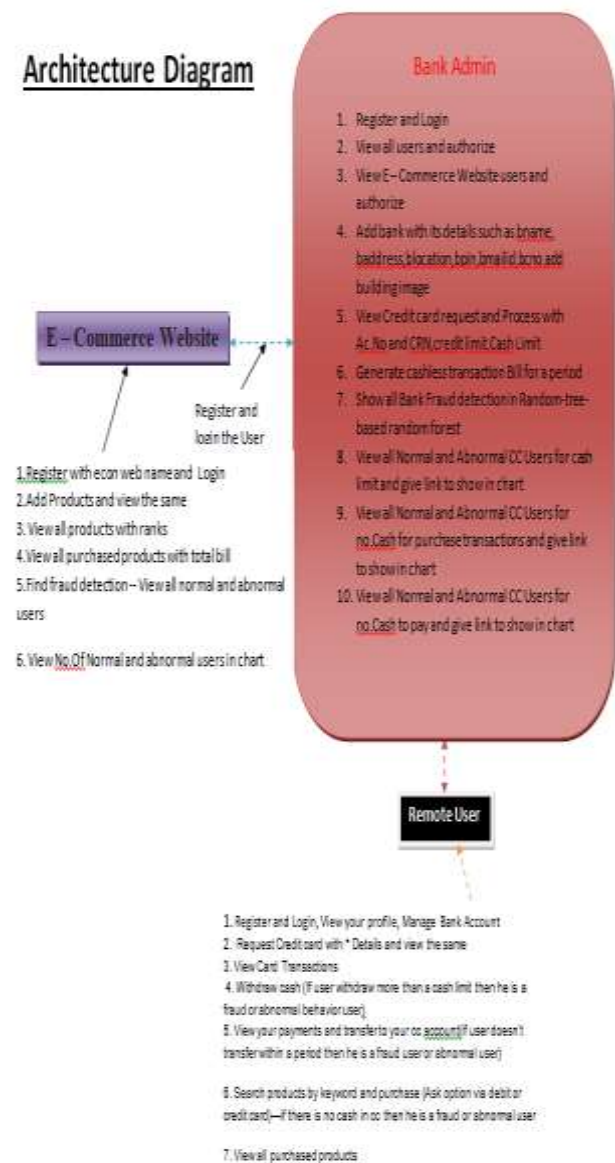
Users will have more knowledge about the systems they are using, hackers will have an exceedingly difficult time fooling the block chain system, and data will be difficult to associate with specific people. A cashless society poses risks for its members because all of their transactions will be tracked online. The members of said cashless society will have to figure out a way to protect their transaction data or risk the threat of organizations collecting mass amounts of data about them, which reduces personal privacy.

2. EXISTING SYSTEM

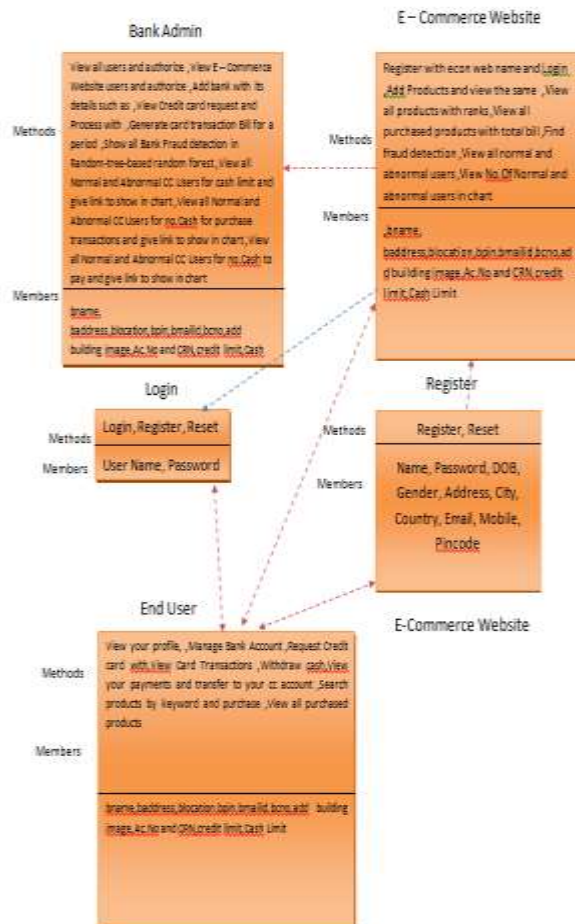
The 2017 report "A Cashless Society - Benefits, Risks and Issues" from a volunteer working party focused on global developments for the topic of a cashless society during the year. This is the 2018 update. It focuses on the trends of that year only. Only countries with substantial events or announcements are talked about, and only new findings are reported for the ones that

featured in the 2017 copy. This copy was collated in the spirit of further developing knowledge, compared to last year. The paper first identifies the driving trends for the year, pointing to structural disruption of the payments ecosystem from conflicting forces. It then reports on regional developments for the topic, with emphasis on India, Kenya, the UK and Australia.

Architecture Diagram



3. CLASS DIAGRAM



The class diagram is the main building block of object oriented modeling. It is used both for general conceptual modeling of the systematic of the application, and for detailed modeling translating the models into programming code. Class diagrams can also be used for modeling. The classes in a class diagram represent both the main objects, interactions in the application and the classes to be programmed.

In the diagram, classes are represented with boxes which contain three parts

- The upper part holds the name of the class

- The middle part contains the attributes of the class
- The bottom part gives the methods or operations the class can take or undertake

In the design of a system, a number of classes are identified and grouped together in a class diagram which helps to determine the static relations between those objects. With detailed modeling, the classes of the conceptual design are often split into a number of subclasses.

4. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only



the customized products had to be purchased.

TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system. SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the

organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

6. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7. CONCLUSION

A cashless society poses risks for its members because data and metadata about their transactions are being collected and used. The members of said cashless society will have to figure out a way to protect their data in order to increase their privacy. Our group has found the idea of a cashless society to involve many systemic complexities. Within the complex system, opportunities arise to implement solutions to privacy and security problems. The various actors in said system have different desires and will respond in unique ways to changes made. Sometimes the best solution to a problem is the culmination of multiple approaches. Spreading information to the



general public helps people learn about the systems they are using and allows for them to make informed decisions. Blockchain helps promote privacy and security through its authentication process. Randomized credit cards help users keep their account numbers private. All three approaches are effective ways of adapting to a dynamic currency system.

8. REFERENCES

- [1] "Bitcoin - Open Source P2P Money." n.d. Accessed December 12, 2019. <https://bitcoin.org/en/>.
- [2] Wolters, Timothy. "'Carry Your Credit in Your Pocket': The Early History of the Credit Card at Bank of America and Chase Manhattan." *Enterprise & Society* 1.2 (2000): 315-54. Print.
- [3] Mercer, Christina. n.d. "History of PayPal: 1998 to Now." *Techworld*. Accessed December 12, 2019. <https://www.techworld.com/picturegallery/business/history-of-paypal-1998-now-3630386/>.
- [4] Meadows, Donella H., and Diana Wright. *Thinking in Systems: a Primer*. Chelsea Green Publishing, 2015.
- [5] Andrew Ferguson, *The rise of big data policing: surveillance, race, and the future of law enforcement*, New York; New York University Press, 2017.
- [6] "The Rise of Big Data Policing — TechCrunch." n.d. Accessed February 5, 2020. <https://techcrunch.com/2017/10/22/the-rise-of-bigdata-policing/>.
- [7] Symanovich, Steve. "What Is a VPN?" Official Site, us.norton.com/internetsecurity-privacy-what-is-a-vpn.html.
- [8] Swan, M. (2015). *Blockchain: Blueprint for a New Economy*. Sebastopol, CA: O'Reilly Media, Inc.
- [9] 2019 Data Breaches - Identity Theft Resource Center. (2020). Retrieved 27 March 2020, from <https://www.idtheftcenter.org/2019-data-breaches/>
- [10] "Leverage Points: Places To Intervene In A System." *The Academy for Systems Change*. N. p., 2020. Web. 3 Feb. 2020.
- [11] "What's New In The 2019 Cost Of A Data Breach Report." *Security Intelligence*. N. p., 2020. Web. 6 Feb. 2020.
- [12] Arthur, W. (2018, March 23). *Lawsuits may be key to tighter US data privacy rules*. Retrieved March 26, 2020, from <https://dailybrief.oxan.com/Analysis/DB230635/Lawsuits-may-bekey-to-tighter-US-data-privacy-rules>
- [13] Arthur, W. (2018, March 23). *Lawsuits may be key to tighter US data privacy rules*. Retrieved March 26, 2020, from <https://dailybrief.oxan.com/Analysis/DB230635/Lawsuits-may-bekey-to-tighter-US-data-privacy-rules>
- [14] "Leverage Points: Places to Intervene in a System." *The Academy for Systems Change*, donellameadows.org/archives/leverage-points-places-to-intervene-in-a-system/.
- [15] Singh, Simon. *The Code Book: the Secrets behind Codebreaking*. Ember, 2016.
- [16] Marwick, Alice E. "How Your Data Are Being Deeply Mined." *The New York Review of Books*, www.nybooks.com/articles/2014/01/09/how-your-data-are-being-deeply-mined/.



IJARST

International Journal For Advanced Research In Science & Technology

A peer reviewed international journal

ISSN: 2457-0362

www.ijarst.in

CROP RECOMMENDATION USING RANDOM FOREST ML ALGORITHM

Chelluboyina Latha Nageswari (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari
District, Andhra Pradesh, India, 534202.

ABSTRACT-: Agriculture and its allied sectors are undoubtedly the largest providers of livelihoods in rural India. The agriculture sector is also a significant contributor factor to the country's Gross Domestic Product (GDP). Blessing to the country is the overwhelming size of the agricultural sector. However, regrettable is the yield per hectare of crops in comparison to international standards. This is one of the possible causes for a higher suicide rate among marginal farmers in India. This paper proposes a viable and user-friendly yield prediction system for the farmers. The proposed system provides connectivity to farmers via a mobile application. GPS helps to identify the user location. The user provides the area & soil type as input. Machine learning algorithms allow choosing the most profitable crop list or predicting the crop yield for a user-selected crop. To predict the crop yield, selected Machine Learning algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), Multivariate Linear Regression (MLR), and K-Nearest Neighbour (KNN) are used. Among them, the Random Forest showed the best results with 95% accuracy. Additionally, the system also suggests the best time to use the fertilizers to boost up the yield.

1. INTRODUCTION

Agriculture has an extensive history in India. Recently, India is ranked second in the farm output worldwide [15]. Agriculture-related industries such as forestry and fisheries contributed for 16.6% of 2009 GDP and around 50% of the total workforce. Agriculture's monetary contribution to India's GDP is decreasing [1]. The crop yield is the significant factor contributing in agricultural monetary. The crop yield depends on multiple factors such as climatic, geographic, organic, and financial elements [6]. It is difficult for farmers to decide when and which crops to plant because of fluctuating market prices [7]. Citing to Wikipedia figures India's suicide rate ranges from 1.4-1.8% per 100,000 populations, over the last 10 years [15]. Farmers are unaware of which crop to grow, and what is the right time and

place to start due to uncertainty in climatic conditions. The usage of various fertilizers is also uncertain due to changes in seasonal climatic conditions and basic assets such as soil, water, and air. In this scenario, the crop yield rate is steadily declining . The solution to the problem is to provide a smart user-friendly recommender system to the farmers.

The crop yield prediction is a significant problem in the agriculture sector . Every farmer tries to know crop yield and whether it meets their expectations, thereby evaluating the previous experience of the farmer on the specific crop predict the yield .Agriculture yields rely primarily on weather conditions, pests, and preparation of harvesting operations. Accurate information on crop history is critical for making decisions on agriculture risk management In this paper, we have proposed a model that addresses these issues. The novelty of the proposed system is to guide the farmers to maximize the crop yield as well as suggest the most profitable crop for the specific region. The proposed model provides crop selection based on economic and environmental conditions, and benefit to maximize the crop yield that will subsequently help to meet the increasing demand for the country's food supplies [8]. The proposed model predicts the crop yield by studying factors such as rainfall, temperature, area, season, soil type etc. The system also helps to determine the best time to use fertilizers. The existing system which recommends crop yield is either hardware-based being costly to maintain, or not easily accessible. The proposed system suggests a mobile-based application that precisely predicts the most profitable crop by predicting the crop yield. The use of GPS helps to identify the user location. The user provides an area under cultivation and soil type as inputs. According to the requirement, the model predicts the crop yield for a specific crop. The model also recommends the most profitable crop and suggests the right time to use the fertilizers. The major contributions of the paper are enlisted below,

1. Prediction of the crop yield for specific regions by executing various Machine Learning algorithms, with a comparison of error rate and accuracy.
2. A user-friendly mobile application to recommend the most profitable crop.
3. A GPS based location identifier to retrieve the rainfall estimation at the given area.
4. A recommender system to suggest the right time for using fertilizers.

The organization of the rest of the paper is as follows. Section II discusses the background work of researchers in the field of agriculture and yield prediction. Section III presents the proposed model for yield prediction and recommends which crop for cultivation.

2. LITERATURE SUREVY

TITTLE : BIoT: Blockchain-based IoT for Agriculture

AUTHORS : Umamaheswari S, Sreeram S, Kritika N, Prasanth DJ

ABSTRACT : Blockchain's most basic promise for the agriculture industry is that it removes the need for third parties otherwise required to ensure trust within buyer-seller relationships, or for that matter any source-destination relationship. In an environment enabled by blockchain technology, transactions become peer-to-peer with no use for intermediaries. Apart from providing the means to transact peer-to-peer, blockchain can create 'smart contracts' that execute the terms of any agreement when specified conditions are met. Every time value changes hands, whether physical products, services or money, the transaction can be documented, creating a permanent history of the product or transaction, from source to ultimate destination. Blockchain can be of great help in this sector. A transparent and trusted system can be built by putting all the information about agricultural events on a blockchain. Farmers can also get instant data related to the seed quality, climate environment related data, payments, soil moisture, demand and sale price, etc. all on a single platform. The intent of this project is to store the sensor data in a blockchain and build a smart contract deployed in the Ethereum blockchain to facilitate buying and selling of crops and land.

TITTLE : Analysis of growth and instability in the area, production, yield, and price of rice in India

AUTHORS : Jain A

ABSTRACT : Agricultural growth with stability has been a matter of concern in India. This paper analyses 41 years data (1970-71 to 2011-12) on area, production and yield under paddy to understand the question of instability in rice production in India. The analysis shows that at all India level compound annual growth rate of area, production and yield of rice were positive but it had been declining gradually over the periods. In the recent decade (2000-01 to 2011-12) there is increase in instability at all India level in area, production and yield of rice. The possible reasons for increase in instability were low percentage of irrigated area to total cropped area, decline in use of seeds and manure and other inputs necessary for agriculture. In the post reform period (1990-91 to 2016-17) the instability has increased in case of

wholesale price of paddy across various states while instability has declined in case of farm harvest price of paddy

TITTLE : A model for prediction of crop yield

AUTHORS : Manjula E, Djodiltachoumy S

ABSTRACT : Data Mining is emerging research field in crop yield analysis. Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region i.e. district of Tamil Nadu in India. The experimental results shows that the proposed work efficiently predict the crop yield production.

3.EXISTING SYSTEM

The existing system which recommends crop yield is either hardware-based being costly to maintain, or not easily accessible. The proposed system suggests a mobile-based application that precisely predicts the most profitable crop by predicting the crop yield. The use of GPS helps to identify the user location. The user provides an area under cultivation and soil type as inputs. According to the requirement, the model predicts the crop yield for a specific crop. The model also recommends the most profitable crop and suggests the right time to use the fertilizers.

DISADVANTAGES OF EXISTING SYSTEM :

- 1) Less accuracy
- 2)low Efficiency

4.PROPOSED SYSTEM

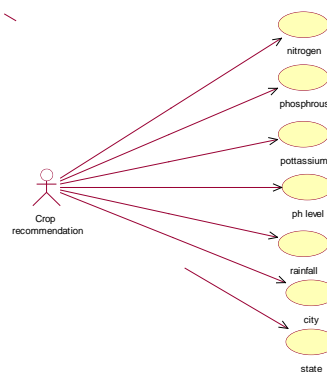
Supervised learning approach is used to implement crop yield prediction system. Established the correlation between multiple attributes selected from the historical which helps the system to increase the crop yield [19]. Rainfall and temperature are two factors which influence the crop yield. Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) algorithms applied on these time series data to enhance the accuracy [20]. ARMA (Auto Regressive Moving Average), SARIMA (Seasonal Auto Regressive Integrated Moving Average) and ARMAX (ARMA with exogenous variables) methods are used to predict the temperature and rainfall using historical data. The best model among them is used in the crop yield prediction system implemented with fuzzy logic. Cloud cover and evapotranspiration are exogenous variables used in the proposed system .

ADVANTAGES OF PROPOSED SYSTEM

- 1) High accuracy
- 2) High efficiency

5. USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



6 .CONCLUSION

This paper highlighted the limitations of current systems and their practical usage on yield prediction. Then walks through a viable yield prediction system to the farmers, a proposed system provides connectivity to farmers via a mobile application. The mobile application includes multiple features that users can leverage for the selection of a crop. The inbuilt predictor system helps the farmers to predict the yield of a given crop. The inbuilt recommender system allows a user exploration of the possible crops and their yield to take more educated decisions. For yield to accuracy, various machine learning algorithms such as Random Forest, ANN, SVM, MLR, and KNN were implemented and tested on the given datasets from the Maharashtra and Karnataka states. The various algorithms are compared with their accuracy. The results obtained indicate that Random Forest Regression is the best among the set of standard algorithms used on the given datasets with an accuracy of 95%. The proposed model also explored the timing of applying fertilizers and recommends appropriate duration.

The future work will be focused on updating the datasets from time to time to produce accurate predictions, and the processes can be automated. Another functionality to be implemented is to provide the correct type of fertilizer for the given crop and location. To implement this thorough study of available fertilizers and their relationship with soil and climate needs to be done. An analysis of available statistical data needs to be done.

7. REFERENCES

- [1] Umamaheswari S, Sreeram S, Kritika N, Prasanth DJ, “BIoT: Blockchain-based IoT for Agriculture”, 11th International Conference on Advanced Computing (ICoAC), 2019 Dec 18 (pp. 324-327). IEEE.
- [2] Jain A. “Analysis of growth and instability in the area, production, yield, and price of rice in India”, Journal of Social Change and Development, 2018;2:46-66
- [3] Manjula E, Djodiltachoumy S, “A model for prediction of crop yield” International Journal of Computational Intelligence and Informatics, 2017 Mar;6(4):2349-6363.
- [4] Sagar BM, Cauvery NK., “Agriculture Data Analytics in Crop Yield Estimation: A Critical Review”, Indonesian Journal of Electrical Engineering and Computer Science, 2018 Dec;12(3):1087-93.
- [5] Wolfert S, Ge L, Verdouw C, Bogaardt MJ, “Big data in smart farming– a review. Agricultural Systems”, 2017 May 1;153:69-80.

[6] Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, Godfray HC, Herrero M, Howitt RE, Janssen S, Keating BA, “Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. *Agricultural systems*”, 2017 Jul 1;155:269-88.

[7] Johnson LK, Bloom JD, Dunning RD, Gunter CC, Boyette MD, Creamer NG, “Farmer harvest decisions and vegetable loss in primary production. *Agricultural Systems*”, 2019 Nov 1;176:102672.

[8] Kumar R, Singh MP, Kumar P, Singh JP, “Crop Selection Method to maximize crop yield rate using a machine learning technique”, *International conference on smart technologies and management for computing, communication, controls, energy, and materials (ICSTM)*, 2015 May 6 (pp. 138-145). IEEE.

[9] Sriram Rakshith.K, Dr.Deepak.G, Rajesh M, Sudharshan K S, Vasanth S, Harish Kumar N, “A Survey on Crop Prediction using Machine Learning Approach”, In *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, April 2019, pp(3231- 3234)

[10] Veenadhari S, Misra B, Singh CD, “Machine learning approach for forecasting crop yield based on climatic parameters”, In *2014 International Conference on Computer Communication and Informatics*, 2014 Jan 3 (pp. 1-5). IEEE.

FLOOD PREDICTION USING MACHINE LEARNING MODELS

Chinthapalli Venkata Ramana, (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract:- Floods are among the most destructive natural disasters, which are highly complex to model. The research on the advancement of flood prediction models contributed to risk reduction, policy suggestion, minimization of the loss of human life, and reduction the property damage associated with floods. To mimic the complex mathematical expressions of physical processes of floods, during the past two decades, machine learning (ML) methods contributed highly in the advancement of prediction systems providing better performance and cost-effective solutions. Due to the vast benefits and potential of ML, its popularity dramatically increased among hydrologists. Researchers through introducing novel ML methods and hybridizing of the existing ones aim at discovering more accurate and efficient prediction models. The main contribution of this paper is to demonstrate the state of the art of ML models in flood prediction and to give insight into the most suitable models. In this paper, the literature where ML models were benchmarked through a qualitative analysis of robustness, accuracy, effectiveness, and speed are particularly investigated to provide an extensive overview on the various ML algorithms used in the field. The performance comparison of ML models presents an in-depth understanding of the different techniques within the framework of a comprehensive evaluation and discussion. As a result, this paper introduces the most promising prediction methods for both long-term and short-term floods. Furthermore, the major trends in improving the quality of the flood prediction models are investigated. Among them, hybridization, data decomposition, algorithm

1. INTRODUCTION

Among the natural disasters, floods are the most destructive, causing massive damage to human life, infrastructure, agriculture, and the socioeconomic system. Governments, therefore, are under pressure to develop reliable and accurate maps of flood risk areas and further plan for sustainable flood risk management focusing on prevention, protection, and preparedness. Flood prediction models are of significant importance for hazard assessment and extreme event management. Robust and accurate prediction contribute highly to water resource management strategies, policy suggestions and analysis, and further evacuation modeling. Thus, the importance of advanced systems for short-term and long-term prediction for flood and other hydrological events is strongly emphasized to alleviate damage. However, the prediction of flood lead time and occurrence location is fundamentally complex due to the dynamic nature of climate condition. Therefore, today's major flood prediction models are mainly data-specific and involve various simplified assumptions. Thus, to mimic the complex mathematical expressions of physical processes and basin behavior, such models benefit from specific techniques e.g., event-driven, empirical black box, lumped and distributed, stochastic, deterministic, continuous, and hybrids. Physically based models were long used to predict hydrological events, such as storm, rainfall/runoff, shallow water condition, hydraulic models of flow, and further global circulation phenomena, including the coupled effects of atmosphere, ocean, and floods. Although physical models showed great capabilities for predicting a diverse predictions [40,41]. In comparison to traditional statistical models, ML models were used for prediction with greater accuracy [42]. Ortiz-García et al. [43] described how ML techniques could efficiently model complex hydrological systems such as floods. Many ML algorithms, e.g., artificial neural networks (ANNs) [44], neuro-fuzzy [45,46], support vector machine (SVM) [47], and support vector regression (SVR) [48,49], were reported as effective for both short-term and long-term flood forecast. In addition, it was shown that the performance of ML could be improved through hybridization with other ML methods, soft computing techniques, numerical simulations, and/or physical models. Such applications provided more robust and efficient models that can effectively learn complex flood systems in an adaptive manner. Although the literature includes numerous evaluation performance analyses of individual ML models [49–52], there is no definite conclusion reported with regards to which models function better in certain applications. In fact, the literature includes only a limited number of surveys on specific ML methods in specific hydrology fields [53–55]. Consequently, there is a research gap for a comprehensive literature

review in the general applications of ML in all flood resource variables from the perspective of ML modeling and data-driven prediction systems.

2. SYSTEM DESIGN

2.1 UML DIAGRAMS:

UML represents Unified Modeling Language. UML is an institutionalized universally useful showing dialect in the subject of article situated programming designing. The fashionable is overseen, and become made by way of, the Object Management Group.

The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components: a Meta-show and documentation. Later on, a few type of method or system can also likewise be brought to; or related with, UML.

The Unified Modeling Language is a popular dialect for indicating, Visualization, Constructing and archiving the curios of programming framework, and for business demonstrating and different non-programming frameworks.

The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks.

The UML is a essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

GOALS:

The Primary goals inside the plan of the UML are as in step with the subsequent:

1. Provide clients a prepared to-utilize, expressive visual showing Language on the way to create and change massive models.
2. Provide extendibility and specialization units to make bigger the middle ideas.
3. Be free of specific programming dialects and advancement manner.

4. Provide a proper cause for understanding the displaying dialect.
5. Encourage the improvement of OO gadgets exhibit.
6. Support large amount advancement thoughts, for example, joint efforts, systems, examples and components.
7. Integrate widespread procedures.

2.2 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



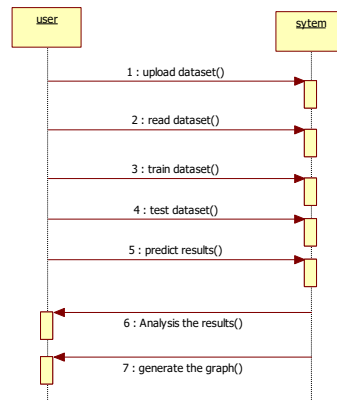
3. CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



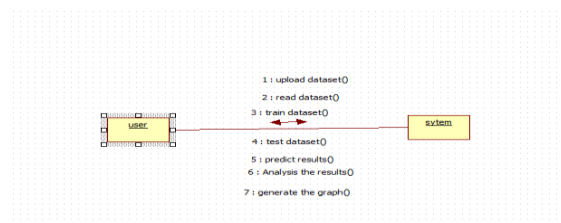
4. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



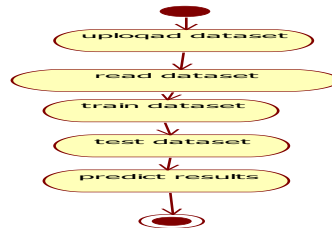
5. COLLABORATION DIAGRAM:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization where as the collaboration diagram shows the object organization.



6. ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



7. DEPLOYMENT DIAGRAM:

Deployment diagram represents the deployment view of a system. It is related to the component diagram. Because the components are deployed using the deployment diagrams. A deployment diagram consists of nodes. Nodes are nothing but physical hardware's used to deploy the application.

8. SYSTEM STUDY

8.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

9. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

10. Conclusions

The current state of ML modeling for flood prediction is quite young and in the early stage of advancement. This paper presents an overview of machine learning models used in flood prediction, and develops a classification scheme to analyze the existing literature. The survey represents the performance analysis and investigation of more than 6000 articles. Among them, we identified 180 original and influential articles where the performance and accuracy of at least two machine learning models were compared. To do so, the prediction models were classified into two categories according to lead time, and further divided into categories of hybrid and single methods. The state of the art of these classes was discussed and analyzed in detail, considering the performance comparison of the methods available in the literature. The performance of the methods was evaluated in terms of R^2 and RMSE, in addition to the generalization ability, robustness, computation cost, and speed. Despite the promising results already reported in implementing the most popular machine learning methods, e.g., ANNs, SVM, SVR, ANFIS, WNN, and DTs, there was significant research and experimentation for further improvement and advancement. In this context, there were four major trends reported in the literature for improving the quality of prediction. The first was novel hybridization, either through the integration of two or more machine learning methods or the integration of a machine learning method(s) with more conventional means, and/or soft computing. The second was the use of data decomposition techniques for the purpose of improving the quality of the dataset, which highly contributed in improving the accuracy of prediction. The third was the use of an ensemble of methods, which dramatically increased the generalization ability of the models and decreased the uncertainty of prediction. The fourth was the use of add-on optimizer algorithms to improve the quality of machine learning algorithms, e.g., for better tuning the ANNs to reach optimal neuronal architectures. It is expected that, through these four key technologies, flood

prediction will witness significant improvements for both short-term and long-term predictions. Surely, the advancement of these novel ML methods depends highly on the proper usage of soft computing techniques in designing novel learning algorithms. This fact was discussed in the paper, and the soft computing techniques were introduced as the main contributors in developing hybrid ML methods of the future.

11. REFERENCES

1. Danso-Amoako, E.; Scholz, M.; Kalimeris, N.; Yang, Q.; Shao, J. Predicting dam failure risk for sustainable flood retention basins: A generic case study for the wider greater manchester area. *Comput. Environ. Urban Syst.* 2012, *36*, 423–433.
2. Xie, K.; Ozbay, K.; Zhu, Y.; Yang, H. Evacuation zone modeling under climate change: A data-driven method. *J. Infrastruct. Syst.* 2017, *23*, 04017013.
3. Pitt, M. *Learning Lessons from the 2007 Floods*; Cabinet Office: London, UK, 2008.
4. Lohani, A.K.; Goel, N.; Bhatia, K. Improving real time flood forecasting using fuzzy inference system. *J. Hydrol.* 2014, *509*, 25–41.
5. Mosavi, A.; Bathla, Y.; Varkonyi-Koczy, A. Predicting the Future Using Web Knowledge: State of the Art Survey. In *Recent Advances in Technology Research and Education*; Springer: Cham, Switzerland, 2017; pp. 341–349.
6. Zhao, M.; Hendon, H.H. Representation and prediction of the indian ocean dipole in the poama seasonal forecast model. *Q. J. R. Meteorol. Soc.* 2009, *135*, 337–352.
7. Borah, D.K. Hydrologic procedures of storm event watershed models: A comprehensive review and comparison. *Hydrol. Process.* 2011, *25*, 3472–3489.
8. Costabile, P.; Costanzo, C.; Macchione, F. A storm event watershed model for surface runoff based on 2D fully dynamic wave equations. *Hydrol. Process.* 2013, *27*, 554–569.
9. Cea, L.; Garrido, M.; Puertas, J. Experimental validation of two-dimensional depth-averaged models for forecasting rainfall–runoff from precipitation data in urban areas. *J. Hydrol.* 2010, *382*, 88–102.

FACIAL EMOTION RECOGNITION SYSTEM THROUGH MACHINE LEARNING APPROACH

Chippada Naga Raja Lakshmi Aishwarya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract-The emotions, set in simple words are what people feel. Emotional aspects have huge impact on Social intelligence like communication understanding, decision making and helps in understanding behavioral aspect of human. Human faces provide various information about emotions. As per psychological researcher, a person expresses his emotions less by verbal talk and more by non-verbal body posture and gestures. Emotion recognition or Affective Computing (AC) being the AI related area imparts intelligence to computers in recognizing human emotions. Emotion recognition is proved a popular research area topic in few decades. The aim of this paper is to report an illustrative and comprehensive study of most popular emotion recognition methods, which are generally used in emotion recognition problems. We are motivated by the lack of detailed study of all possible technique's implementations in available literature. This paper provides an up-to-date comprehensive survey of techniques available for emotion recognition.

Keywords- emotions, images, emotion recognition, facial image, human computer interaction, facial emotion recognition.

1. INTRODUCTION

Emotions entail different components, such as subjective experience, cognitive processes, expressive behavior, psychophysiological changes, and behavior. These various components of emotion are categorized in a different way depending on the academic discipline. In psychology and philosophy, emotion includes a subjective, conscious experience characterized by psychophysiological expressions, biological reactions, and mental states. The research on emotion has increased significantly greater than the past two decades. There are many fields contributing that include psychology, neuroscience, endocrinology, medicine, history, sociology, and computer science.

There are abundant theories that attempt to explicate the origin, experience, and function of emotions and have fostered more intense research on this topic. Current areas of research in the concept of emotion include the development of materials that motivate and elicit emotion [1]. Charles Darwin's (1872/1965) book "The Expression of the Emotions in Man and Animals" has been highly important for research on emotions. This book was intended to counteract the claim by Sir Charles Bell (1844), that certain muscles were created so as to give humans the ability to express their feelings. Darwin's basic message was that emotion expressions are evolved and adaptive. For Darwin,



emotional expressions not only originated as part of an emotion process but also had an important communicative function [2]. The cross-cultural studies conducted by Ekman and his collaborators and by Izard strongly suggested universality in interpreting facial expressions of emotion. These findings countered customary ideas of cultural relativism, and suggested that the study of facial expression is relevant to central questions regarding human nature. Then, researcher developed measures of facial emotion recognition, which some emotion researchers used to measure facial activity itself directly, rather than studying the observers' judgments of the emotions they saw in an expression. Whereas formerly facial activity were measured via electromyography, it is far more invasive and less precise than scoring systems measuring the changes in the appearance of the face. The purpose of emotion recognition systems is the appliance of emotion related knowledge in such a way that human computer communication will be enhanced and furthermore the user's experience will become more satisfying. By enabling computers to sense the emotional state of the user and react accordingly, this communication can be renovated to a satisfying one. Refining the communication with computers is not the only application of emotion recognition. There can be specialized systems that can be developed and can be used for even more serious problems like in various medical applications aggression detection, stress detection, autistic disorder, as per

gersyndrome, hepatolenticular degeneration, frustration detection.

2. LITERATURE SURVEY

The paper titled "A Novel Approach for Face Expressions Recognition" focus on a new method for face expression recognition. Haar functions is used for face, eyes and mouth detection; edge detection method for extracting the eyes correctly and Bezier curves is applied to approximate the extracted regions. Then, a set of distances for varied face type is extracted and it is serve as training input for a multilayer neural network. The novel factor of this approach consists in applying Bezier curves to efficiently extract the distances between facial parts. The pre classification is done using K-means algorithm. A two layered feed-forward neural network created is then used as a classifying tool for the input images. The consistency of the results is demonstrated by the median value. The performance achieved here is 82%. The method is not able to treat situations when the eyes are closed. Strong illumination variations affect the results

D. Drume, introduced and evaluated multi-level classification framework for the emotion classification. This framework include three phases, face localization, facial feature extraction and training & classification. This paper uses principal component analysis at level-1 and support vector machine at level-2 for the training and classification. Results show that this approach successfully recognize facial emotion with 93% recognition rate. The results suggest that the method introduced is

able to support the more accurate classification of emotion from the images. The Neural network classifying method is used in this work to perform facial expression recognition. The expressions classified include the six facial expressions and the neutral one. A neural network, trained using Zernike moments, was applied to the set of the Yale and JAFFE database images in order to perform face detection. Then detected faces were

be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

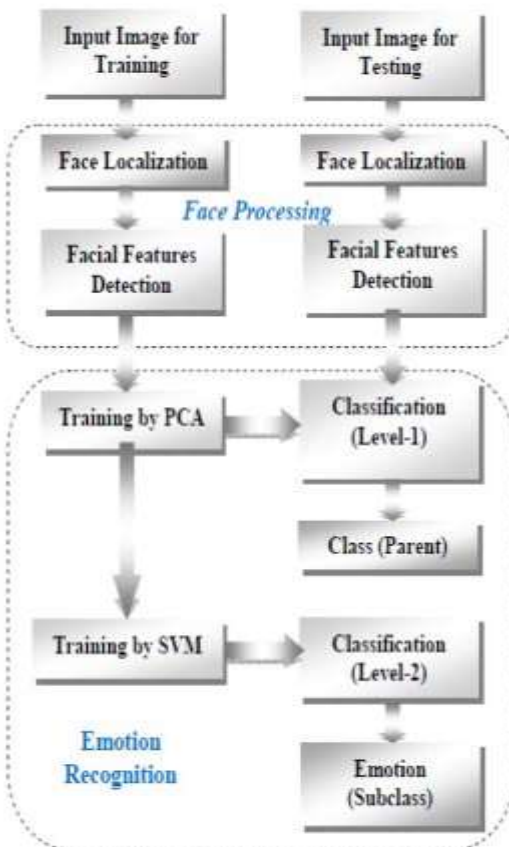
ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY



3. SYSTEM STUDY

3.1 FEASIBILITY STUDY The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to



The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4.1 TYPES OF TESTS

4.2 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an

individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

4.3 Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successful unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

4.4 Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.



Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked. Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as

specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or –



one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

5.CONCLUSION

There is increasing integration of computers and computer interfaces in our lives, due to the rise in the need of computers in order to be able to recognize and respond to human communication and behavioral cues of emotions and mental states. The automated analysis of expressions is a challenging endeavor because of the uncertainty inherent in the inference of hidden mental states from behavioral cues. As the facial expression recognition systems are becoming robust and effective in communications, many other innovative applications and uses are yet to be seen. The objective of this research paper is to give brief overview towards the process, various techniques, and application of facial emotion recognition system.

6. REFERENCES

[1]

https://en.wikipedia.org/wiki/Emotion#Basic_emotions.

[2]U. Hess and P. Thibault, “Darwin and Emotion Expression”, American Psychological Association, 2009, vol. 64, no. 2, 120-128.

[3] S. M. Banu et. al., “A Novel Approach for Face Expressions Recognition”, IEEE 10th Jubilee International Symposium on Intelligent Systems and Informatics, 2012.

[4] D. Drume and A. S. Jalal, “A Multi-level Classification Approach for Facial Emotion Recognition”, IEEE International Conference on Computational Intelligence and Computing Research, 2012.

[5] M. Saaidia et. al., “Facial Expression Recognition Using Neural Network Trained with Zernike Moments”, IEEE International Conference on Artificial Intelligence with Applications in Engineering and Technology, 2014.

[6] Zhiding Yu et. al., “Image based Static Facial Expression Recognition with Multiple Deep Network Learning”, ACM, 2015.

[7] V. D. Bharate et. al., “Human Emotions Recognition using Adaptive Sublayer Compensation and various Feature Extraction Mechanism”, IEEE WiSPNET, 2016

[8]M. Aziz et. al., “Facial Expression Recognition using Multiple Feature Sets”, IEEE, 2015.

[9]D. Datcu, and J. M. Rothkrantz, “Facial Expression Recognition with Relevance Vector Machines”, Interactive Collaborative Information Systems (ICIS).

A MACHINE LEARNING BASED CLASSIFICATION AND PREDICTION TECHNIQUE FOR DDOS ATTACKS

Chirapa Vijaya Priya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Distributed network attacks are referred to, usually, as Distributed Denial of Service (DDoS) attacks. These attacks take advantage of specific limitations that apply to any arrangement asset, such as the framework of the authorized organization's site. In the existing research study, the author worked on an old KDD dataset. It is necessary to work with the latest dataset to identify the current state of DDoS attacks. This paper, used a machine learning approach for DDoS attack types classification and prediction. For this purpose, used Random Forest and XGBoost classification algorithms. To access the research proposed a complete framework for DDoS attacks prediction. For the proposed work, the UNWS-np-15 dataset was extracted from the GitHub repository and Python was used as a simulator. After applying the machine learning models, we generated a confusion matrix for identification of the model performance. In the first classification, the results showed that both Precision (PR) and Recall (RE) are 89% for the Random Forest algorithm. The average Accuracy (AC) of our proposed model is 89% which is superb and enough good. In the second classification, the results showed that both Precision (PR) and Recall (RE) are approximately 90% for the XGBoost algorithm. The average Accuracy (AC) of our suggested model is 90%. By comparing our work to the existing research works, the accuracy of the defect determination was significantly improved which is approximately 85% and 79%, respectively.

1. EXISTING SYSTEM

We studied the latest research papers of the past two years for this research work and also Gozde Karatas et al. proposed a machine learning approach for attacks classification. They used different machine learning algorithms and found that the KNN model is best for classification as compared to other research work. Nuno Martins et al. proposed intrusion detection using machine learning approaches. They used the KDD dataset which is available on the UCI

repository. They performed different supervised models to balance un classification algorithm for better performance. In this work, a comparative study was proposed by the use of different classification algorithms and found good results in their work. Laurens D'hooge et al. proposed a systematic review for malware detection using machine learning models. They compared different malware datasets from online resources as well as approaches for the dataset. They found that machine learning supervised models are very effective for malware detection to make a better decision in less time. Xianwei Gao et al. proposed a comparative work for network traffic classification. They used machine learning classifiers for intrusion detection. The dataset is taken is CICIDS and KDD from the UCI repository. They found support vector machine SVM one of the best algorithms as compare to others. Tongtong Su et al. proposed adaptive learning for intrusion detection. They used the KDD dataset from an online repository. These models are Dtree, R-forest, and KNN classifiers. In this study, the authors found that Dtree and ensemble models are good for classification results. The overall accuracy of the proposed work is 85%. Kaiyuan Jiang et al. proposed deep learning models for intrusion detection. The dataset is KDD and the models are Convention neural network (CNN), BAT-MC, BAT, and Recurrent neural network. The overall model's performance was very good. They found CNN as best for learning. The accuracy is improved from 82% to 85%. Arun Nagaraja *et al.* [5] proposed a hybrid model deep learning model for intrusion detection. They combined two deep learning models for the classification of CNNC LSTM from the RNN model. The dataset was used in this work is KDD. They found an 85.14% average accuracy for the proposed. Yanqing Yang *et al.* [8] proposed a similarity-based approach for anomaly detection using machine learning. They used k mean cluster model for feature similarity detection and naïve Bayes model used for classification. Hui Jiang *et al.* [4] used an auto-encoder for labels and performed deep learning classification models on the KDD dataset. They found an 85% average accuracy for the proposed model SANA ULLAH JAN *et al.* proposed a PSO-Xgboost model because it is higher than the overall classification accuracy alternative models, e.g. Xgboost, Random-Forest, Bagging, and Adaboost. First, establish a classification model based on Xgboost, and then use the adaptive search PSO optimal structure Xgboost. NSL-KDD, reference dataset used for the proposed model evaluation. Our results show that, PSO-Xgboost model of precision, recall, and macro-average average accuracy, especially in determining the U2R and R2L attacks. This work also provides an experimental basis for the application group NIDS in intelligence.

Disadvantages

- 1) .The system doesn't have the accuracy and effectiveness.
- 2). There is no real-world datasets to evaluate OFDPI's exhibitions on the Ryu SDN regulator and Mininet stage.

2. PROPOSED SYSTEM

In this research, we design a framework for the DDoS attack classification and prediction based on the existing dataset that used machine learning methods. This framework involves the following main steps.

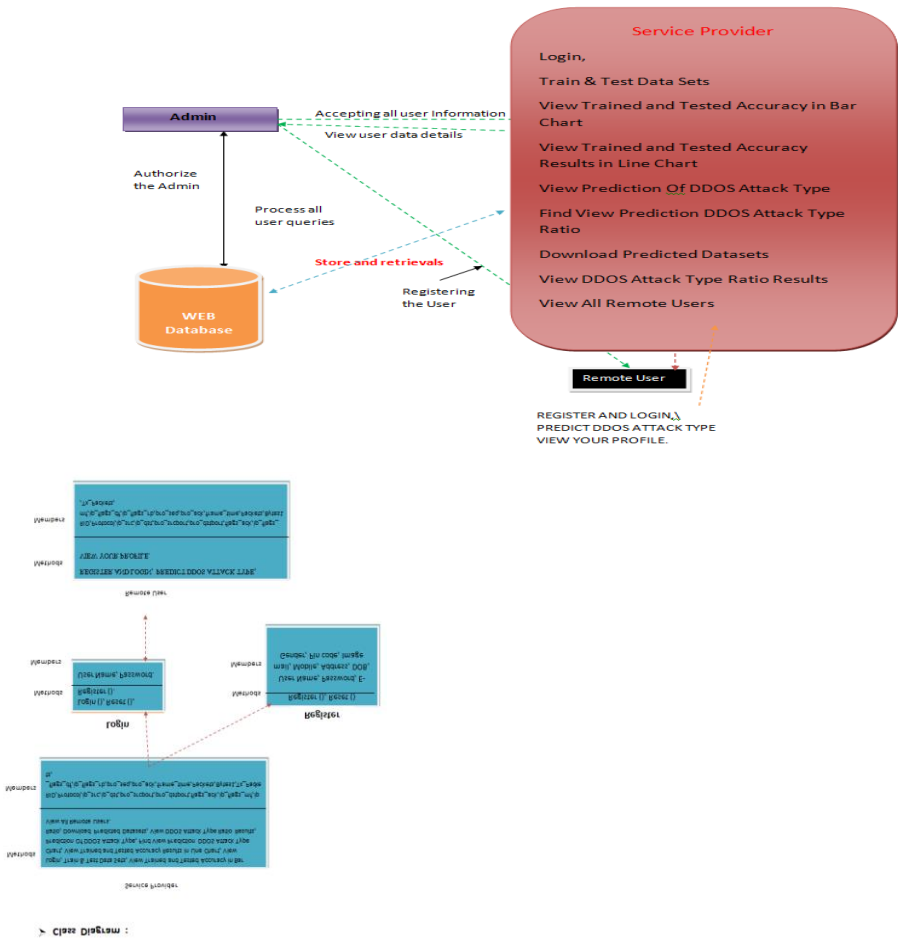
- 1) The first step involves the selection of dataset for utilization.
- 2) The second step involves the selection of tools and language.
- 3) The third step involves data pre-processing techniques to handle irrelevant data from the dataset. In the fourth step feature extraction and label.
- 4) Encoding is performed to convert symbolical data into numerical data.
- 5) In the fifth step, the data splitting is performed into a train and test set for the model. In this step, we build and train our proposed model. However, model optimization is also performed on the trained model in terms of kernel scaling and kernel hyper-parameter tuning to improve model efficiency. When the model optimizes then we will generate output results from the model.

Advantages

The system is designed and developed an approach using supervised machine learning classifiers for DDoS attack detection based on different techniques.

The proposed system is designed a step-by-step framework for data utilization.

Architecture Diagram



3. SYSTEM STUDY

3.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

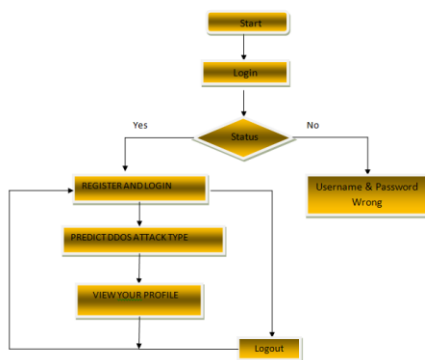
Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

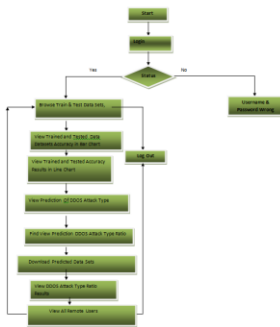
SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

➤ Flow Chart : Remote User



➤ Flow Chart :Service Provider



4. CONCLUSION

In this paper, we proposed a complete systematic approach for detection of the DDOS attack. First, we selected the UNSW-nb15 dataset from the GitHub repository that contains information about the DDOS attacks. This dataset was provided by the Australian Centre for Cyber Security (ACCS) [29], [30]. Then, Python and jupyter notebook were used to work on data wrangling. Secondly, we divided the dataset into two classes i.e. the dependent class and the independent class. Moreover, we normalized the dataset for the algorithm. After data normalization, we applied the proposed, supervised, machine learning approach. The model

generated prediction and classification outcomes from the supervised algorithm. Then, we used Random Forest and XG Boost classification algorithms. In the first classification, we observed that both the Random Forest Precision (PR) and Recall (RE) are approximately 89% accurate. Furthermore, we noted approximately 89% average Accuracy (AC) for the proposed model that is enough good and extremely awesome. Note that the average Accuracy illustrates the F1 score as 89%. For the second classification, we noted that both the XG Boost Precision (PR) and Recall (RE) are approximately 90% accurate. We noted approximately 90% average Accuracy (AC) of the suggested model that is wonderful and extremely brilliant. Again, the average Accuracy illustrates the F1 score as 90%. By comparing the proposal to existing research works, the defect determination accuracy of the existing research [4] which was 85% and 79% were also significantly improved.

Looking to the future, for functional applications, it is important to provide a more user-friendly, faster alternative to deep learning calculations, and produce better results with a shorter burning time. It is important to work on unsupervised learning toward supervised learning for unlabeled and labeled datasets. Moreover, we will investigate how non-supervised learning algorithms will affect the DDOS attacks detection, in particular, we non-labeled datasets are taken into account.

5. REFERENCES

- [1] N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, "Adversarial machine learning applied to intrusion and malware scenarios: A systematic review," *IEEE Access*, vol. 8, pp. 35403_35419, 2020.
- [2] G. Karatas, O. Demir, and O. K. Sahingoz, "Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset," *IEEE Access*, vol. 8, pp. 32150_32162, 2020.
- [3] T. Su, H. Sun, J. Zhu, S. Wang, and Y. Li, "BAT: Deep learning methods on network intrusion detection using NSL-KDD dataset," *IEEE Access*, vol. 8, pp. 29575_29585, 2020.

- [4] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network intrusion detection based on PSO-xgboost model," *IEEE Access*, vol. 8, pp. 58392_58401, 2020.
- [5] A. Nagaraja, U. Boregowda, K. Khatatneh, R. Vangipuram, R. Nuvvusetty, and V. S. Kiran, "Similarity based feature transformation for network anomaly detection," *IEEE Access*, vol. 8, pp. 39184_39196, 2020.
- [6] L. D'hooge, T. Wauters, B. Volckaert, and F. De Turck, "Classification hardness for supervised learners on 20 years of intrusion detection data," *IEEE Access*, vol. 7, pp. 167455_167469, 2019.
- [7] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512_82521, 2019.
- [8] Y. Yang, K. Zheng, B. Wu, Y. Yang, and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," *IEEE Access*, vol. 8, pp. 42169_42184, 2020.
- [9] C. Liu, Y. Liu, Y. Yan, and J. Wang, "An intrusion detection model with hierarchical attention mechanism," *IEEE Access*, vol. 8, pp. 67542_67554, 2020.
- [10] S. U. Jan, S. Ahmed, V. Shakhov, and I. Koo, "Toward a lightweight intrusion detection system for the Internet of Things," *IEEE Access*, vol. 7, pp. 42450_42471, 2019.

WATERNET A NETWORK FOR MONITORING AND ASSESSING WATER QUALITY FOR DRINKING AND IRRIGATION PURPOSES

Chitti Swathi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Water is a fundamental requirement for human, animal, and plant survival. Despite its importance, quality water is not always t for drinking, domestic and/or industrial use. Numerous factors such as industrialization, mining, pollution, and natural occurrences impact the quality of water, as they introduce or alter various parameters present therein, thus, affecting its suitability for human consumption or general use. The World Health Organization has guidelines which stipulate the threshold levels of various parameters present in water samples intended for consumption or irrigation. The Water Quality Index (WQI) and Irrigation WQI (IWQI) are metrics used to express the level of these parameters to determine the overall water quality. Collecting water samples from different sources, measuring the various parameters present, and bench-marking these measurements against pre-set standards, while adhering to various guidelines during transportation and measurement can be extremely daunting. To this end this study proposes a network architecture to collect data on water parameters in real-time and use Machine Learning (ML) tools to automatically determine suitability of water samples for drinking and irrigation purposes. The developed monitoring network is based on LoRa and takes the land topology into consideration. Results of simulations done in Radio Mobile revealed a partial mesh network topology as the most adequate. Due to the absence of large and open datasets on drinking and irrigation water, datasets usable for training ML models were developed. Three ML models - Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) were considered for the water classification process and results obtained showed that LR performed best for drinking water, while SVM was better suited for irrigation water. Recursive feature elimination was then combined with the three ML models to reveal

which of the water parameters had the greatest influence on the classification accuracies of the respective model.

1. INTRODUCTION

Access to water is a critical component of human lives and is now considered a basic human right. Access to clean water is also one of the 17 Sustainable Development Goals (SDG) set up by the United Nations in 2015 to achieve a better future for all. Specifically, the sixth goal, which is to ensure and sustain the availability of water and sanitation to all. Potable water can also be linked to the third SDG goal – good health and well-being, as contaminated water can be a transmission medium for diseases such as cholera, typhoid, and diarrhoea, which are jointly the highest cause of mortality (especially children) in developing nations of Africa and Asia. Water is also important in agriculture and food production. Recent statistics shows that about 10% of the world population is malnourished, with developing countries being hit the hardest, with starvation resulting in about 45% of infant mortality. Ensuring global food security is thus of utmost importance. Food security has been recognized as a critical requirement, hence its inclusion as one of the SDG (goal 2), with specific focus on ending hunger, by promoting sustainable agriculture and improving food distribution. Food production and agriculture in general rely heavily on water, both for irrigation and for animal consumption. It is thus pertinent to ensure the availability and sustainable management of water for agricultural use. There are several sources of water for both drinking and irrigation use, including rivers, streams, rain, and groundwater (accessed through wells and boreholes). The nature and characteristics of a source of water are often critical factors that influence the constituents of water samples obtained therein. Beyond natural factors, chemical wastes from human activities such as mining, crude oil extraction, and industrial wastes, most often end up in streams, rivers, and other sources of water, changing the nature and properties of these waters. These waters then end up in homes or farms, where they are used for domestic purposes, drunk, fed to livestock, or used to water crops. Consuming this type of water can have dire health consequences or result in death. It is therefore paramount that a proper process be put in place to ensure end-to-end monitoring of the water right from the source to its last point of use. At each monitoring point, samples of water need to be collected to assess the quality or "fitness for use" for human (and animal) consumption, irrigation and domestic (or industrial) uses.

2. EXISTING SYSTEM

In a network for measuring and monitoring water parameters in a metal producing city in Brazil was developed. Twelve water monitoring stations were setup to measure several physico-chemical water parameters, including pH, dissolved solids, Zinc, Lead etc. Finally, obtained results were analysed using principal component analysis. In a similar manner, developed a system to monitor water quality in Limpopo River Basin in Mozambique and set up 23 monitoring stations to measure physico-chemical and microbiological parameters, and ultimately assess the quality of water in the river basin. To address the challenges of optimal placement of gauges and sampling frequencies, which are often faced when developing water monitoring systems, the authors in developed an economically viable model that combined genetic algorithm with 1-D water quality simulation. Though the work was only simulated by using genetic algorithm, the authors were able to solve the NP hard problem of optimally placing monitoring stations. Monitoring water parameters often entails periodically sampling a body of water to capture relevant metrics. These metrics might include physico-chemical and microbiological measurements, such as potential of hydrogen (pH), temperature, sodium levels etc. In a water monitoring network, measured parameters need to be transferred to a base station where relevant decision(s) would be taken. Due to the sparse nature of transmitted data, light weight communication protocols capable of transmitting relatively small data over long distance are required for water monitoring networks. From literature, Low Power Wide Area Network (LPWAN) technologies have been favoured for such applications. An extensive discussion on LPWAN technologies was done in . The work compared a few sub-GHz solutions including SigFox, LoRa, Ingenu and Telensa, with respect to their range, transmission rate, and channel count. Ingenu was reported to have the longest range in city settings at 15 km, followed by SigFox at 10 km (in cities) and 50 km (in rural areas); then LoRa at 5 km (in cities), and 15 km in rural settings. Regarding the assessment of communication technologies, there has been a long-drawn debate over the efficacy of software simulations versus real-world testing. Though this debate still rages, several researchers have shown that simulation results are often at par with real-world tests. For instance, using LoRa, the authors in compared simulation results with real world test for intervehicle communication. They used NS3 as a simulation platform and an Arduino UNO C Dragino LoRa module for the real-world tests, while Propagation loss, coverage Packet Inter-

reception (PIR), Packet Delivery Ratio (PDR) and Received Signal Strength Indicator (RSSI) level were used as benchmark metrics. They concluded that the results of the simulator were consistent with those of the real-world tests. In a similar work, Hassan also compared the efficacy of simulation results (from Radio Mobile simulator) with real-world tests (using micro controllers C LoRa modules) when using LoRa as a bridge for Wi-Fi. Unlike did not give a side-by-side comparison of simulated vs. real-world results for each metric considered but concluded that the simulator performed well. set up seven pairs of XBee modules and compared communication performance using both the 800/900MHz and 2.4GHz frequencies. They concluded that simulation results from the Radio Mobile simulator corroborated with those of real-world tests.

Disadvantages

An existing methodology doesn't implement DATA PRE-PROCESSING & LABELLING method.

The system not implemented Calculating WQI for Irrigation Water for prediction in the datasets.

3. PROPOSED SYSTEM

The water monitoring network proposed in this work is to be deployed in the City of Cape Town in Western Cape, South Africa, with the intention of monitoring water parameters in water storage dams and/or water treatment plants across the city. Data gathered by the monitoring network are then passed through Machine Learning (ML) models to determine their suitability for consumption or irrigation purposes.

1) Build a network for real-time collection and monitoring of water quality across water storage dams in the city of Cape Town. This network takes into consideration the unique geographical features of Cape Town, such as mountains and elevations that might obstruct radio frequency propagation.

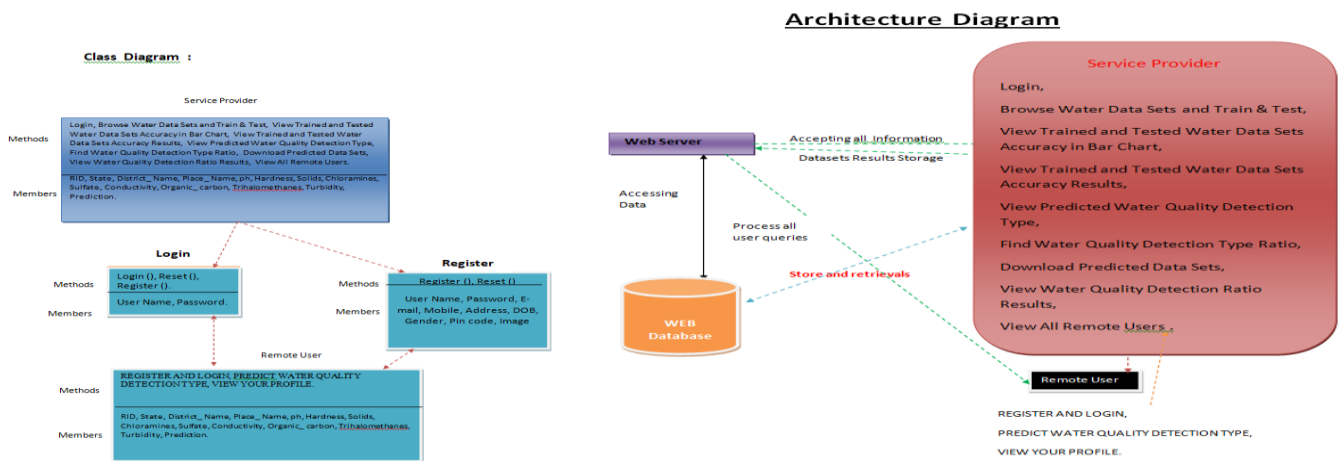
2) Curate ample sized datasets on drinking and irrigation water that can be used to train (and test) machine learning models to automatically determine the "fitness for use" of a sample of water for drinking and/or irrigation purposes.

3) Build models that determine the most critical parameters that influence the accuracy of machine learning models in analyzing water for drinking or irrigation.

Advantages

The purpose of WaterNet is to gather data on water parameters from dams across the city. These parameters are then used to assess the quality of water with regards "fitness for use" for drinking and irrigation purposes.

In this work, rather than relying on instrumental and physico-chemical analysis carried out in laboratories to assess water parameters, we propose the use of machine learning (ML) models, which take the numerous water parameters into consideration and automatically determine if a sample of water is potable or fit for agricultural use.



4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is

not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSION

This work focused on two major concept, firstly, the proposal of a real-time water monitoring network for gathering data on water parameters from water bodies. Secondly, the application of machine learning (ML) models as means of assessing water quality. The developed water monitoring network is based on Lo Ra, a low power long range protocol for data transmission, and was developed using the City of Cape Town as case study. Results of the simulation done in Radio Mobile, revealed a partial mesh network topology as the most adequate network to cover the city. Data gathered from this monitoring network would ideally be aggregated on a Cloud server, where ML models can then be applied to assess the water's fitness of use for drinking or irrigation purposes. Due to the absence of relevant datasets, two suitable datasets were built in this work and used to training and testing three ML models considered, which are Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM). Results of the test showed that LR performed best for drinking water, as it gave the highest classification accuracy and lowest false positive and negative values, while SVM was better suited for irrigation water.

Finally, a model for identifying the most influential water parameter(s) w.r.t classification accuracies of the ML models was then explored using recursive feature elimination (RFE). Obtained results showed that pH, and total hardness were the least influential parameters in drinking water, while SSP was the least for irrigation water.

7. REFERENCES

- [1] B. X. Lee, F. Kjaerulf, S. Turner, L. Cohen, P. D. Donnelly, R. Muggah, R. Davis, A. Realini, B. Kieselbach, L. S. MacGregor, I. Waller, R. Gordon, M. Moloney-Kitts, G. Lee, and J. Gilligan, "Transforming our world: Implementing the 2030 agenda through sustainable development goal indicators," *J. Public Health Policy*, vol. 37, no. S1, pp. 13_31, Sep. 2016.
- [2] *Integrated Approaches for Sustainable Development Goals Planning: The Case of Goal 6 on Water and Sanitation*, U. ESCAP, Bangkok, Thailand, 2017.
- [3] WHO. Water. *Protection of the Human Environment*. Accessed: Jan. 24, 2022. [Online]. Available: www.afro.who.int/health-topics/water
- [4] L. Ho, A. Alonso, M. A. E. Forio, M. Vanclooster, and P. L. M. Goethals, "Water research in support of the sustainable development goal 6: A case study in Belgium," *J. Cleaner Prod.*, vol. 277, Dec. 2020, Art. no. 124082.
- [5] *Global Nutrition Report 2016: From Promise to Impact: Ending Malnutrition by 2030*, International Food Policy Research Institute, Washington, DC, USA, 2016, doi: 10.2499/9780896295841.
- [6] N. Akhtar, M. I. S. Ishak, M. I. Ahmad, K. Umar, M. S. Md Yusuff, M. T. Anees, A. Qadir, and Y. K. A. Almanasir, "Modification of the water quality index (WQI) process for simple calculation using the multicriteria decision-making (MCDM) method: A review," *Water*, vol. 13, no. 7, p. 905, Mar. 2021.
- [7] World Health Organization. (1993). *Guidelines for Drinking-Water Quality*. World Health Organization. Accessed: Jan. 12, 2022. [Online]. Available: <http://apps.who.int/iris/bitstream/handle/10665/44584/9789241548151-eng.pdf>
- [8] *Standard Methods for the Examination of Water and Wastewater*, Federation WE, APH Association, American Public Health Association (APHA), Washington, DC, USA, 2005.

[9] L. S. Clesceri, A. E. Greenberg, and A. D. Eaton, "Standard methods for the examination of water and wastewater," Amer. Public Health Assoc. (APHA), Washington, DC, USA. Tech. Rep.21, 2005.

[10] M. F. Howladar, M. A. Al Numanbakth, and M. O. Faruque, "An application of water quality index (WQI) and multivariate statistics to evaluate the water quality around Maddhapara granite mining industrial area, Dinajpur, Bangladesh," *Environ. Syst. Res.*, vol. 6, no. 1, pp. 1_8, Jan. 2018.



COMPOSITE BEHAVIORAL MODELING FOR IDENTITY THEFT DETECTION IN ONLINE SOCIAL NETWORKS

Dasari Kusuma Prabha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. V. Bhaskara Murthy, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- In this work, we aim at building a bridge from coarse behavioral data to an effective, quick-response, and robust behavioral model for online identity theft detection. We concentrate on this issue in online social networks (OSNs) where users usually have composite behavioral records, consisting of multidimensional low-quality data, e.g., offline check-ins and online user-generated content (UGC). As an insightful result, we validate that there is a complementary effect among different dimensions of records for modeling users' behavioral patterns. To deeply exploit such a complementary effect, we propose a *joint* (instead of *fused*) model to capture both online and offline features of a user's composite behavior. We evaluate the proposed joint model by comparing it with typical models and their fused model on two real-world datasets: Foursquare and Yelp. The experimental results show that our model outperforms the existing ones, with the area under the receiver operating characteristic curve (AUC) values 0.956 in Foursquare and 0.947 in Yelp, respectively. Particularly, the *recall* (true positive rate) can reach up to 65.3% in Foursquare and 72.2% in Yelp with the corresponding *disturbance rate* (false-positive rate) below 1%. It is worth mentioning that these performances can be achieved by examining only one composite behavior, which guarantees the low response latency of our method. This study would give the cybersecurity community new insights into whether and how real-time online identity authentication can be improved via modeling users' composite behavioral patterns.

1. INTRODUCTION

The rapid development of the Internet, more and more affairs, e.g., mailing health caring shopping booking hotels, and purchasing tickets, are handled online. Meanwhile, the Internet also brings sundry potential risks of invasions, such as losing financial identity theft and privacy leakage. Online accounts serve as the agents of users in the cyber world. Online identity theft is a typical online crime which is the deliberate use of another person's account usually as a method to gain a financial advantage or

obtain credit and other benefits in another person's name. As a matter of fact, compromised accounts are usually the portals of most cybercrimes such as blackmail fraud and spam. Thus, identity theft detection is essential to guarantee users' security in the cyber world. Traditional identity authentication methods are mostly based on access control schemes, e.g., passwords and tokens. But users have some overheads in managing dedicated passwords or tokens. Accordingly, the biometric identification is delicately



introduced to start the era of password-free. However, some disadvantages make these access control schemes still incapable of being effective in real-time online services

1) They are not *nonintrusive*. Users have to spend extra time in the authentication.

2) They are not *continuous*. The defending system will fail to take further protection once the access control is broken. Behavior-based suspicious account detection [16], [18], [19] is a highly anticipated solution to pursue a nonintrusive and continuous identity authentication for online services. It depends on capturing users' suspicious behavior patterns to discriminate the suspicious accounts. The problem can be divided into two categories: fake/sybil account detection [20] and compromised account detection [21]. The fake/Sybil account's behaviors usually do not conform to the behavioral pattern of the majority. Meantime, the compromised account usually behaves in a pattern that does not conform to its previous one, even behaves like fake/sybil accounts. It can be solved by capturing *mutations* of users' behavioral patterns. Comparing with detecting compromised accounts, detecting fake/sybil accounts is relatively easy since the latter's behaviors are generally more detectable than the former's. It has been extensively studied and can be realized by various population-level approaches, e.g., clustering [22], [23], classification [5], [24]–[26] and statistical or empirical rules [8], [27], [28]. Thus, we *only* focus on the compromised account detection, commonly called *identity theft detection*, based on individual-level behavioral models.

2. EXISTING SYSTEM

introduced hand movement, orientation, and grasp (HMOG), a set of behavioral features to continuously authenticate smartphone users. Rajoub and Zwiggelaar used thermal imaging to monitor the periorbital region's thermal variations and test whether it can offer a discriminative signature for detecting deception. However, these biometric technologies usually require expensive hardware devices which makes it inconvenient and difficult to popularize. explored a multimodal deception detection approach that relied on a novel dataset of 149 multimodal recordings, and integrated multiple physiological, linguistic, and thermal features. These works indicated that users' behavior patterns can represent their identities. Many studies turn to utilize users' behavior patterns for identifications. Behavior-based methods were born at the right moment, which plays important roles in a wide range of tasks including preventing and detecting identity theft. Typically, behavior-based user identification includes two phases: user profiling and user identifying. User profiling is a process to characterize a user with his/her history behavioral data. Some works focus on statistical characteristics, such as the mean, variance, median, or frequency of a variable, to establish the user profile. Naini *et al.* [55] studied the task of identifying the users by matching the histograms of their data in the anonymous dataset with the histograms from the original dataset. But it mainly relied on experts' experience since different cases usually have different characteristics. proposed a behavior-based method to identify compromises of individual high-



profile accounts. However, it required high-profile accounts which were difficult to obtain. Other researchers discovered other features, such as tracing patterns, topic and spatial distributions, to describe user identity. conducted a study on online user behavior by collecting and analyzing user clickstreams of a well-known OSN. developed a topic model extending the LDA to identify the active users. presented a technique based on principal component analysis (PCA) that accurately modeled the “like” behavior of normal users in Facebook and identified significant deviations from it as anomalous behaviors. proposed an approach that involved the novel collection of online news stories and reports on the topic of identity theft. Lichman and Smyth [48] proposed MKDE model to accurately characterize and predict the spatial pattern of an individual’s events. Tsikerdekis and Zeadally presented a detection method based on nonverbal behavior for identity deception, which can be applied to many types of social media. These methods above mainly concentrated on a specific dimension of the composite behavior and seldom thought about utilizing multidimensional behavior data. explored the complex interaction between social and geospatial behavior and demonstrated that social behavior can be predicted with high precision. It indicated that composite behavior features can identify one’s proposed a probabilistic generative model combining the use of spatiotemporal data and semantic information to predict user’s behavior. presented POISED, a system that leverages the differences in propagation between benign and malicious messages on social

networks to identify spam and other unwanted content. These studies implied that composite behavior features are possibly helpful for user identification.

Disadvantages

- 1) LDA model performs poorly in both datasets which may indicate its performance is strongly sensitive to the data quality.
- 2) CF-KDE and LDA model performs not well in Yelp dataset comparing to Foursquare dataset, but the fused model [17] observes a surprising reversion.
- 3) The joint model based on *relative anomalous score* S_r outperforms the model based on *logarithmic anomalous score* S_l .
- 4) The joint model (i.e., JOINT-SR, the joint model in the following content of the system all refer to the joint model based on S_r) is indeed superior to the fused model.

3. PROPOSED SYSTEM

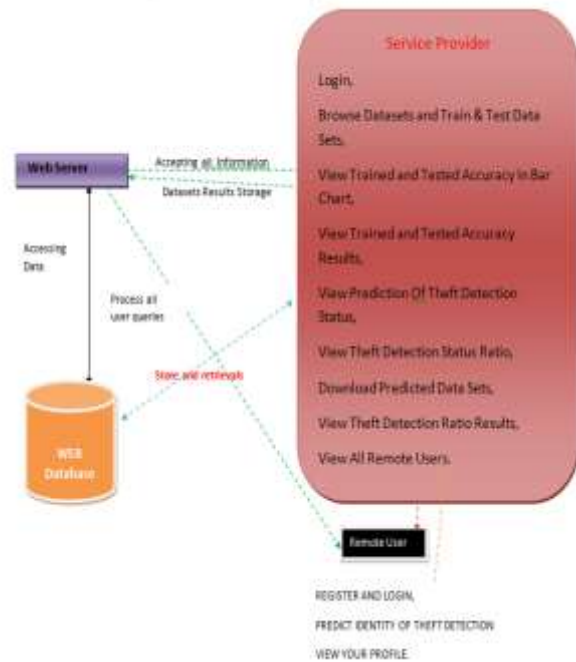
In this article, we propose an approach to detect identity theft by using multidimensional behavioral records which are possibly insufficient in each dimension. According to such characteristics, we choose the online social network (OSN) as a typical scenario where most users’ behaviors are coarsely recorded [39]. In the Internet era, users’ behaviors are composited by offline behaviors, online behaviors, social behaviors, and perceptual/cognitive behaviors. The behavioral data can be collected in many applications, such as offline check-ins in location-based services (LBSs), online tips-posting in instant messaging services, and social relationship making in online social services. Accordingly, we design our method based on users’ composite behaviors by these categories. In OSNs, user behavioral data

that can be used for online identity theft detection are often too low-quality or restricted to build qualified behavioral models due to the difficulty of data collection, the requirement of user privacy, and the fact that some users have a few several behavioral records. We devote ourselves to proving that a high-quality (effective, quick response, and robust) behavioral model can be obtained by integrally using multidimensional behavioral data, even though the data is extremely insufficient in each dimension.

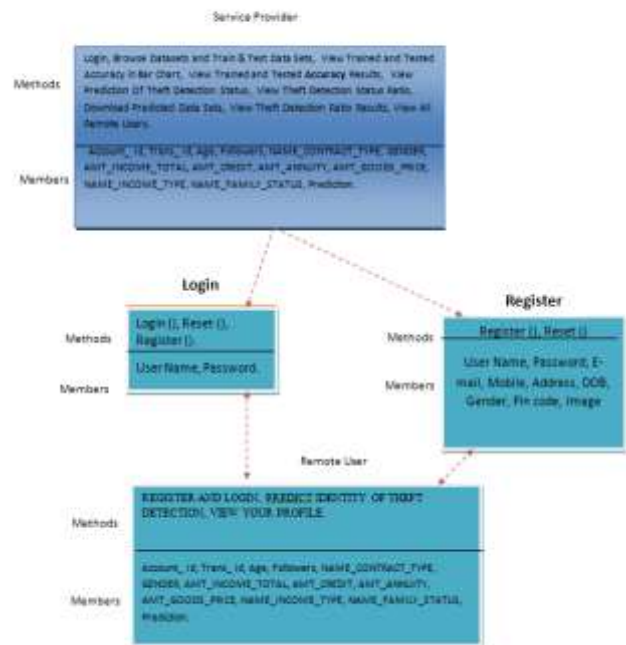
Advantages

- 1) We propose a joint model, CBM, to capture both online and offline features of a user's composite behavior to fully exploit coarse behavioral data.
- 2) We devise a relative anomalous score S_r to measure the occurrence rate of each composite behavior for realizing real-time identity theft detection.
- 3) We perform experiments on two real-world datasets to demonstrate the effectiveness of CBM. The results show that our model outperforms the existing models and has the low response latency.

Architecture Diagram



Class Diagram :



4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth



with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSION

We investigate the feasibility of building a ladder from low-quality behavioral data to a high-performance behavioral model for user identification in OSNs. By deeply exploiting the complementary effect among OSN users' multidimensional behaviors, we propose a joint probabilistic generative model by integrating online and offline behaviors. When the designed joint model is applied to identity theft detection in OSNs, its comprehensive performance, in terms of the detection efficacy, response latency, and

robustness, is validated by extensive evaluations on real-life OSN datasets. Particularly, the joint model significantly outperforms the existing fused model. Our behavior-based method mainly aims at detecting identity

7. REFERENCES

- [1] J. Onaolapo, E. Mariconti, and G. Stringhini, "What happens after you are pwned: Understanding the use of leaked Webmail credentials in the wild," in *Proc. Internet Meas. Conf.*, Nov. 2016, pp. 65–79.
- [2] A. Mohan, "A medical domain collaborative anomaly detection framework for identifying medical identity theft," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2014, pp. 428–435.
- [3] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, Jan. 2015.
- [4] P. Hyman, "Cybercrime: It's serious, but exactly how serious?" *Commun. ACM*, vol. 56, no. 3, pp. 18–20, Mar. 2013.
- [5] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirida, "All your contacts are belong to us: Automated identity theft attacks on social networks," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, 2009, pp. 551–560.
- [6] J. Lynch, "Identity theft in cyberspace: Crime control methods and their effectiveness in combating phishing attacks," *Berkeley Technol. Law J.*, vol. 20, no. 1, pp. 259–300, 2005.
- [7] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," *IEEE Trans. Dependable Secure Comput.*, vol. 14, no. 4, pp. 447–460, Jul. 2017.



- [8] T. C. Pratt, K. Holtfreter, and M. D. Reisig, "Routine online activity and Internet fraud targeting: Extending the generality of routine activity theory," *J. Res. Crime Delinquency*, vol. 47, no. 3, pp. 267–296, Aug. 2010.
- [9] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song, "Design and evaluation of a real-time URL spam filtering service," in *Proc. IEEE Symp. Secur. Privacy*, May 2011, pp. 447–462.
- [10] H. Li *et al.*, "Bimodal distribution and co-bursting in review spam detection," in *Proc. 26th Int. Conf. World Wide Web*, Apr. 2017, pp. 1063–1072.
- [11] A. M. Marshall and B. C. Tompsett, "Identity theft in an online world," *Comput. Law Secur. Rev.*, vol. 21, no. 2, pp. 128–137, Jan. 2005.
- [12] B. Schneier, "Two-factor authentication: Too little, too late," *Commun. ACM*, vol. 48, no. 4, p. 136, Apr. 2005.
- [13] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "CEREBRE: A novel method for very high accuracy event-related potential biometric identification," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1618–1629, Jul. 2016.
- [14] R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, F. Scotti, and G. Sforza, "Biometric recognition in automated border control: A survey," *ACM Comput. Surv.*, vol. 49, no. 2, p. 24, 2016.
- [15] B. A. Rajoub and R. Zwiggelaar, "Thermal facial analysis for deception detection," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 6, pp. 1015–1023, Jun. 2014.
- [16] M. M. Waldrop, "How to hack the hackers: The human side of cybercrime," *Nature*, vol. 533, no. 7602, pp. 164–167, May 2016.
- [17] C. Wang, B. Yang, J. Cui, and C. Wang, "Fusing behavioral projection models for identity theft detection in online social networks," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 4, pp. 637–648, Aug. 2019.
- [18] C. Shen, Y. Li, Y. Chen, X. Guan, and R. A. Maxion, "Performance analysis of multi-motion sensor behavior for active smartphone authentication," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 48–62, Jan. 2018.
- [19] C. Wang and H. Zhu, "Representing fine-grained co-occurrences for behavior-based fraud detection in online payment services," *IEEE Trans. Dependable Secure Comput.*, early access, May 4, 2020, doi: 10.1109/TDSC.2020.2991872.
- [20] H. Zheng *et al.*, "Smoke screener or straight shooter: Detecting elite sybil attacks in user-review social networks," in *Proc. 25th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Feb. 2018, pp. 259–300

AN ARTIFICIAL INTELLIGENCE AND CLOUD BASED COLLABORATIVE PLATFORM FOR PLANT DISEASE IDENTIFICATION, TRACKING AND FORECASTING FOR FARMERS

Dasari Srinath (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- Plant diseases are a major threat to farmers, consumers, environment and the global economy. In India alone, 35% of field crops are lost to pathogens and pests causing losses to farmers. Indiscriminate use of pesticides is also a serious health concern as many are toxic and biomagnified. These adverse effects can be avoided by early disease detection, crop surveillance and targeted treatments. Most diseases are diagnosed by agricultural experts by examining external symptoms. However, farmers have limited access to experts. Our project is the first integrated and collaborative platform for automated disease diagnosis, tracking and forecasting. Farmers can instantly and accurately identify diseases and get solutions with a mobile app by photographing affected plant parts. Realtime diagnosis is enabled using the latest Artificial Intelligence (AI) algorithms for Cloud-based image processing. The AI model continuously learns from user uploaded images and expert suggestions to enhance its accuracy. Farmers can also interact with local experts through the platform. For preventive measures, disease density maps with spread forecasting are rendered from a Cloud based repository of geo-tagged images and micro-climatic factors. A web interface allows experts to perform disease analytics with geographical visualizations. In our experiments, the AI model (CNN) was trained with large disease datasets, created with plant images self-collected from many farms over 7 months. Test images were diagnosed using the automated CNN model and the results were validated by plant pathologists. Over 95% disease identification accuracy was achieved. Our solution is a novel, scalable and accessible tool for disease management of diverse agricultural crop plants and can be deployed as a Cloud based service for farmers and experts for ecologically sustainable crop production.

Keywords - Crop Diseases, Agriculture, Artificial Intelligence, Cloud, CNN, Mobile, Plant Pathology, Neural Networks

1.INTRODUCTION

Agriculture is fundamental to human survival. For populated developing countries like India, it is even more imperative to increase the productivity of crops, fruits and vegetables. Not only productivity, the quality of produce needs to stay high for better public health. However, both productivity and quality of food gets hampered by factors such as spread of diseases that could have been prevented with early diagnosis. Many of these diseases are infectious leading to total loss of crop yield. Given the vast geographical spread of agricultural lands, low education levels of farmers coupled with limited awareness and lack of access to plant pathologists, human assisted disease diagnosis is not effective and cannot keep up with the exorbitant requirements. To overcome the shortfall of human assisted disease diagnosis, it is imperative to build automation around crop disease diagnosis with technology and introduce low cost and accurate machine assisted diagnosis easily accessible to farmers. Some strides have been made in applying technologies such as robotics and computer vision systems to solve myriad problems in the agricultural domain. The potential of image processing has been explored to assist with precision agriculture practices, weed and herbicide technologies, monitoring plant growth and plant nutrition management \However, progress on automating plant disease diagnosis is still rudimentary in spite of the fact that many plant diseases can be identified by plant pathologists by visual inspection of physical symptoms such as detectable change in color, wilting, appearance of spots and lesions etc. along with soil and climatic conditions. Overall, the commercial level of investment in bridging agriculture and technology remains lower as compared to investments done in more lucrative fields such as human health and education. Promising research efforts have not been able to productize due to challenges such as access and linkage for farmers to plant pathologists, high cost of deployment and scalability of solution. Recent developments in the fields of Mobile technology, Cloud computing and Artificial Intelligence (AI) create a perfect opportunity for creating a scalable low-cost solution for crop diseases that can be widely deployed. In developing countries such as India, mobile phones with internet connectivity have become ubiquitous. Camera and GPS enabled low cost mobile phones are widely available that can be leveraged by individuals to upload images with geolocation. Over widely available mobile networks, they can communicate with more sophisticated Cloud

based backend services which can perform the compute heavy tasks, maintain a centralized database, and perform data analytics. Another leap of technology in recent years is AI based image analysis which has surpassed human eye capabilities and can accurately identify and classify images. The underlying AI algorithms use Neural Networks (NN) which have layers of neurons with a connectivity pattern inspired by the visual cortex. These networks get “trained” on a large set of pre-classified “labeled” images to achieve high accuracy of image classification on new unseen images. Since 2012 with “AlexNet” winning the ImageNet competition, deep Convolutional Neural Networks (CNNs) have consistently been the winning architecture for computer vision and image analysis [3]. The breakthrough in the capabilities of CNNs have come with a combination of improved compute capabilities, large data sets of images available and improved NN algorithms. Besides accuracy, AI has evolved and become more affordable and accessible with open source platforms such as TensorFlow [4]. Prior art related to our project includes initiatives to gather healthy and diseased crop images [5], image analysis using feature extraction [6], RGB images [7], spectral patterns [8] and fluorescence imaging spectroscopy [9]. Neural Networks have been used in the past for plant disease identification but the approach was to identify texture features. Our proposal takes advantage of the evolution of Mobile, Cloud and AI to develop an end-to-end crop diagnosis solution that simulates the expertise (“intelligence”) of plant pathologists and brings it to farmers. It also enables a collaborative approach towards continually increasing the disease database and seeking expert advice when needed for improved NN classification accuracy and tracking for outbreaks .

AN END-TO- END SOLUTION FOR CROP DIAGNOSIS

Our proposed solution brings plant disease diagnostics to farmers through a Cloud based scalable collaborative platform. The platform is accessible through a mobile app that enables users to upload images of multiple parts of their plant and get the plant disease automatically diagnosed in real-time. They can also view “disease-density” map for their neighborhood showing geographical spread of diseases. The uploaded image gets classified by our AI engine into the appropriate category of disease for which a previously identified best-known method solution is provided to the individual. Simultaneously, the geo-location of the image and a time stamp is used to tag the presence of the particular disease in that location. A collective density of diseases stored in a Cloud database is displayed on a map to show its location relative to the user. This allows the user to take preventive measures based on diseases in their neighborhood and serves

as an alert for any spreading epidemic. The major components in the end-to-end system architecture of the proposed solution is shown in Fig. 1

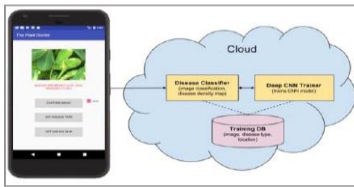


Fig. 1. System architecture with Cloud and Mobile components

and the description of the components is provided below.

□ **Mobile App** - The mobile app contains a simplified frontend for the farmer that is easy to use and hides the complexity of the backend. It enables the user to take images of the plant (*live* mode) or choose existing images from the gallery (*offline* mode) and upload them to the Cloud backend for analysis. It allows them to get the disease type of the uploaded images with a score reflecting the probability or accuracy of classification. It also enables the user to view a disease density map of the local area (if location service is enabled on the phone). Overall, the mobile app has 8 screens (sign-in with mobile number, main page with options, capture new image, load existing image, get disease type, get disease maps, history and expert connect). Android Studio 3.1.3 was used to develop the mobile app in Java with usage of Google Camera API and Maps API. The mobile app communicates with the Cloud backend running on Amazon Web Services (AWS) over the cellular network using AWS Mobile SDK for Android.

□ **Disease Classifier** – The Classifier is a standalone application running in the Cloud platform that receives the images uploaded via the mobile app and uses a trained deep Convolutional Neural Network (CNN) model to classify the disease type. The CNN model is computed by the Deep CNN Trainer and is used by the Classifier to automatically classify the uploaded images into the correct disease type. The Classifier also performs post-processing such as making a decision on whether the uploaded images should be added to the Training Database based on the classification score or sent to an agricultural expert registered on the platform for further analysis. When the classification score is greater than a preconfigured threshold, the images along with their metadata such as disease type and location of the images get added to the Training Database. In case of low classification score, the system forwards the case and seeks assistance from agricultural expert teams for manual classification which are then sent to the farmer and stored in the Training Database. Low accuracy typically occurs if the user uploads an image with an underlying disease that is so far not known to the trained CNN model, or the

image quality is poor. Expert intervention in case of low classification score allows addition of new disease types which can be stored for future training runs. After the Training Database has sufficiently large number of images of the new disease category and a high classification accuracy is achieved, the Classifier can start recognizing the new disease automatically. Over time as more farmers collaborate and contribute images, it enables us to improve the accuracy for automated response to covered diseases, while using the limited expert resources to expand coverage for new diseases.

□ **Deep CNN Trainer** - This Cloud application is responsible for the more intensive work of training the neural network and builds the deep CNN model that is used by the Classifier to classify images into the correct disease types. This application is run asynchronously (without any interference to the Classifier) whenever the number of new images added to the Training Database goes beyond a pre-configured threshold. Each subsequent run of this training application works on a larger training dataset, and hence continually improves the deep CNN model used by the Classifier for more accurate disease classification. AWS was used to build the entire Cloud platform. The Disease Classifier and the Deep CNN Trainer are applications developed in Python. To make these Python applications accessible over mobile internet, they were developed using a web framework called FLASK and deployed behind an Apache Web Server running on an AWS EC2 machine (Ubuntu 16.04.2 LTS, 2 GiB memory, 8 GiB EBS volume). Disease Classifier and Deep CNN Trainer are built with TensorFlow [4], which is an open source library for Artificial Intelligence by Google.



Fig. 2. Expert dashboard with disease data visualizations

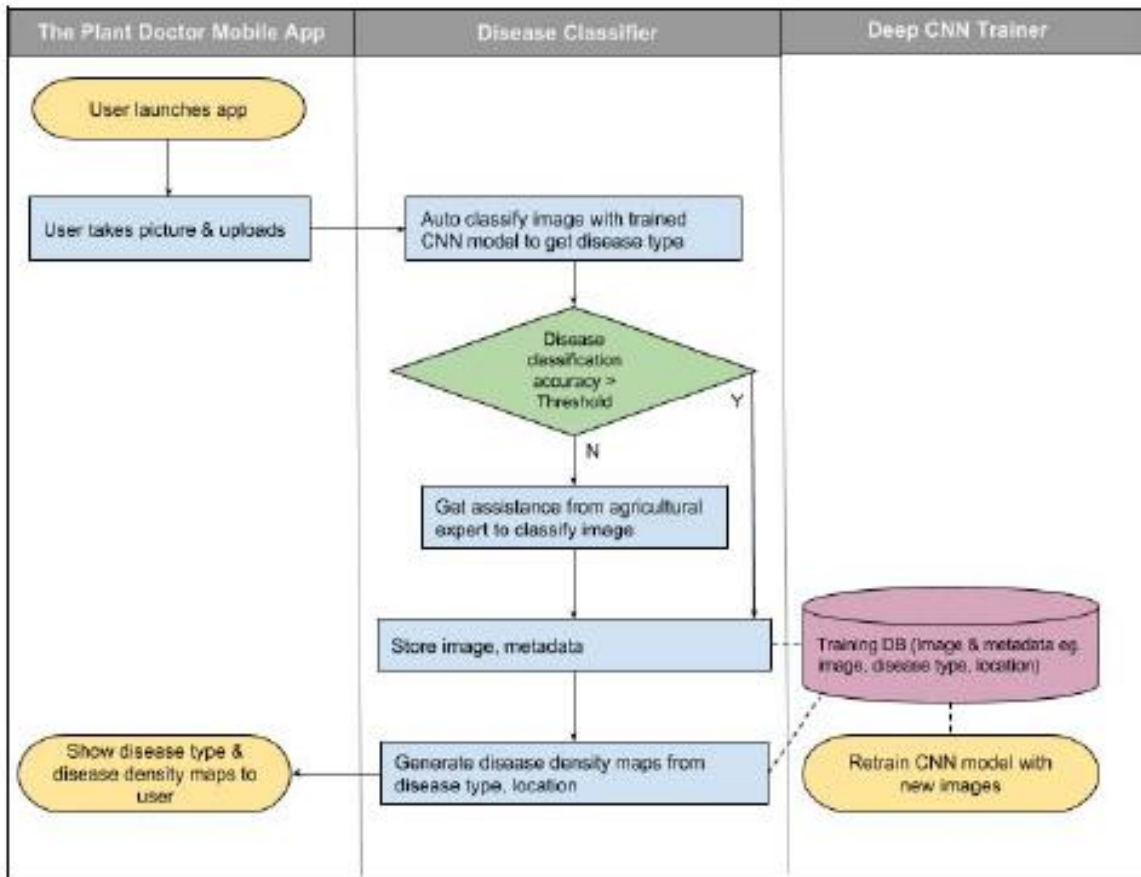


Fig. 3. Process flow of the components

2. FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

2.1.ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development

of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

2.2 TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

2.3 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3. SYSTEM DESIGN

3.1 UML DIAGRAMS:UML represents Unified Modeling Language. UML is an institutionalized universally useful showing dialect in the subject of article situated programming designing. The fashionable is overseen, and become made by way of, the Object Management Group.

The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components: a Meta-show and documentation. Later on, a few type of method or system can also likewise be brought to; or related with, UML.

The Unified Modeling Language is a popular dialect for indicating, Visualization, Constructing and archiving the curios of programming framework, and for business demonstrating and different non-programming frameworks.

The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks.

The UML is a essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

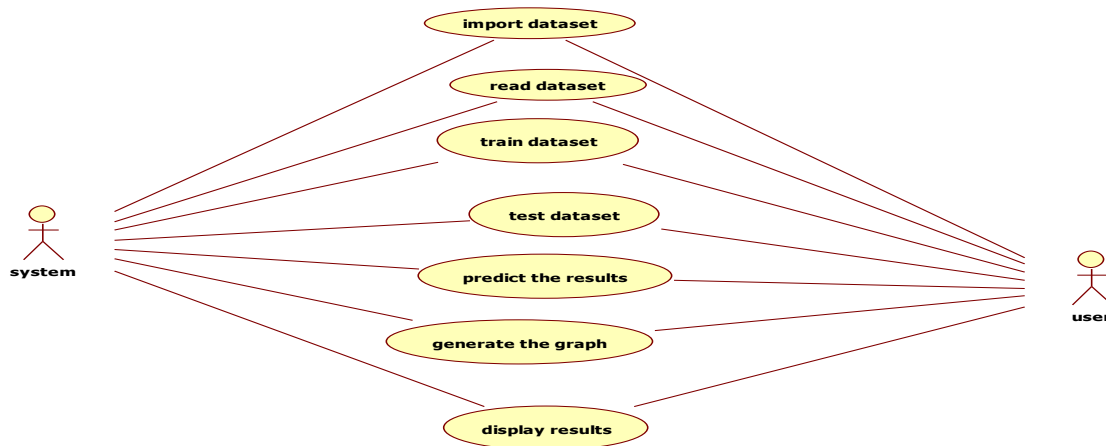
GOALS:

The Primary goals inside the plan of the UML are as in step with the subsequent:

1. Provide clients a prepared to-utilize, expressive visual showing Language on the way to create and change massive models.
2. Provide extendibility and specialization units to make bigger the middle ideas.
3. Be free of specific programming dialects and advancement manner.
4. Provide a proper cause for understanding the displaying dialect.
5. Encourage the improvement of OO gadgets exhibit.
6. Support large amount advancement thoughts, for example, joint efforts, systems, examples and components.
7. Integrate widespread procedures.

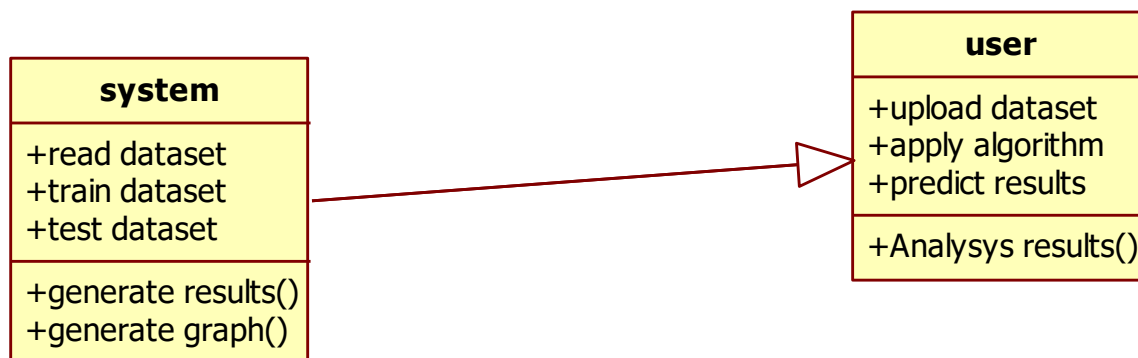
USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



4. CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



5.SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6.CONCLUSION

This paper presents an automated, low cost and easy to use end-to-end solution to one of the biggest challenges in the agricultural domain for farmers – precise, instant and early diagnosis of crop diseases and knowledge of disease outbreaks - which would be helpful in quick decision making for measures to be adopted for disease control. This proposal innovates on known prior art with the application of deep Convolutional Neural Networks (CNNs) for disease classification, introduction of social collaborative platform for progressively improved accuracy, usage of geocoded images for disease density maps and expert interface for analytics. High performing deep CNN model “Inception” enables real time classification of diseases in the

Cloud platform via a user facing mobile app. Collaborative model enables continuous improvement in disease classification accuracy by automatically growing the Cloud based training dataset with user added images for retraining the CNN model. User added images in the Cloud repository also enable rendering of disease density maps based on collective disease classification data and availability of geolocation information within the images. Overall, the results of our experiments demonstrate that the proposal has significant potential for practical deployment due to multiple dimensions – the Cloud based infrastructure is highly scalable and the underlying algorithm works accurately even with large number of disease categories, performs better with high fidelity real-life training data, improves accuracy with increase in the training dataset, is capable of detecting early symptoms of diseases and is able to successfully differentiate between diseases of the same family.

7. REFERENCES

- [1] L. Saxena and L. Armstrong, “A survey of image processing techniques for agriculture,” in *Proceedings of Asian Federation for Information Technology in Agriculture*, 2014, pp. 401-413.
- [2] E. L. Stewart and B. A. McDonald, “Measuring quantitative virulence in the wheat pathogen *Zymoseptoria tritici* using high-throughput automated image analysis,” in *Phytopathology* 104 9, 2014, pp. 985– 992.
- [3] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012.
- [4] TensorFlow.[Online].Available: <https://www.tensorflow.org/>
- [5] D. P. Hughes and M. Salathé, “An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing,” in *CoRR abs/1511.08060*, 2015.
- [6] S. Raza, G. Prince, J. P. Clarkson and N. M. Rajpoot, “Automatic detection of diseased tomato plants using thermal and stereo visible light images,” in *PLoS ONE*, 2015.
- [7] D. L. Hernández-Rabadán, F. Ramos-Quintana and J. Guerrero Juk, “Integrating soms and a bayesian classifier for segmenting diseased plants in uncontrolled environments,” 2014, in *the Scientific World Journal*, 2014.
- [8] S. Sankaran, A. Mishra, J. M. Maja and R. Ehsani, “Visible-near infrared spectroscopy for detection of huanglongbing in citrus orchards,” in *Computers and Electronics in. Agriculture* 77, 2011, pp. 127–134.

- [9] C. B. Wetterich, R. Kumar, S. Sankaran, J. B. Junior, R. Ehsani and L. G. Marcassa, "A comparative study on application of computer vision and fluorescence imaging spectroscopy for detection of huanglongbing citrus disease in the USA and Brazil," in *Journal of Spectroscopy*, 2013.
- [10] C. Szegedy, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818-2826.
- [11] Mango Diseases and Symptoms. [Online]. Available <http://vikaspedia.in/agriculture/crop-production/integrated-pestmanagment/ipm-for-fruit-crops/ipm-strategies-for-mango/mangodiseases- and-symptoms>
- [12] P. Subrahmanyam, S. Wongkaew, D. V. R. Reddy, J. W. Demski, D. McDonald, S. B. Sharma and D. H. Smith, "Field Diagnosis of Groundnut Diseases". *Monograph. International Crops Research Institute for the Semi-Arid Tropics*, 1992.

GROUNDWATER LEVEL PREDICTION USING HYBRID ARTIFICIAL NEURAL NETWORK WITH GENETIC ALGORITHM

Dhulipala Leela Sai Pavan Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

Abstract: - In recent years, the growth of the economy has led to the increasing exploitation of water resources and groundwater. Due to heavy abstraction of groundwater its importance increases, with the requirements at present as well as in future. Accurate estimates of groundwater level have a valuable effect in improving decision support systems of groundwater resources exploitation. This paper investigates the ability of a hybrid model of artificial neural network (ANN) and genetic algorithm (GA) in predicting groundwater levels in an observation well from Udipi district. The ground water level for a period of ten years and rainfall data for the same period is used to train the model. A standard feed forward network is utilized for performing the prediction task. A groundwater level forecasting model is developed using artificial neural network. The Genetic Algorithm is used to determine the optimized weights for ANN. This study indicates that the ANN-GA model can be used successfully to predict groundwater levels of observation well. In addition, a comparative study indicates that the ANN-GA hybrid model performs better than the traditional ANN back-propagation approach.

Keywords: Artificial neural network, feedforward network, genetic algorithm, ground water level, hybrid model.

1. INTRODUCTION

Groundwater is one of the major sources of supply for domestic, industrial and agricultural purposes. Estimation of groundwater level is very important in hydrogeology studies and aquifer management. In many cases, groundwater level fluctuations have resulted in damage to engineering structures [1]. With considerable amounts of these fluctuations, appropriatedecisions

can be presented in terms of hydrogeology, water quality and its management [2]. For this, a constant monitoring of the groundwater levels is extremely important. The water levels, if forecast well in advance, helps administrators to better plan the groundwater utilization. A continuous forecast of groundwater levels is required to effective use of any simulation model for water management and overall development [1]. In this regard, it is important to develop a fast and cost-effective method for aquifer simulation with an acceptable accuracy. Towards this goal, many researchers have used intelligent systems including, Coulibaly et al., Daliakopoulos et al., Lallahem et al., Dogan et al., Nourani et al, Yang et al., Sreekanth et al. These researchers used ANN for aquifer modelling in a variety of basins. ANN is an information-processing paradigm, that is inspired by the way biological nervous systems, such as the brain, processes information. It determines the relationship between inputs and outputs of physical systems by a network of interconnecting nodes adjusted by connecting weights based on the training samples, and extracts patterns and detects trends that are too complex to be noticed by either humans or other computational techniques. Neural networks take a different approach to problem solving than that of conventional computers. It has remarkable ability to learn and derive meanings from complicated and imprecise data. It has an ability to learn and apply the knowledge based on the data given for training or initial experience.

2.Literature review

A detailed review of artificial neural network applications can be found in Maier et al. They reviewed forty-three papers dealing with the use of neural network models for the prediction of water resources variables. In recent years, Nourani et al. [13] evaluate a hybrid of the ANN-Geostatic methodology for spatiotemporal prediction of groundwater levels in a coastal aquifer system. Jalalkamali and Jalalkamali employed a hybrid model of Artificial Neural Network and Genetic Algorithm (ANN-GA) for forecasting groundwater levels in an individual well. The hybrid ANN-GA model was designed to find an optimal number of neurons for hidden layers. Their research admitted the superiority of the ANN-GA model in prediction of groundwater levels. Taormina et al. employed an ANN for simulation of hourly groundwater levels in a coastal aquifer system. They confirmed that the developed feedforward neural network (FNN) can accurately reproduce groundwater depths of the shallow aquifer for several months. Moreover, a combined method of discrete wavelet transform method and different mother wavelets with

ANN (WANN) was proposed by Nakhaei and for the prediction of groundwater level fluctuations. Furthermore, a hybrid model of NeuroFuzzy Inference System with Wavelet (Wavelet- ANFIS) was proposed by Moosavi et al. for groundwater level forecasting in different prediction periods. These studies demonstrated that the wavelet transform can improve accuracy of groundwater level forecasting. The back-propagation algorithm (BP) is the most popular in the domain of neural networks, which is utilized in the most frequently mentioned studies for aquifers simulation. BP is the standard of the Gradient Descent algorithm (GDA). The gradient descent method, its algorithms, easily become stuck in local minimum and often need a longer training time.] showed the stochastic optimization method (GA) to train a FNN; therefore, numerical weights of neuron connections and biases represent the solution components of the optimization problem. In fact, a combination of genetic algorithm to adjust the neural network weights was proposed in several researches on artificial intelligence Montana Genetic Algorithm is one type of stochastic algorithms that is capable of solving multi-dimensional complex problems, especially non-smooth, noncontiguous, nondifferentiable objective function to find the global optimum, to escape the local optima and acquire a global optima solution. This combination would be an efficient method of training neural networks because, it takes advantage of the strengths of genetic algorithms and back propagation (the fast initial convergence of stochastic algorithms and the powerful local search of back propagation), and circumvents the weaknesses of the two methods (the weak fine-tuning capability of stochastic algorithms and a flat spot in back propagation). developed a Feed forward Neural Network coupled with Genetic Algorithm to simulate the rainfall field. The technique implemented to forecast rainfall for a number of times using hyetograph of recording rain gauges. The results showed that when Feed forward neural network coupled with Genetic Algorithm, the model performed better compared to similar work of using ANN alone. ANN applications in hydrology vary, from real time to event base modelling. They have been used for groundwater modelling, level estimation A comprehensive review of the applications of ANNs in hydrology can be found in the ASCE Task Committee report have systematically appraised the feat of the ANN model and the standard FNN trained with Levenberg algorithm, was tested for predicting groundwater level at Maheshwaram watershed, Hyderabad, India. The model competence and correctness were estimated according to the Root Mean Square Error (RMSE) and regression coefficient (R²). The model furnished the best fit and the forecast trend was hand in glove with the

experiential data. have competently conceived a technique to predict the monthly maximum, minimum, mean and cumulative precipitation totals within a period of the next four successive months, by means of ANNs. The precipitation datasets represent monthly totals recorded at four meteorological stations in Greece. For the appraisal of the outcomes and the competencies of the designed prognostic methods, suitable statistical indexes like the coefficient of determination (R²), the index of agreement (IA) and the RMSE were employed. The observations from this appraisal demonstrated that the technique of ANN furnishes ample precipitation totals in four successive months and these outcomes emerge as superior ones in relation to those gathered by means of traditional statistical methods.

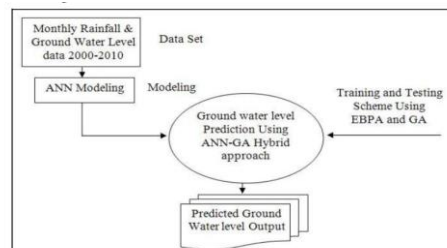


Figure 1. Methodology for the proposed model

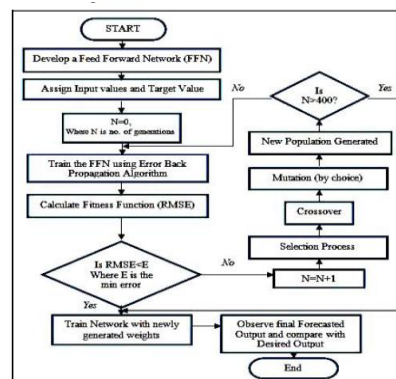


Figure 2. Flowchart of the Hybrid ANN-GA model

3. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

3.1 ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

3.2 TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3.3 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising

software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5. CONCLUSION

In this paper, two approaches of soft computing have been developed for predicting groundwater level in an observation well identified in Udupi district. Initially ANN modelling was carried out using feed forward neural network architecture to predict groundwater level. The inputs of the ANN model were monthly rainfall record and water level for period of 10 years. The hybrid ANN-GA model was developed and the results are compared with the ANN gradient descent algorithm. The performance of ANN and ANN-GA algorithms was evaluated. It is observed that the performance of ANN-GA is considered superior than ANN model. Thus, ANN-GA hybrid algorithm can be used for predicting ground water levels over the study area. Further, more investigations needed on the field generated data in groundwater level forecasting to have a precise statement.

6. REFERENCES

- [1] Sreekanth P.D., Geethanjali N., Sreedevi P.D, Shakeel Ahmed, Ravi Kumar N. and Kamala Jayanthi P.D., Forecasting groundwater level using artificial neural networks., Journal of Current Science, Vol. 96, No. 7, 2009.
- [2] Hosseini Z. and Nakhaei M., Estimation of ground water level using a hybrid genetic algorithm Neural network., Journal of pollution, Vol 1(1), Winter , pp.9-12, 2015.
- [3] Banerjee P., Prasad R.K. and Singh V.S., Forecasting of groundwater level in hard rock region using artificial neural network., Journal of Environmental Geology, 58(6),pp 1239-1246, 2009.
- [4] M.Nasseri, K.Asghari and M.J.Abedini, ,Optimized scenario for rainfall forecasting using GA coupled with artificial neural network, Science direct, Elsevier, Vol 35, pp. 1415-1421,2008.

[5] Coulibaly P, Anctil F, Aravena R, Bobee B., Artificial neural network modeling of water table depth fluctuations, *Water Resources Research*, 37(4): 885-896,2001.

[6] Lallahem, S., and Mania, J., Evaluation and forecasting of daily groundwater inflow in a small chalky watershed., *Hydrological Process.*, 17(8), 1561- 1577,2003.

[7] Chau, K. W., Application of a PSO-based neural network in analysis of outcomes of construction claims. *Automation*

FRAUD DETECTION IN ONLINE PRODUCT REVIEW SYSTEMS VIA HETEROGENEOUS GRAPH TRANSFORMER

Dodda Diana (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- In online product review systems, users are allowed to submit reviews about their purchased items or services. However, fake reviews posted by fraudulent users often mislead consumers and bring losses to enterprises. Traditional fraud detection algorithm mainly utilizes rule-based methods, which is insufficient for the rich user interactions and graph-structured data. In recent years, graph-based methods have been proposed to handle this situation, but few prior works have noticed the camouflage fraudster's behavior and inconsistency heterogeneous nature. Existing methods have either not addressed these two problems or only partially, which results in poor performance. Alternatively, we propose a new model named Fraud Aware Heterogeneous Graph Transformer (FAHGT), to address camouflages and inconsistency problems in a unified manner. FAHGT adopts a type-aware feature mapping mechanism to handle heterogeneous graph data, then implementing various relation scoring methods to alleviate inconsistency and discover camouflage. Finally, the neighbors' features are aggregated together to build an informative representation. Experimental results on different types of real-world datasets demonstrate that FAHGT outperforms the state-of-the-art baselines.

1. INTRODUCTION

Internet services have brought human beings with ecommerce, social networking, and entertainment platforms, which not only facilitate information exchange but also provide chances to fraudsters. Fraudsters disguise themselves as ordinary users to publish spam information or collect user privacy, compromising the interest of both platforms and users. In addition, multiple entities on the Internet are connected with multiple relationships. Traditional machine learning algorithms cannot handle this complicated heterogeneous graph data well. The current approach is to model the data as a heterogeneous information network so that similarities in characteristics and structure of

fraudsters can be discovered. Due to the effectiveness in learning the graph representation, graph neural networks (GNNs) have already been introduced into fraud detection areas including product review mobile application distribution cyber crime identification and financial services. However, most existing GNN based solutions just directly apply homogeneous GNNs, ignoring the underlying heterogeneous graph nature and camouflage node behaviors. This problem has drawn great attention with many solutions proposed. Graph Consis found that there are three inconsistency problems in fraud detection and further proposed two camouflage behaviors. These problems could be Camouflage: Previous work showed that



crowd workers could adjust their behavior to alleviate their suspicion via connecting to benign entities like connecting to highly reputable users, disguise fraudulent URLs with special or generate domain-independent fake reviews via generative language model to conceal their suspicious activities. Inconsistency: Two users with distinct interests could be connected via reviewing a common product such as food or movies. Direct aggregation makes GNNs hardly distinguish the unique semantic user pattern. Also, if a User r is suspicious, then the other one should be more likely to be distrustful if they are connected by common activity relation since fraudulent users tend to post many fraudulent reviews in the same short period. To address the above two problems, many methods have been proposed. Graph Consis addresses the inconsistency problem by computing the similarity score between node embeddings, which cannot distinguish nodes with different types. CAREGNN enhances GNN-based fraud detectors against camouflaged fraudsters by reinforcement learning based neighbor selector and relation aware aggregator. Its performance still suffers from the heterogeneous graph. In this paper, we introduce the Fraud Aware Heterogeneous Graph Transformer (FAHGT), where we propose heterogeneous mutual attention to address the inconsistency problem and design a label-aware neighbor selector to solve the camouflage problem. Both are implemented in a unified manner called the “score head mechanism”. We demonstrate the effectiveness and efficiency of FAHGT on many real world datasets. Experimental

results suggest that FAHGT can significantly improve KS and AUC over state-of-the-art GNNs as well as GNN-based fraud detectors. The advantages of FAHGT can be summarized as follows: – Heterogeneity: FAHGT is able to handle heterogeneous graphs with multi-relation and multi-node type without designing meta-path manually – Adaptability: FAHGT attentively selects neighbors given a noise graph from real-world data. The selected neighbors are either informative for feature aggregation or risky for fraud detection. Efficiency: FAHGT admits a low computational complexity via a parallelizable multi-head mechanism in relation scoring and feature aggregation. Flexibility: FAHGT injects domain knowledge by introducing a flexible relation scoring mechanism. The score of a relation connecting two nodes not only comes from direct feature interaction but is also constrained by domain knowledge.

2. EXISTING SYSTEM

For GNNs on spatial domain, samples a tree rooted at each node and computes the root’s hidden representation by hierarchically aggregating hidden node representations from the bottom to top. further proposes to learn in the spatial domain by computing different importance of neighbor nodes via the masked selfattention mechanism. All these methods are designed for homogeneous graphs. They cannot be directly applied to a heterogeneous graph with multiple types of entities and relations. In recent years, lots of heterogeneous GNN based methods have been developed. and Deep- transforms a heterogeneous graph into



several homogeneous graphs based on handcrafted meta-paths, applies GNN separately on each graph, and aggregates the output representations by attention mechanism. constructs meta-paths between nodes with the same object type. first samples a fixed number of neighbors via random walk strategy. Then it applies a hierarchical aggregation mechanism for intra-type and intertype aggregation. extends transformer architecture to heterogeneous graphs. They directly calculate attention scores for all the neighbors of a target node and perform aggregation accordingly without considering domain knowledge. For relation-aware graph fraud detectors, their main solution is to build multiple homogeneous graphs based on edge type information of the original graph then perform type independent node level aggregation and graph level concatenation learns weighting parameters for different homogeneous subgraph. both adopt attention mechanism in feature aggregation and SemiGNN further leverages a structure loss to guarantee the node embeddings homophily. Some works directly aggregate heterogeneous information in the graph. For instance, under a user-review-item heterogeneous graph, learns a unique set of aggregators for different node types and updates the embeddings of each node type iteratively.

Disadvantages

In the existing work, the system did not implement Fraud Aware Heterogeneous Graph Transformer(FAHGT) to measure frauds exactly.

This system is less performance due to lack of META RELATION SCORING.

3. PROPOSED SYSTEM

GraphConsis addresses the inconsistency problem by computing the similarity score between node embeddings, which cannot distinguish nodes with different types. CAREGNN enhances GNN-based fraud detectors against camouflaged fraudsters by reinforcement learning based neighbor selector and relation aware aggregator. Its performance still suffers from the heterogeneous graph. In this paper, the system introduces the Fraud Aware Heterogeneous Graph Transformer(FAHGT), where we propose heterogeneous mutual attention to address the inconsistency problem and design a label-aware neighbor selector to solve the camouflage problem. Both are implemented in a unified manner called the “score head mechanism”. We demonstrate the effectiveness and efficiency of FAHGT on many real world datasets. Experimental results suggest that FAHGT can significantly improve KS and AUC over state-of-the-art GNNs as well as GNN-based fraud detectors.

Advantages

The advantages of FAHGT can be summarized as follows.

Heterogeneity: FAHGT is able to handle heterogeneous graphs with multi-relation and multi-node type without designing meta-path manually.

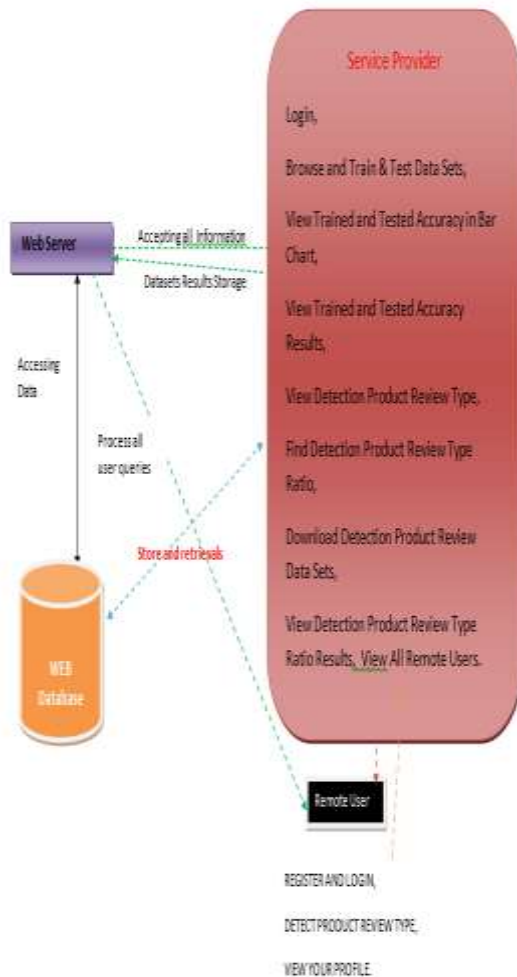
Adaptability: FAHGT attentively selects neighbors given a noise graph from real-world data. The selected neighbors are either

informative for feature aggregation or risky for fraud detection.

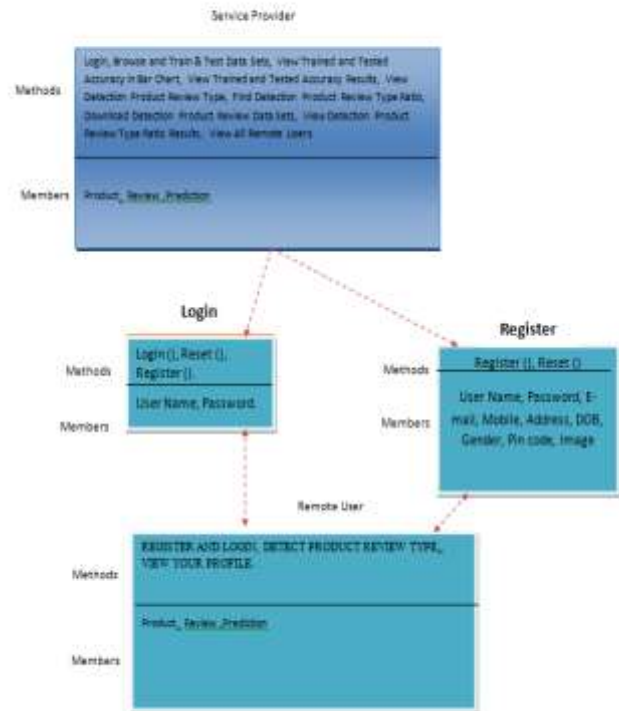
Efficiency: FAHGT admits a low computational complexity via a parallelizable multi-head mechanism in relation scoring and feature aggregation.

Flexibility: FAHGT injects domain knowledge by introducing a flexible relation scoring mechanism. The score of a relation connecting two nodes not only comes from direct feature interaction but is also constrained by domain knowledge.

Architecture Diagram



Class Diagram :



4. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine, address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

5. REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an



investigation being considered, the project request must be examined to determine precisely what the system requires. Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

6. FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

Operational Feasibility

Economic Feasibility

Technical Feasibility

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of

purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

6.1 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.



7. SYSTEM DESIGN AND DEVELOPMENT

7.1 INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations. This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design. Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases. Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent

pages after completing all the entries in the current page.

8. CONCLUSIONS

In this paper, we propose FAHGT, a novel heterogeneous graph neural network for fraudulent user detection in online review systems. To handle inconsistent features, we adopt heterogeneous mutual attention for automatic meta path construction. To detect camouflage behaviors, we design the label aware scoring to filter noisy neighbors. Two neural modules are combined in a unified manner called “score head mechanism” and both contribute to edge weight computation in final feature aggregation. Experiment results on real-world business datasets validate the excellent effect on fraud detection of FAHGT. The hyper-parameter sensitivity and visual analysis further show the stability and efficiency of our model. In summary, FAHGT is capable of alleviating inconsistency and discover camouflage and thus achieves state-of-art performance in most scenarios. In the future, we plan to extend our model in handling dynamic graphs data and incorporate fraud detection into other areas, such as robust item recommendation in E-commerce or loan default prediction in financial services.

9. REFERENCES

- [1] V. S. Tseng, J. Ying, C. Huang, Y. Kao, and K. Chen, “Fraudetector: A graph-mining-based framework for fraudulent phone call detection,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015, L. Cao, C. Zhang, T. Joachims, G. I. Webb, D. D. Margineantu,



- and G. Williams, Eds. ACM, 2015, pp. 2157–2166. [Online]. Available: <https://doi.org/10.1145/2783258.2788623>
- [2] J. Wang, R. Wen, and C. Wu, “Fdgars: Fraudster detection via graph convolutional networks in online app review system,” in WWW Workshops, 2019.
- [3] A. Li, Z. Qin, R. Liu, Y. Yang, and D. Li, “Spam review detection with graph convolutional networks,” in CIKM, 2019.
- [4] Z. Liu, Y. Dou, P. S. Yu, Y. Deng, and H. Peng, “Alleviating the inconsistency problem of applying graph neural network to fraud detection,” in SIGIR, 2020.
- [5] Y. Dou, Z. Liu, L. Sun, Y. Deng, H. Peng, and P. S. Yu, “Enhancing graph neural network-based fraud detectors against camouflaged fraudsters,” in CIKM, 2020.
- [6] R. Wen, J. Wang, C. Wu, and J. Xiong, “Asa: Adversary situation awareness via heterogeneous graph convolutional networks,” in WWW Workshops, 2020.
- [7] Y. Zhang, Y. Fan, Y. Ye, L. Zhao, and C. Shi, “Key player identification in underground forums over attributed heterogeneous information network embedding framework,” in CIKM, 2019.
- [8] D. Wang, J. Lin, P. Cui, Q. Jia, Z. Wang, Y. Fang, Q. Yu, and J. Zhou, “A semi-supervised graph attentive network for fraud detection,” in ICDM, 2019.
- [9] Z. Liu, C. Chen, X. Yang, J. Zhou, X. Li, and L. Song, “Heterogeneous graph neural networks for malicious account detection,” in CIKM, 2018.
- [10] Y. Dou, G. Ma, P. S. Yu, and S. Xie, “Robust spammer detection by nash reinforcement learning,” in KDD, 2020.
- [11] P. Kaghazgaran, M. Alfifi, and J. Caverlee, “Wide-ranging review manipulation attacks: Model, empirical study, and countermeasures,” in CIKM, 2019.
- [12] Z. Zhang, P. Cui, and W. Zhu, “Deep learning on graphs: A survey,” TKDE, 2020

PREGBOT: A SYSTEM BASED ON MI AND NIP FOR SUPPORTING WOMEN AND FAMILIES DURING PREGNANCY

Dola Vishali (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Artificial intelligence is transforming healthcare with a profound paradigm shift impacting diagnostic techniques, drug discovery, health analytics, interventions and much more. In this paper we focus on exploiting AI-based Pregbot systems, mainly based on machine learning algorithms and Natural Language Processing, to understand and respond to needs of patients and their families. In particular, we describe an application scenario for an AI-Pregbot delivering support to pregnant women, mothers, and families with young children, by giving them help and instructions in relevant situations.

1. INTRODUCTION

This work gives a general introduction to Pregbots by explaining what they are, what they can be used for and how to develop them. No previous domain-specific knowledge is required. Lately as of writing topics around Pregbots have received increasing attention from media and also numerous investments from different actors in the industry. At the same time not many potential users know about the existence of Pregbots or about areas in which Pregbots could be helpful assistance. The topic is equally unknown to developers. While the term Pregbot is commonly used in media, the meaning mostly remains ambiguous. There is a need for further explanation of what Pregbots are and further analysis to identify well suited applications for Pregbots. Additionally to spreading knowledge about the potentials of Pregbots and their use cases, more developers should be enabled to create new, innovative Pregbots. The lack of knowledge can be solved by providing answers to the questions of what Pregbots are, what benefits they bring and how to create them. An appropriate definition of Pregbots can be given by analyzing the fundamental meaning of the term Pregbot and by exploring past and current applications. Use cases of Pregbots can be identified in existing products. Market trends and attributes of media and technology can be analyzed to find new potential scenarios for the usage of Pregbots. Development is best explained by creating a

real Pregbot and by using it to present the general principles of the development process. Explaining what Pregbots are, demystifying what to use them for and presenting how to create them, will help more people to be able to use and create Pregbots, and thereby, accelerate the development of the Pregbot ecosystem. Innovation in technology and the creation of new solutions can help automating and simplifying more tasks, which gives people the opportunity to focus on more interesting issues and accomplish more things. Pregbots have the potential to simplify and automate many existing tasks and thereby accelerate the overall technological progress.

2. LITERATURE SURVEY

TITLE 1: A Pregbot for Perinatal Women's and Partners' Obstetric and Mental Health Care: Development and Usability Evaluation Study

AUTHOR: Kyungmi Chung , Orcid Image ; Hee Young Cho Orcid Image ; Jin Young Park , Orcid Image

The objectives of this study are to develop and evaluate a user-friendly question-and-answer (Q&A) knowledge database-based Pregbot (Dr. Joy) for perinatal women's and their partners' obstetric and mental health care by applying a text-mining technique and implementing contextual usability testing (UT), respectively, thus determining whether this medical Pregbot built on mobile instant messenger (KakaoTalk) can provide its male and female users with good user experience. Methods: Two men aged 38 and 40 years and 13 women aged 27 to 43 years in pregnancy preparation or different pregnancy stages were enrolled. All participants completed the 7-day-long UT, during which they were given the daily tasks of asking Dr. Joy at least 3 questions at any time and place and then giving the Pregbot either positive or negative feedback with emoji, using at least one feature of the Pregbot, and finally, sending a facilitator all screenshots for the history of the day's use via KakaoTalk before midnight. One day after the UT completion, all participants were asked to fill out a questionnaire on the evaluation of usability, perceived benefits and risks, intention to seek and share health information on the Pregbot, and strengths and weaknesses of its use, as well as demographic characteristics.

TITLE 2: Artificial Intelligence in Pregnancy: A Scoping Review AUTHOR M. C. Romero-Tenero ;Andreea Madalina Oprescu ;Gloria Miró Amarante Artificial Intelligence has been widely applied to a majority of research areas, including health and medicine. Certain

complications or disorders that can appear during pregnancy can endanger the life of both mother and fetus. There is enough scientific literature to support the idea that emotional aspects can be a relevant risk factor in pregnancy (such as anxiety, stress or depression, for instance). This paper presents a scoping review of the scientific literature from the past 12 years (2008-2020) to identify which methodologies, techniques, algorithms and frameworks are used in Artificial Intelligence and Affective Computing for pregnancy health and well-being. The methodology proposed by Arksey and O'Malley, in conjunction with PRISMA-ScR framework has been used to create this review. Despite the relevance that emotional status can have as a risk factor during pregnancy, one of the main findings of this study is that there is still not a significant amount of literature on automatic analysis of emotion. Health enhancement and well-being for pregnant women can be achieved with artificial intelligence or affective computing based devices, hence future work on this topic is strongly suggested

3.EXISTING SYSTEM

Pregbots receive increasing attention from media and industry, but at the same time it is not yet well known what Pregbots really are, what they can be used for and how to create them. The goal of this work is to answer these three questions by analyzing existing platforms, products and technologies, and additionally developing an exemplary Pregbot. Explaining what Pregbots are, demystifying what to use them for and showing how to create them will help more people to be able to use and create Pregbots and thereby accelerate the development of the Pregbot ecosystem. Starting by defining fundamental terms, the first half of the work focuses on showing available platforms, products and technologies, while the second half guides through the development of an exemplary Pregbot, including user interaction design and software architecture.

EXISTING SYSTEM ADVANTAGES:

It is used for general conversation not for the specific task.

4. PROPOSED SYSTEM

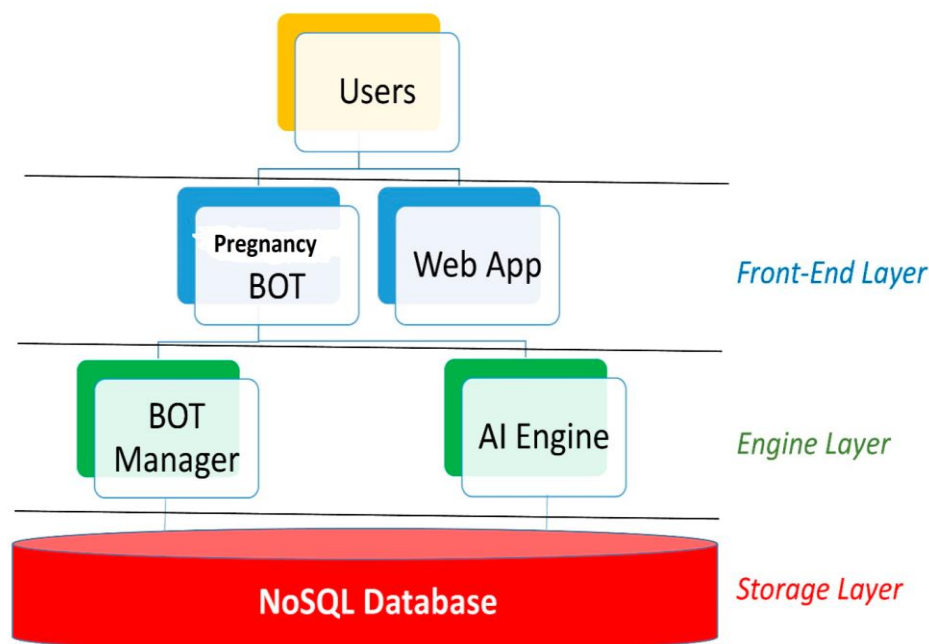
This is an automated chat robot design to answer users frequently asked questions, earlier natural language processing techniques were using to design this robots but its accuracy of giving correct answer was less and now due to Deep Learning algorithms accuracy of giving correct answer increase, so here using python deep learning project we are building PREGBOT application to answer users questions. To implement this technique first we train

deep learning models with the train data (all possible question's answers) and whenever users give any question then application will apply this test question on train model to predict exact answer for given question. Earlier companies were hiring humans to answer user's queries but by using this application we can answer user's question without using any man power. Chabot can be described as software that can chat with people using artificial intelligence. Chabot 's are generally used to respond quickly to users. Chabot's, a common name for automated conversational interfaces, present a new way for individuals to interact with computer systems. Traditionally, to get a question answered by a software program involves using a search engine, or filling out a form. A Chabot allows a user to simply ask questions in the same manner that they would address a human. There are many well-known voice-based chatbots currently available in the market: Google Assistant, Alexa and Siri. Chabot's are currently being adopted at a high rate on computer chat platforms. To implement this project we are using python deep learning neural networks and NLTK (natural Language Processing API) to process train and test text data.

PROPOSED SYSTEM ADVANTAGES:

It will be more useful for the pregnancy women.

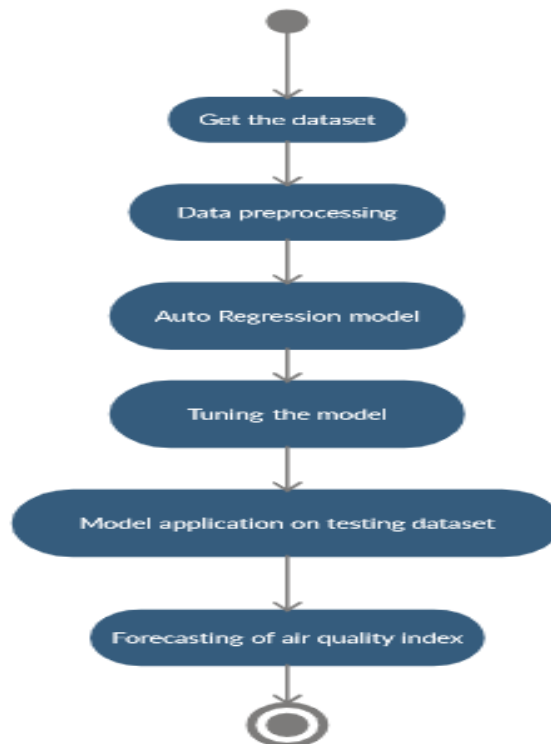
5. SYSTEM ARCHITECTURE



5.1 UML DIAGRAMS

ACTIVITY DIAGRAM

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes (i.e., workflows), well as the data flows intersecting with the related activities



6.TE

STINGS

6.1 Unit Testing

Unit testing includes the design of test belongings that validate that the internal program logic is operative properly, and that program input produces valid outputs. All choice branches and internal code flow should be authorized. It is the testing of separate software units of the request. It is done after the close of an individual unit before integration. This is a structural testing, that relies on data of its structure and is invasive. Unit tests achieve basic tests at factor level and test a specific commercial process, application, and/or system formation. Unit tests ensure that each single path of business process completes accurately to the documented provisions and contains clearly defined inputs and probable results.

6.2 Integration Testing

Integration tests are calculated to test integrated software components to regulate if they actually run as one program. Testing is occasion driven and is more concerned with the basic result of screens or fields. Integration tests validate that although the workings were individually approval, as shown by positively unit testing, the grouping of components is correct and dependable. Integration testing is specifically aimed at revealing the problems that arise from the grouping of components.

7.CONCLUSION

Pregbots can offer a lot of benefits in the mHealth domain both for healthcare providers and patients without having to download and install an app. All they have to do is chat with the bot to get relevant answers to their queries. Pregbots cannot replace humans but they can provide an interesting channel to support patients in delivering useful information and services through a simple conversation delivering personalized care while cutting down wait time. If well-designed and implemented, Pregbots can increase users' engagement and self-empowerment, by providing a better experience and save costs for the healthcare system (by reducing the number of unnecessary consultations). There are still several challenges in using Pregbots (e.g. conversations generally cannot be very complex and require increasing resources when expanding the Pregbot domain focus). Moreover, synonyms, hypernyms and hyponyms which are NLP and ontology challenges are among the complex limitations that most Pregbots suffer today. Other challenges are related to the privacy and security of the data collected. Pregbots indeed have to adhere to regulatory rules to avoid the exposure of patient information. In this paper we have presented an AI-based Pregbot useful to support and help pregnant women, mothers and families with young children about any doubts or problems that may incur during the pregnancy/childhood. The prototype needs further testing under real conditions, but its current status suggests that deployment will be straightforward. In general, we have found that as the intents grow, the use of similar words in different contexts can lead to a 7 <https://www.issalute.it/index.php> reduction in the accuracy of the system in identifying the specific intent.

8.REFERENCES

[1] PricewaterhouseCoopers LLP, 2016.

- [2] W. Hochfeld, J. Riffell, N. Levinson. Four trends that will transform healthcare in Europe in 2016. *European Pharmaceutical Review* 21(1) 2016.
- [3] A.S. Mosa, I. Yoo, L. Sheets, A systematic review of healthcare applications for smartphone, *BMC Med. Inform. Decis. Mak.* 2012;12:67.
- [4] L. Bellina, E. Missoni, Mobile cell-phones (M-phones) in telemedicine: Increasing connectivity of isolated laboratories, *Diagn. Pathol.* 2009;4:19.
- [5] L. Dayer, S. Heldenbrand, P. Anderson, P.O. Gubbins, B.C. Martin, Smartphone medication adherence apps: Potential benefits to patients and providers, *J. Am. Pharm. Assoc.* 2013; 53:172.
- [6] N. Tripp, K. Hailey, A. Liu, A. Poulton, M. Peek, J. Kim, R. Nanan, An emerging model of maternity care: Smartphone, midwife, doctor?, *Women Birth.* 2014;27:64–67.
- [7] A.P. Demidowich, K. Lu, R. Tamler, Z. Bloomgarden, An evaluation of diabetes self-management applications for Android smartphones, *J. Telemed. Telecare.* 2012;18:235–238.
- [8] A. Rao, P. Hou, T. Golnik, J. Flaherty, S. Vu, Evolution of data management tools for managing self-monitoring of blood glucose results: A survey of iPhone applications, *J. Diabetes Sci. Technol.* 2010;4:949–957.
- [9] S. Wallace, M. Clark, J. White, “It’s on my iPhone”: Attitudes to the use of mobile computing devices in medical education, a mixed-methods study, *BMJ Open.* 2012;2:e001099
- [10] K.E. Muessig, E.C. Pike, S. Legrand, L.B. Hightow-Weidman, Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: A review, *J. Med. Internet Res.* 2013;15:e1
- [10] K.E. Muessig, E.C. Pike, S. Legrand, L.B. Hightow-Weidman, Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: A review, *J. Med. Internet Res.* 2013;15:e1.

HEARSMOKING SMOKING DETECTION IN DRIVING ENVIRONMENT VIA ACOUSTIC SENSING ON SMARTPHONES

Donga Siva Sankar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Driving safety has drawn much public attention in recent years due to the fast-growing number of cars. Smoking is one of the threats to driving safety but is often ignored by drivers. Existing works on smoking detection either work in contact manner or need additional devices. This motivates us to explore the practicability of using smartphones to detect smoking events in driving environment. In this paper, we propose a cigarette smoking detection system, named HearSmoking, which only uses acoustic sensors on smartphones to improve driving safety. After investigating typical smoking habits of drivers, including hand movement and chest fluctuation, we design an acoustic signal to be emitted by the speaker and received by the microphone. We calculate Relative Correlation Coefficient of received signals to obtain movement patterns of hands and chest. The processed data is sent into a trained Convolutional Neural Network for classification of hand movement. We also design a method to detect respiration at the same time. To improve system performance, we further analyse the periodicity of the composite smoking motion. Through extensive experiments in real driving environments, HearSmoking detects smoking events with an average total accuracy of 93:44% in real-time.

I. INTRODUCTION

Driving state detection using smartphones. With the increase of public awareness about road safety, many works on driving state detection using smartphones emerge to improve the quality of daily driving. SenSpeed [11] is a system for accurate vehicle speed estimation, which can estimate vehicle speed by integrating the readings of accelerometers in smartphone. D3-Guard [13] proposes a drowsy driving detection system, which leverages audio sensors in smartphones, to detect drowsy actions and alert drowsy drivers. TEXIVE [14] uses smartphones to distinguish

drivers from passengers and detect texting operations during driving according to irregular and rich micro-movements of users. V-Sense [9] develops a vehicle steering detection middleware that can run on commodity smartphones to detect various vehicle maneuvers, including lane-changes, turns, and driving on curvy roads. Various kinds of works indicate the powerful capability of smartphones and embedded sensors. However, research about smoking detection in driving environment using smartphones is absent. This motivate us to propose HearSmoking to detect and alert drivers' smoking behavior.

1.1. Existing system

Smoking detection in contact manner. Technologies and studies on smoking detection using specialized devices, e.g., smart bracelets, smartwatches and chest belts, have been developed for some time. HLSDA [15] is a smoking detection algorithm, which collects various sensor data from a smartwatch and recognizes smoking behavior. PACT [16] is another wearable sensor system based on support vector machines. It detects smoking events by monitoring cigarette-to-mouth hand gestures in a contact manner. By capturing arm movements and breath puffs from 6-axis inertial sensors worn on two wrists of the user, puffMarker [17] builds a model based on 10-fold cross-validation to detect cigarette smoking. Another study [18] investigates the differences in brain signals of craving smokers, noncraving smokers, and non-smokers. This study uses data from resting-state EEG devices to train predictive models based on residual neural networks, and can distinguish the three groups. These works are all based on contact manner that need users to wear additional devices. Thus, they either cost high price to deploy or suffer inconvenience in daily using. Non-contact and device-free methods are needed for the smoking detection.

Smoking detection in non-contact manner. Non-contact methods are proposed by using civil cameras, gas sensors, Wi-Fi devices, etc. Smokey [8] is a smoking detection system that depends on Wi-Fi infrastructure. It leverages the smoking patterns leaving on Wi-Fi signals to identify the smoking activity even in the through-wall environments. A self-determined mechanism [19] is proposed to analyse smoking related events directly from videos by combining color re-projection techniques, Gaussian mixture models and hierarchical holographic modeling framework. Besides cameras, gas sensors are widely used. UbiLighter [20] detects cigarette smoking by using a gas sensor embedded in lighters to capture the gas from burning tobacco. A smoking monitoring method [21] uses a microphone to distinguish smoking breath from non-smoking breath. However, due to the great influence on accuracy that ambient noises would

bring by only using the microphone, the system is not suitable to be used in cars. The methods based on computer vision heavily depend on lighting and weather condition. Moreover, companies are not allowed to use the cameras and other recording devices due to privacy regulations in some areas [4]. Other methods based on specific sensors are costly or difficult to be deployed in cars.

HearSmoking. Different from existing works, HearSmoking detects drivers' smoking behaviors only using smartphones. HearSmoking can be applied in many ways. It can help to supervise the drivers of nosmoking vehicles, such as taxis and buses. In particular, HearSmoking is very suitable to be used in Uber and Lyft, since if a passenger complaint a smoking driver, it is easier for Uber and Lyft to obtain evidence from HearSmoking. It can also work with other systems to improve driving safety. For example, it can be integrated with the ubiquitous driving modes and navigation systems on the smartphones. Furthermore, if the detection results are uploaded to the transportation department, the police can further understand the driver's state when dealing with traffic accidents.

Disadvantages

An existing system that doesn't focus on improving the quality of daily driving by using smartphones emerge in quantity.

An existing system doesn't focus on multiple body movements when a driver is smoking during driving, e.g., steering with one hand, holding cigarette with another hand, putting up and down the cigarette, inhaling and exhaling smoke with chest expanding and shrinking.

1.2. Proposed System

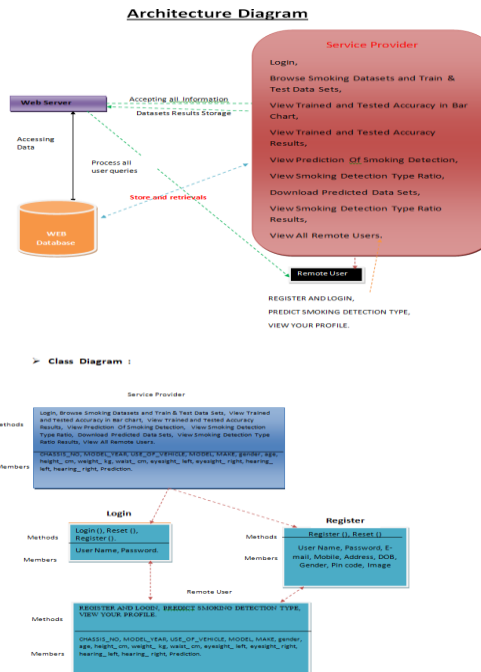
we propose a smoking detection system, named HearSmoking, which only uses acoustic sensors on smartphones in driving environments. We first analyse the smoking behaviors of 17 drivers, and find the typical smoking steps of drivers. To perceive motions, we let the smartphone speaker sends designed acoustic signals. The acoustic signals are reflected by surrounding objects and then received by the smartphone microphone. To get distances between reflectors and the smartphone, we calculate Relative Correlation Coefficient (RCC) of the collected data. Further, we get a set of sequence profiles from RCC profiles. Each sequence profile describes distance changes between moving objects and the smartphone over a period of time. According to our observations, when a driver is smoking, his/her main moving parts are hands and chest, so HearSmoking focuses on detecting movements of hands and chest. For hand movement

detection, we innovatively transform a sequence profile into a two-dimension image, and then send the image to a carefully designed Convolutional Neural Network (CNN) to identify whether there is a movement that matches the smoking hand movement pattern in the sequence profile. For chest movement detection, we perform Fast Fourier Transform (FFT) to find out waveforms in sequence profiles that fit human breath rate. Then a major breath path is selected to eliminate multipath interference. We analyse the amplitude and period of the waveform to determine whether there is a breath similar to smoking breath. If both hand movement and breath pattern fit the characteristic of those in a smoking event, we then analyse the periodicity of the detected composite motion to improve system performance. Finally, we get an analysis result whether the driver is smoking or not. To meet realistic demands, we collect training data using smartphones for 5 months to build the system model. We implement HearSmoking on different versions of Android platforms and comprehensively evaluate its performance in various environments. Experiment results show that HearSmoking is reliable and efficient in real driving environments. We study the unique patterns of smoking behaviors during driving. Based on our findings, we propose a smoking detection system, HearSmoking, which uses acoustic sensors embedded in smartphones to detect smoking events of drivers. To the best of our knowledge, we are the first to design a smoking detection system by only using smartphones. We divide the smoking detection into hand movement classification and respiration identification. We innovatively combine acoustic signal processing with CNN-based image classification into HearSmoking. After that, we design the methods of composite analysis and periodicity analysis to obtain the final detection result. We conduct extensive experiments in real driving environments. HearSmoking achieves an average total accuracy of 93.44% for smoking event detection.

Advantages

The system proposes a smoking detection system, named HearSmoking, which only uses acoustic sensors on smartphones in driving environments.

The system analyses the amplitude and period of the waveform to determine whether there is a breath similar to smoking breath. If both hand movement and breath pattern fit the characteristic of those in a smoking event, we then analyse the periodicity of the detected composite motion to improve system performance.



3. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

3.1 TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system

configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

4. CONCLUSION

In this paper, we address how to detect cigarette smoking activity during driving to improve road safety. Through literature survey and experimental verification, we find some characteristic smoking patterns in driving environment. We propose a smoking detection system, named Hear Smoking, which leverages acoustic sensors on smart phones to detect cigarette smoking events of drivers when they are driving. Hear Smoking takes advantages of RCC and CNN to detect both hand movements and respirations of the driver. Methods of composite analysis and periodicity analysis are designed to improve system performance. We conduct extensive experiments in different driving environments. Hear Smoking can detect smoking events with an average accuracy of 93:44% in real-time, which indicates that it works efficiently and reliably.

5. REFERENCES

- [1] D. L. Group, "Smoking while driving causes accidents in clearwater," <https://www.dolmanlaw.com/smoking-driving-causes-distracted-driving-accidents-clearwater-fl>, 2019.
- [2] S. Gupta and V. Kumar, "A study on effects of smoking on society: a case study," *MOJ Public Health*, vol. 7, no. 4, pp. 192–194, 2018.
- [3] "Smoke free law and vehicles," <http://www.smokefreeengland.co.uk/faq/vehicles/>, 2007.
- [4] Lyft, "Safety policies," <https://help.lyft.com/hc/en-us/articles/115012923127-Safety-policies#nosmoking>, 2018.
- [5] W.Wu and C. Chen, "Detection system of smoking behavior based on face analysis," in *ICGEC*, 2010, pp. 184–187.
- [6] R. W. Bukowski, R. D. Peacock, J. D. Averill, T. G. Cleary, W. D. Walton, P. A. Reneke, and E. D. Kuligowski, "Performance of home smoke alarms, analysis of the response of several available technologies in residential fire settings," National Institute of Standards and Technology, Tech. Rep., 2008.
- [7] M. Ahrens, "Home smoke alarms: The data as context for decision," *Fire Technology*, vol. 44, no. 4, pp. 313–327, 2008.
- [8] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, and Y. Liu, "Smokey: Ubiquitous smoking detection with commercial wifi infrastructures," in *IEEE INFOCOM*, 2016, pp. 1–9.

SLIDING WINDOW BLOCKCHAIN ARCHITECTURE FOR INTERNET OF THINGS

Gadde Ajay Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract- Internet of Things (IoT) refers to the concept of enabling Internet connectivity and associated services to nontraditional computers formed by integrating essential computing and communication capability to physical things for everyday usage. Security and privacy are two of the major challenges in IoT. The essential security requirements of IoT cannot be ensured by the existing security frameworks due to the constraints in CPU, memory, and energy resources of the IoT devices. Also, the centralized security architectures are not suitable for IoT because they are subjected to single point of attacks. Defending against targeted attacks on centralized resources is expensive. Therefore, the security architecture for IoT needs to be decentralized and designed to meet the limitations in resources. Blockchain is a decentralized security framework suitable for a variety of applications. However, blockchain in its original form is not suitable for IoT, due to its high computational complexity and low scalability. In this paper, we propose a sliding window blockchain (SWBC) architecture that modifies the traditional blockchain architecture to suit IoT applications. The proposed sliding window blockchain uses previous $(n - 1)$ blocks to form the next block hash with limited difficulty in Proof-of-Work. The performance of SWBC is analyzed on a real-time data stream generated from a smart home testbed. The results show that the proposed blockchain architecture increases security and minimizes memory overhead while consuming fewer resources.

Index Terms—Blockchain, Internet of Things, smart home, security, sliding window

1. INTRODUCTION

Blockchain is a distributed ledger used to record transactions between two or more parties. Unlike relational database systems, blockchain is a data structure where new entries get appended at the end of the ledger, and there exist no administrator permissions within a blockchain which allow modification of the data. Also, the addition of a new block to the chain needs to be verified by all other parties through a consensus algorithm. Since there exists a

distributed control over the blockchain, it is difficult for attackers to modify the data compared to a relational database system. Relational databases are primarily designed for centralized data storage and blockchain are specifically designed for decentralized data storage. There exist two types of blockchains: (i) permissioned and (ii) permissionless. A permissioned blockchain is a private blockchain which requires pre-verification of the participants within the network who are assumed to know each other whereas, a permissionless

blockchain is a public blockchain [1]. Traditional blockchain approach is not suitable for IoT with real-time data streams due to their computationally complex Proof-of-Work (PoW) [2]. As the computational time increases, blockchain security becomes infeasible to be used for IoT.

The two major challenges involved in applying blockchain to IoT environments include: (i) computational complexity and (ii) scalability. The computational complexity depends on

difficulty level and Merkle tree size. Merkle tree is a tree in which every leaf node is labeled with the hash of a transaction data and every non-leaf node is labeled with the cryptographic

hash of the labels of its child nodes. Merkle tree grows with the number of transactions made and, thereby, increasing the time consumed for Proof-of-Work, which is less favorable for

an IoT network. Scalability refers to the limits on the number of transactions a blockchain can process within a specific time period. Bitcoin is a popular example of a blockchain. Bitcoin blockchain is a payment system that does not rely on a central authority to secure and control its money supply. Each block in a Bitcoin blockchain has limited block size. In Bitcoin, the block size is limited to 1 MB and a block is mined every ten minutes. Interestingly, the existing literature [3] suggests blockchain as one of the data security and privacy algorithms that can be implemented for IoT applications due to its distributed architecture. In this paper, we propose a new blockchain architecture for

IoT environments, especially in the context of smart home applications. A smart home monitors, analyzes, and reports the state of the home. Smart homes use devices connected to IoT to automate and monitor in-home systems [4]. Smart home can be considered as the smallest unit of a smart city. The security standardization of a smart home supports a smart city and vice versa. In a smart home, the real-time data streams are generated

by sensors which help us to monitor the current status of the home, analyze energy consumption, and investigate any accidents inside a smart home. The volume of data generated by a smart home depends on the number of sensors deployed and the frequency of data acquisition. Therefore, proper sampling of sensor data is required to produce meaningful information which can be later stored in the blockchain. The volume of data stored in a blockchain decides the packet overhead, memory overhead, and computational overhead. In this context,

our proposed sliding window blockchain architecture tries to improve the security and reduce the memory overhead of IoT in a smart home environment.

ARCHITECTURE

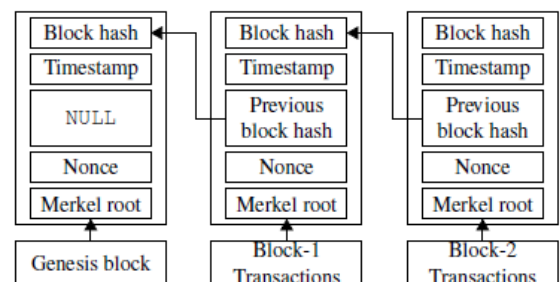


Figure 1: Blockchain architecture.

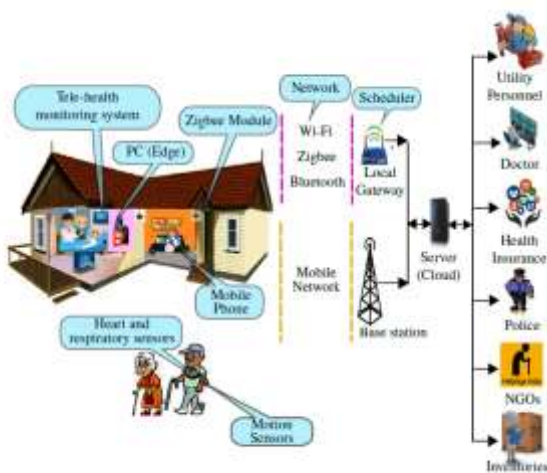


Figure 2: A typical smart home system for assisted living.

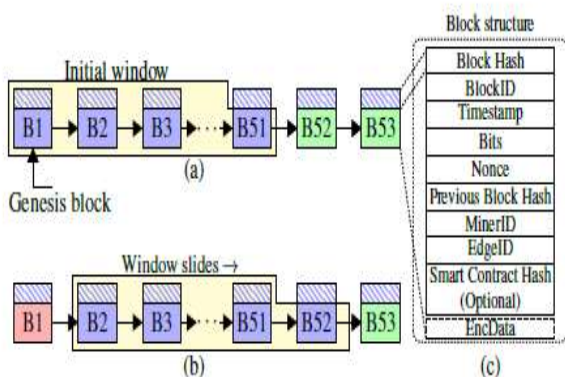


Figure 3: Sliding window blockchain.

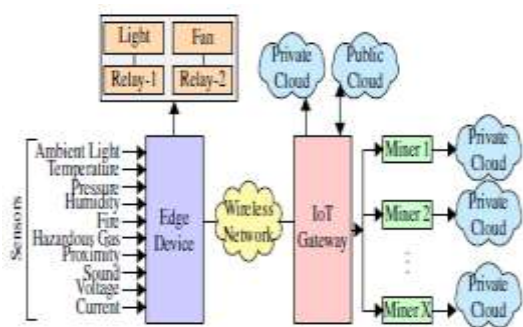


Figure 4: Smart home testbed used for studying sliding window blockchain.

2.SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth

with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY



The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

3.1 TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural

testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.



Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as

specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or –



one step up – software applications at the company level – interact without error.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

4.CONCLUSION

IoT devices face constraints on resources such as computational capability, energy sources, and memory. Therefore, the standard security algorithms are not feasible for IoT. We proposed a sliding window blockchain that meets the requirements of a resource constrained IoT network by reducing the memory overhead and limiting the computational overhead. The memory overhead is reduced by storing only a limited part of the blockchain, as defined by the sliding window size

in the IoT device and maintaining the whole blockchain in the private cloud. Computational overhead is limited by using the difficulty level between 1 and 5 and by eliminating the Merkle

tree. The security is increased by generating the block hash using the properties of n blocks in the sliding window. A false miner cannot mine a block unless he gets the previous $(n-1)$ blocks and the window size information. From the experimental results, we observed the following: (i) The computational time of PoW for each level of difficulty increases exponentially. (ii) The total block addition time increases with the increase in the number of miners in the group. (iii) As the window size increases, the hash computation time increases

linearly. (iv) A random selection of difficulty for each block in a blockchain reduces the total block addition time. Future work can be carried out to analyze the impact of a variable size sliding window. New consensus algorithms can be developed to suit the IoT environment. Furthermore, energy consumption of the blockchain can also be analyzed to draw more insights on energy resources required for an IoT device.

5.REFERENCES

- [1] S. Kulkarni, "The beauty of the blockchain," Open Source for You, vol. 06, pp. 22–24, June 2018.
- [2] T. M. F. Carames and P. F. Lamas, "A review on the use of blockchain for the Internet of Things," IEEE Access, vol. 6, pp. 32 979–33 001, May 2018.
- [3] A. Dorri, S. S. Kanhere, and R. Jurdak, "Blockchain in Internet of Things: challenges and solutions," arXiv preprint arXiv:1608.05187, August 2016.
- [4] IoT Agenda, "Smart home or building," April 2018. [Online]. Available: <https://internetofthingsagenda.techtarget.com/definition/smart-home-or-building>
- [5] L. Jiang, D. Y. Liu, and B. Yang, "Smart home research," in Proceedings of 2004 International Conference on Machine Learning and Cybernetics, vol. 2, August 2004, pp. 659–663.
- [6] theinstitute.ieee.org, "Towards a definition of the Internet of Things (IoT)," May 2015. [Online]. Available: [https://iot.ieee.org/images/files/pdf/IEEE IoT Towards Definition Internet of Things Revision1 27MAY15.pdf](https://iot.ieee.org/images/files/pdf/IEEE_IoT_Towards_Definition_Internet_of_Things_Revision1_27MAY15.pdf)
- [7] J. Wan, X. Gu, L. Chen, and J. Wang, "Internet of Things for ambient assisted



living: Challenges and future opportunities,” in International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), October 2017, pp. 354–357.

[8] D. Abbasinezhad-Mood, A. Ostad-Sharif, and M. Nikooghadam, “Novel anonymous key establishment protocol for isolated smart meters,” *IEEE Transactions on Industrial Electronics*, vol. 67, no. 4, pp. 2844–2851, April 2020.

[9] S. K. Das, D. J. Cook, A. Battacharya, E. O. Heierman, and T. Y. Lin, “The role of prediction algorithms in the MavHome smart home architecture,” *IEEE Wireless Communications*, vol. 9, no. 6, pp. 77–84, December 2002.

[10] C. Qu, M. Tao, J. Zhang, X. Hong, and R. Yuan, “Blockchain based credibility verification method for IoT entities,” *Security and Communication Networks*, vol. 2018, pp. 1–11, June 2018.

PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS

Gadde Kavya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT- Phishing is a social engineering cyberattack where criminals deceive users to obtain their credentials through a login form that submits the data to a malicious server. In this paper, we compare machine learning and deep learning techniques to present a method capable of detecting phishing websites through URL analysis. In most current state-of-the-art solutions dealing with phishing detection, the legitimate class is made up of homepages without including login forms. On the contrary, we use URLs from the login page in both classes because we consider it is much more representative of a real case scenario and we demonstrate that existing techniques obtain a high false-positive rate when tested with URLs from legitimate login pages. Additionally, we use datasets from different years to show how models decrease their accuracy over time by training a base model with old datasets and testing it with recent URLs. Also, we perform a frequency analysis over current phishing domains to identify different techniques carried out by phishers in their campaigns. To prove these statements, we have created a new dataset named Phishing Index Login URL (PILU-90K), which is composed of 60K legitimate URLs, including index and login websites, and 30K phishing URLs. Finally, we present a Logistic Regression model which, combined with Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction, obtains 96:50% accuracy on the introduced login URL dataset.

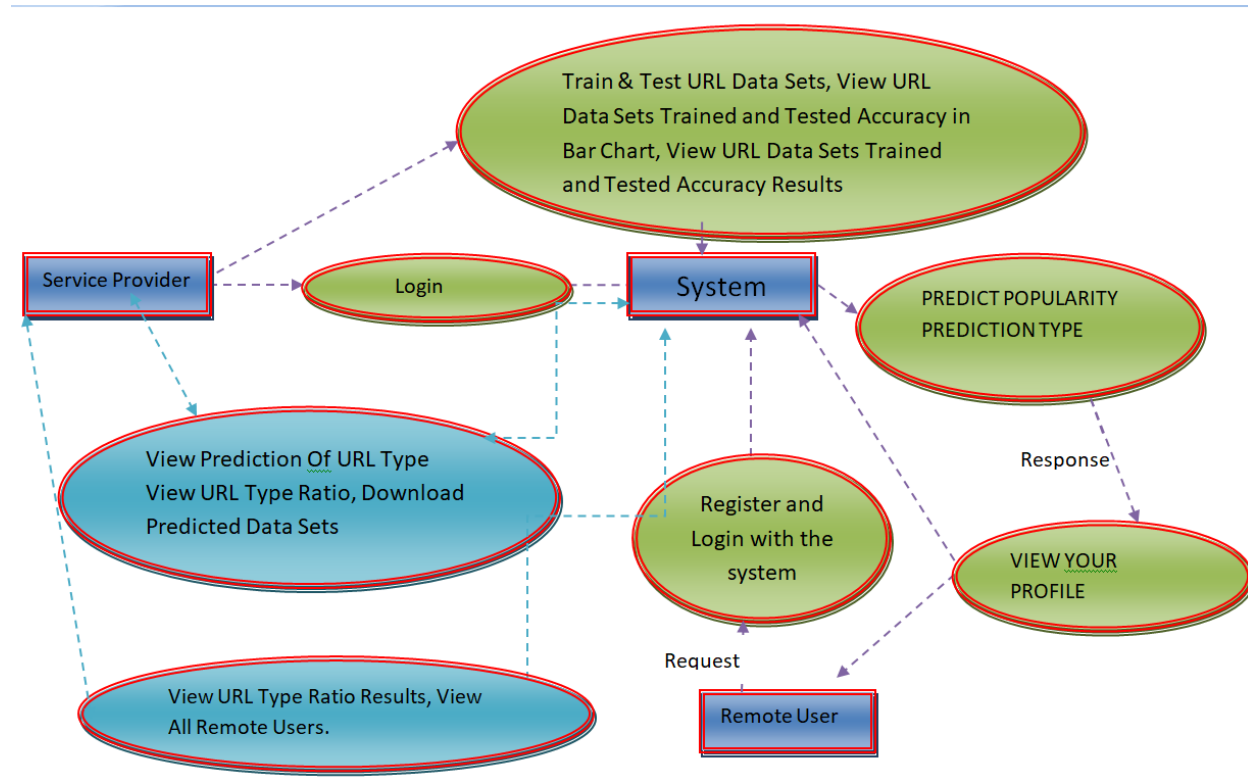
1. INTRODUCTION

In the last years, web services usage has grown drastically due to the current digital transformation. Companies motivate the change by providing their services online, like e-banking, e-commerce or SaaS (Software as a Service) Nowadays, due to the COVID-19 pandemic, restrictions have spread out the work-from-home model, which implies extra millions of workers, students, and teachers developing their activities remotely leading to a substantial additional workload for services such as email, student platforms, VPNs or company portals.

Therefore, there are even more potential targets exposed to phishing attacks, where phishers try to mimic legitimate websites to steal users' credentials or payment information. Recent studies concluded that phishing is one of the most significant attacks based on social engineering during the COVID-19 pandemic, together with spam emails and websites to execute these attacks. Identifying phishing sites through their HTTP protocol is no longer a valid rule. In the 3rd quarter of 2017 the APWG reported that less than 25% of phishing websites were hosted under HTTPS protocol, whilst this amount has increased up to 83% in 1st quarter of 2021. These websites provide secure end-to-end communication, which transmits a false safe impression to the user while making an online transaction. Furthermore, the Anti-Phishing Working Group has reported a significant increase in phishing attacks websites, just between the first quarter of 2020 and 2021 respectively. A reason behind this increase might be that people have resorted (and still are) to online services during the COVID-19 pandemic. One of the most popular solutions for phishing detection is the list-based approach, which analyzes the requested URL against a phishing database. Some examples of this solution are Google SafeBrowsing,¹ PhishTank,² OpenPhish³ or SmartScreen.⁴ If a requested URL matches any record, the request is blocked, and a warning is displayed to the user before visiting the website. However, despite the capabilities of the list-based approach, it would fail if the phishing URL was not reported previously and it will require a continuous effort to update the database with newer phishing data. Bell and Komisarczuk [11] observed that many phishing URLs were removed after day five from Phishtank while Open Phish removed all URLs after seven days from its report. This issue allows attackers to reuse the same URL when it is removed from different lists. Due to the mentioned drawbacks with the blacklist-based methods, automatic detection of phishing URLs based on machine learning, have attracted attention in research. These approaches can be grouped into four classes according to the type of data used for the detection: the text of the URL, the page content, the visual features and networking information. Methods based on the page content and visual features require visiting the website to collect the source code and render it, which is a time-consuming task. Other availability limitations can be found in studies that rely on networking and 3rd party information such as WHOIS or search engine rankings. To overcome these limitations, we focus on phishing detection through URLs since it implies advantages such as

fast computation -because no websites are loaded- and 3rd party and language independent, since features are extracted only from the URLs.

1.1 DATAFLOE DIAGRAM:



2.SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

3 .SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4.CONCLUSION

Phishing detection mechanism aims to improve current blacklist methods, protecting users from malicious login forms. Our work provides an updated dataset PILU-90K for researchers to train and test their approaches. This dataset includes legitimate login URLs which are the most representative scenario for real-world phishing detection. We explored several URL-based detection models using deep learning and machine learning solutions trained with phishing and legitimate home URLs. The main advantage of our approach is the low false-positive rate when classifying this type of URL. Among the different evaluated models, TFIDF combined with N-gram and LR algorithm obtained the best results with a 96:50% accuracy. In comparison with the current state-of-the-art, reviewed in Section II, our approach present three main advantages:

No dependence on external services. A limitation of the description methods that use features such as WHOIS domain age, page ranking on Google or Alexa or online blacklists, is their dependence on those services. Network slowdowns and service shortages can negatively impact analysis time, making real-time execution infeasible. Since phishing websites have a short lifespan [12], low detection times are required to warn users before accessing phishing websites.

Login website detection. Unlike other methods, which are trained with homepage URLs as representatives of the legitimate class, our model was trained with legitimate login websites. This ensures the correct classification of those websites. Therefore, our approach can be applied to the real case scenario where users have to predict whether a login form page is legitimate or phishing.

Updated and real-world dataset. PLU-60K is focused on using updated legitimate login URLs. As demonstrated, models trained with old datasets were not able to endure their performance

over time. We provide an updated phishing URL dataset for models to learn from nowadays phishing URLs and trends, which are crucial for real-world performance. We demonstrated that phishing URL detection systems trained with legitimate land page URLs fail to classify legitimate login URLs correctly. The best-tested models could only classify 69:50% of these URLs correctly, which implies a high false-positive rate. For this reason, we recommend that a phishing detector, which intends to be used in a real situation, should be trained using *legitimate login* websites (such as PLU-60K) instead of homepages. The main drawback of using login websites for training is that, due to the similarity between phishing and legitimate samples, overall accuracy is slightly reduced. The tradeoff against the state-of-the-art methods is still fair due to their high false-positive rate. Different categories for current phishing attacks were identified by using a domain frequency analysis. While standalone and compromised domains were the most common approaches, free hosting services, cloud web servers and malware blog posts represent many current phishing attacks due to their cost and effectiveness for phishing campaigns. Finally, we demonstrated that machine learning models using handcrafted URL features decreased their performance over time, up to 10:42% accuracy in the case of the Light GBM algorithm from the year 2016 to 2020. For this reason, machine learning methods should be trained with recent URLs to prevent substantial ageing from the date of its release. In the future, we will add more information about the samples into the analysis, such as the source code of the website and a screenshot of its content, which could be useful to increase the phishing detection performance. In addition, we will enlarge our dataset, including such information. Finally, observing that deep learning techniques and automatic feature extraction obtained promising results over traditional feature extraction, we intend to explore different URL codifications to improve.

5. REFERENCES

- [1] Statista. (2020). *Adoption Rate of Emerging Technologies in Organizations Worldwide as of 2020*. Accessed: Sep. 12, 2021. [Online]. Available: <https://www.statista.com/statistics/661164/worldwide-cio-surveyoperati%onal-priorities/>
- [2] R. De', N. Pandey, and A. Pal, "Impact of digital surge during COVID- 19 pandemic: A viewpoint on research and practice," *Int. J. Inf. Manage.*, vol. 55, Dec. 2020, Art. no. 102171.

- [3] P. Patel, D. M. Sarno, J. E. Lewis, M. Shoss, M. B. Neider, and C. J. Bohil, "Perceptual representation of spam and phishing emails," *Appl. Cognit. Psychol.*, vol. 33, no. 6, pp. 1296_1304, Nov. 2019.
- [4] J. A. Chaudhry, S. A. Chaudhry, and R. G. Rittenhouse, "Phishing attacks and defenses," *Int. J. Secur. Appl.*, vol. 10, no. 1, pp. 247_256, 2016.
- [5] M. Hijji and G. Alam, "A multivocal literature review on growing social engineering based cyber-attacks/threats during the COVID-19 pandemic: Challenges and prospective solutions," *IEEE Access*, vol. 9, pp. 7152_7169, 2021.
- [6] A. Alzahrani, "Coronavirus social engineering attacks: Issues and recommendations," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 154_161, 2020.
- [7] *Phishing Activity Trends Report 3Q*, Anti-Phishing Working Group, International, 2017. Accessed: Sep. 12, 2021.
- [8] *Phishing Activity Trends Report 1Q*, Anti-Phishing Working Group, International, 2021. Accessed: Sep. 14, 2021.
- [9] R. Chen, J. Gaia, and H. R. Rao, "An examination of the effect of recent phishing encounters on phishing susceptibility," *Decis. Support Syst.*, vol. 133, Jun. 2020, Art. no. 113287.
- [10] *Phishing Activity Trends Report 4Q*, Anti-Phishing Working Group, International, 2020. Accessed: Sep. 12, 2021.
- [11] S. Bell and P. Komisarczuk, "An analysis of phishing blacklists: Google safe browsing, OpenPhish, and PhishTank," in *Proc. Australas. Comput. Sci. Week Multiconf.*, Feb. 2020, pp. 1_11.
- [12] A. Oest, Y. Safaei, P. Zhang, B. Wardman, K. Tyers, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, "Phishtime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 379_396.

SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Gadiraju Sowmya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract-Sign language is a lingua among the speech and hearing-impaired community. It is hard for most people who are not familiar with sign language to communicate without an interpreter. Sign language recognition appertains to track and recognize the meaningful motion of human made with head, arms, hands, fingers, etc. The technique that has been implemented here, transcribes the gestures from sign language to a spoken language which is easily understood by the listening. The gestures that have been translated include alphabets, words from static images. This becomes more important for the people who completely rely on gestural sign language for communication tries to communicate with a person who does not understand the sign language. Most of the systems that are under use face a recognition problem with the skin tone, by introducing a filter it will identify the symbols irrespective of the skin tone. The aim is to represent features that will be learned by a system known as convolutional neural networks (CNN), which contains four types of layers: convolution layers, pooling/subsampling layers, nonlinear layers, and fully connected layers.

Keywords: Sign language, CNN, training, dataset, filters

1. INTRODUCTION

A sign language interpreter is a significant step toward improving contact between the deaf and the general population. Sign language is a natural language used by hearing and speech impaired people to communicate. It uses hand gestures instead of sound to convey messages or information. Sign language can vary from one part of the world to another. Due to this, people find difficulty in communicating with normal people because normal people cannot understand sign languages. There arises a need for sign philological translators, which can translate sign language to spoken language. However, the

availability of translators is limited when considering the sign language translators and these translators have many limitations. This led to the development of a sign language recognition system, which can automatically translate sign language into the text as well as a speech by effective pre-processing and accurate classification of the signs. According to recent developments in the area of deep learning, neural networks may have far-reaching implications and implementations for sign language analysis. In the proposed system, Convolutional Neural Network (CNN) is used to classify images of sign language because convolutional networks are faster in feature extraction and classification of images over other classifiers. The environment may also recognise a sign as a compression technique for information transmission, which is then reconstructed by the receiver. The signs are divided into two categories: static and dynamic signs. The movement of body parts is frequently included in dynamic signs. Depending on the meaning of the gesture, it may also include emotions. Depending on the situation of the context, the gesture may be widely classified as:

- Arm gestures
- Facial / Head gestures
- Body gestures

2. EXISTING SYSTEM

This research suggested the use of filters in the sign language translation algorithm because the existing system has low accuracy as it faced issues with skin tone identification. Sign language conversion can reach a maximum of 96% of accuracy but achieving that can be a tedious task. The current system failed to obtain this accuracy as it lagged to identify the skin tone under the low light areas.

3. PROPOSED SYSTEM

In this article, the filtering of images plays an important role. It improves the accuracy of identifying the symbols even in low light areas. Before the process of saturation and grey scaling the image is sent to the filtering system where it tries to find the symbol shown in the hands, after recognizing the symbol the image is further processed and final result which is the word is obtained.

4. BACKGROUND AND RELATED WORK

Building efficient sign language processing systems necessitates both a knowledge of Deaf culture and knowledge of sign languages in order to construct systems that account for sign

languages' diverse linguistic aspects. This paper outlines the context and addresses recent reviews of sign language processing that do not take a systematic approach to the issue.

5. IMAGE RECOGNITION

The software requirements of this system are python, Open Source Computer Vision Library (OpenCV), TensorFlow, and NumPy. As python is the fastest when compared to other languages, it is used by this system. TensorFlow is an open-source machine learning tool that is used to train the sign images from start to finish. This framework makes use of OpenCV is a free and open-source software library for computer vision and machine learning. Numpy is a Python library that adds support for big, multidimensional arrays and matrices, as well as a wide range of high-level mathematical functions that can be used for them. [1]. The proposed system not only recognizes the digits and alphabets but also recognizes the words of sign language. Background elimination is done by giving a range that lies between the color range of the human hand. Thus this dynamically recognizes the hand region. The lighting condition problem is corrected by using the color of the human hand.

6. IMAGE RECOGNITION SYSTEM

The acquisition of sign data is the first step in the sign language recognition system. The data can be obtained in a variety of ways.

7 IMPLEMENTATION

The above diagram is the basic architecture that was proposed for the system. This consists of the basic system modules such as image acquisition, pre-processing, feature extraction, and finally producing output based on pattern recognition

8. CONCLUSION

The proposed system successfully predicts the signs of sign and some common words under different lighting conditions and different speeds. Accurate masking of the images is being done by giving a range of values that could detect human hand dynamically. The proposed system uses CNN for the training and classification of images. For classification and training, more informative features from the images are finely extracted and being used. A total of 1750 static images for each sign is used for training to get the accurate output. Finally, the output of the recognized sign is shown in the form of text as well as converted into speech. The system is

capable of recognizing 125 words including alphabets. Thus this is a userfriendly system that can be easily accessed by all the deaf and people

9.REFERENCES

- [1] S. C. W. Ong and S. Ranganath, —*Automatic sign language analysis: A survey and the future beyond lexical meaning*,|| IEEE Trans. Pattern Anal. Mach. Intell., vol. **27**, no. 6, pp. 873–891, Jun. 2005.
- [2] L. Ding and A. M. Martinez, —*Modelling and recognition of the linguistic components in American sign language*,|| Image Vis. Comput., vol. **27**, no. 12, pp. 1826– 1844, Nov. 2009.
- [3] D. Kelly, R. Delannoy, J. Mc Donald, and C. Markham, —*A framework for continuous multimodal sign language recognition*,|| in Proc. Int. Conf. Multimodal Interfaces, Cambridge, MA, 2009, pp. 351–358
- [4] G. Fang, W. Gao, and D. Zhao, —*Large vocabulary sign language recognition based on fuzzy decision trees*,|| IEEE Trans. Syst., Man, Cybern. A Syst. Humans, vol. **34**, no. 3, pp. 305–314, May 2004.
- [5] Haldorai, A. Ramu, and S. Murugan, Social Aware Cognitive Radio Networks, Social Network Analytics for Contemporary Business Organizations, pp. 188–202. doi:10.4018/978-1-5225-5097-6.ch010
- [6] R. Arulmurugan and H. Anandakumar, Region-based seed point cell segmentation and detection for biomedical image analysis, International Journal of Biomedical Engineering and Technology, vol. **27**, no. 4, p. 273, 2018.
- [7] N. Purva, K. Vaishali, *Indian Sign language Recognition: A Review*, IEEE proceedings on International Conference on Electronics and Communication Systems, pp. 452-456, 2014.
- [8] F. Pravin, D. Rajiv, HASTA MUDRA *An Interpretation of Indian Sign Hand Gestures*, 3rd International conference on Electronics Computer technology, vol. **2**, pp.377-380, 2011



Block Hunter: Federated Learning for Cyber Threat Hunting in Blockchain-based IIoT Networks

Ganesna Jaya Datta Sri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Nowadays, blockchain-based technologies are being developed in various industries to improve data security. In the context of the Industrial Internet of Things (IIoT), a chain-based network is one of the most notable applications of blockchain technology. IIoT devices have become increasingly prevalent in our digital world, especially in support of developing smart factories. Although blockchain is a powerful tool, it is vulnerable to cyber attacks. Detecting anomalies in blockchain-based IIoT networks in smart factories is crucial in protecting networks and systems from unexpected attacks. In this paper, we use Federated Learning (FL) to build a threat hunting framework called Block Hunter to automatically hunt for attacks in blockchain-based IIoT networks. Block Hunter utilizes a cluster-based architecture for anomaly detection combined with several machine learning models in a federated environment. To the best of our knowledge, Block Hunter is the first federated threat hunting model in IIoT networks that identifies anomalous behavior while preserving privacy. Our results prove the efficiency of the Block Hunter in detecting anomalous activities with high accuracy and minimum required bandwidth. Index Terms—Federated Learning, Anomaly Detection, Threat Hunting, Blockchain, Industrial Internet of Things, IIoT, IoT.

1. INTRODUCTION

THE technological trajectory of block chain makes it a valuable tool in many areas, including healthcare, military, finance and networking, via its immutable and tamperproof data security advantages. With the ever-increasing use of Industrial Internet of Things (IIoT) devices, the world is inevitably becoming a smarter interconnected environment; especially factories are becoming more intelligent and efficient as technology advances [1]. IIoT is considered a subcategory of the Internet of Things (IoT). There are, however,

differences between IOT and IIOT in terms of security requirements. While the IIOT makes consumers' lives easier and more convenient, the IIOT aims to increase production safety and efficiency. IIOT devices are mainly used in B2B (business-to-business) settings, while IOT devices are mostly considered in B2C (business-to-consumer) environments. This would lead to a different threat profile for IIOT networks compared to their IOT counterparts where device-to-device transactions are of utmost importance.



IIOT networks provide an umbrella for supporting many applications and arm us to respond to users' needs, especially in an industry setting such as smart factories [1]. Block chain technology advantages lead to its wide adoption in IIOT based networks such as smart factories, smart homes/buildings, smart farms, smart cities, connected drones, and healthcare systems [1], [2]. While the focus of this paper is on the security of block chain-based IIOT networks in smart factories [3], [4], the suggested framework may be used in other IIOT settings as well.

In modern smart factories, many devices are connected to the public networks, and many activities are supported by smart systems such as temperature monitoring systems, Internet-enabled lights, IP cameras, and IP phones. These devices are storing private and sensitive data and may offer safety-critical services [3], [1]. As the number of IIOT devices in smart factories increases, the main issue will be storing, collecting, and sharing data securely. Industrial, critical, and personal data are therefore at risk in such a situation. Block chain technology can ensure data integrity inside and outside of smart factories through strong authentication and ensure the availability of communication backbones. Despite this, privacy and security issues are significant challenges in IIOT [3], [4]. The probability of fraudulent activity occurring in block chain-based networks [2], [4] is an important issue. Even though block chain technology is a powerful tool, it is not protected from cyber attacks either. For example, a 51% cyber attack [2]

on Ethereum Classic, and three consecutive attacks in August of 2020 [5], which resulted in the theft of over \$5M worth of crypto currency, have exposed the vulnerabilities of this block chain network.

Smart factories should protect users' data privacy during transmission, usage, and storage [4]. Stored data are vulnerable to tampering by fraudsters seeking to access, alter or use the data with malicious motives. Statistically speaking, these attacks can be viewed as anomalous events, exhibiting a strong deviation from usual behavior [2], [6]. Detecting out-of-norm events are essential for threat hunting programs and protecting systems from unauthorized access by automatically identifying and filtering anomalous activities. [6], [7].

The main objective of this paper is to detect suspicious users and transactions in a block chain-based IIOT network specifically for smart factories. Here, abnormal behavior serves as a proxy for suspicious behavior as well [4]. By identifying outliers and patterns, we can leverage Machine Learning (ML) algorithms to identify out-of-norm patterns to detect attacks and anomalies on block chain. Because deep neural networks learn representations automatically from data that they are trained on, they are the candidate solution for detecting anomalies [4], [7]. However, there are challenges with any ML and deep learning-based anomaly detection techniques. These methods suffer from training data scarcity problems, and privacy issues [7].



A novel and practical approach would be to employ Federated learning (FL) models to detect anomalies while preserving data privacy, and monitoring data quality [7], [9]. FL allows edge devices to collaborate during the training stage while all data stays on the device. We can train the model on the device itself instead of sending the data to another place, and only the updates of the model are shared across the network.

FL has become a trend in ML where smart edge devices can simultaneously develop a mutual prediction between each other [7], [10]. In addition, FL ensures multiple actors construct robust machine learning models without sharing data, addressing fundamental privacy, data security, and digital rights management challenges. Considering these characteristics, this paper uses an FL-based anomaly-detection framework called Block Hunter capable of detecting attack payloads in block chain-based IIOT networks

The main contributions of the paper are summarized as follows:

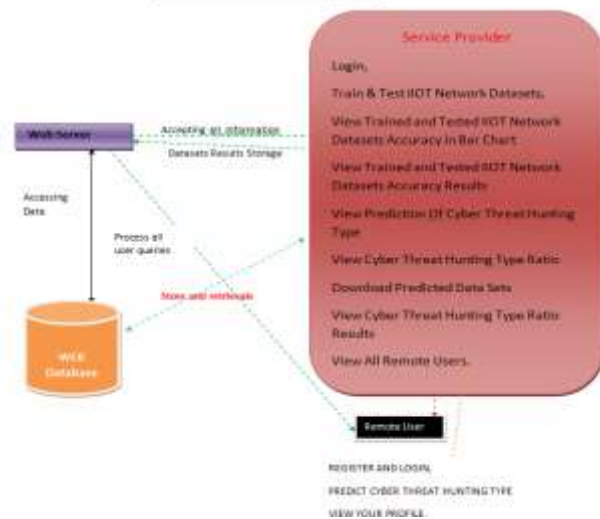
- 1) Utilize a cluster-based architecture to formulate an anomaly detection problem in block chain-based smart factories. The cluster-based approach increase hunting efficiency in terms of bandwidth reduction and throughput in IIoT networks.
- 2) Apply a federated design model to detect anomalous behaviour in IIoT devices related to blockchain-based smart factories. This provides a privacy-preserving feature when using machine learning models in a federated framework.

3) Implementation of various anomaly detection algorithms such as clustering-based, statistical, subspace-based, classifier-based, and tree-based for efficient anomaly detection in smart factories.

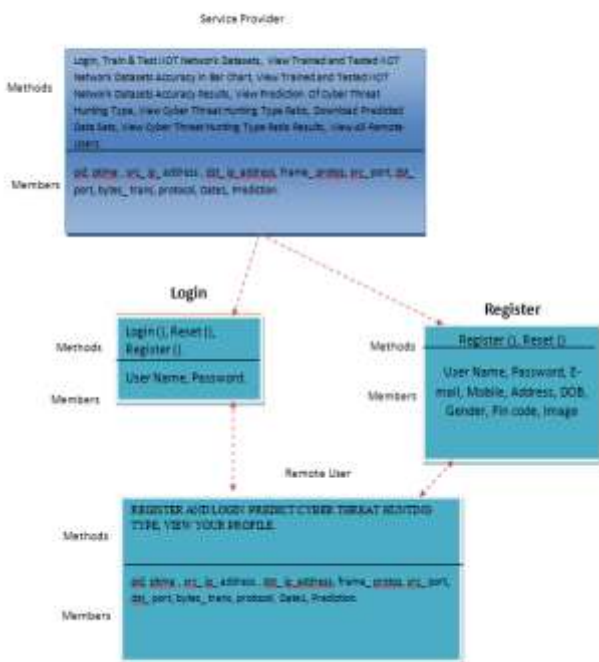
4) The impact of block generation, block size, and miners on the Block Hunter framework are considered. Moreover, the performance measurements like Accuracy, Precision, Recall, F1-score, and True Positive Rate (TPR) anomaly detection are discussed.

Here is a breakdown of the rest of the paper. Section II discusses anomaly detection works in the block chain and FL. Section III describes the Block Hunter framework and presents the network model and topology design. In Section IV, methodology and machine learning approaches to identify anomalies are discussed. In Section V, we present the assessment of the Block Hunter framework. Finally, In Section VI, we conclude the paper and point out future work directions.

Architecture Diagram



➤ Class Diagram :



2. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

3. CONCLUSION

In this paper, we developed the Block Hunter framework to hunt anomalies in

block chain-based IIOT smart factories using a federated learning approach. Block Hunter uses a cluster-based architecture to reduce resources and improve the throughput of block chain-based IIOT networks hunting. The Block Hunter framework was evaluated using a variety of machine learning algorithms (NED, IF, CBLOF, K-means, PCA) to detect anomalies. We also examined the impacts of block generation interval, block size, and different miners on the performance of the Block Hunter. Using generative adversarial networks (GAN) to design and implement a block hunter like framework would be an interesting future research work. Furthermore, designing and applying IIOT-related block chain networks with different consensus algorithms would also be worth investigating in the future.

REFERENCES

- [1] J. Wan, J. Li, M. Imran, D. Li, and F. e Amin, "A blockchain-based solution for enhancing security and privacy in smart factory," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 6, pp. 3652–3660, 2019.
- [2] F. Scicchitano, A. Liguori, M. Guarascio, E. Ritacco, and G. Manco, "Blockchain attack discovery via anomaly detection," *Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)*, 2019, 2019.
- [3] Q. Xu, Z. He, Z. Li, M. Xiao, R. S. M. Goh, and Y. Li, "An effective blockchain-based, decentralized application for smart building system management," in *Real-Time*



Data Analytics for Large Scale Sensor Data. Elsevier, 2020, pp. 157–181.

[4] B. Podgorelec, M. Turkanović, and S. Karakatić, “A machine learningbased method for automated blockchain transaction signing including personalized anomaly detection,” *Sensors*, vol. 20, no. 1, p. 147, 2020.

[5] A. Quintal, “Veriblock foundation discloses mess vulnerability in ethereum classic blockchain,” VeriBlock Foundation. [Online]. Available:

<https://www.prnewswire.com/news-releases/veriblock-foundation-discloses-mess-vulnerability-in-ethereum-classic-blockchain-301327998.html>

com/news-releases/veriblock-foundation-discloses-mess-vulnerability-in-ethereum-classic-blockchain-301327998.html

[6] M. Saad, J. Spaulding, L. Njilla, C. Kamhoua, S. Shetty, D. Nyang, and D. Mohaisen, “Exploring the attack surface of blockchain: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1977–2008, 2020.

[7] R. A. Sater and A. B. Hamza, “A federated learning approach to anomaly detection in smart buildings,” *arXiv preprint arXiv:2010.10293*, 2020.

[8] O. Shafiq, “Anomaly detection in blockchain,” Master’s thesis, Tampere University, 2019.

[9] A. Yazdinejadna, R. M. Parizi, A. Dehghantanha, and H. Karimipour, “Federated learning for drone authentication,” *Ad Hoc Networks*, p. 102574, 2021.

[10] D. Preuveneers, V. Rimmer, I. Tsingenopoulos, J. Spooren, W. Joosen, and E. Ilie-Zudor, “Chained anomaly detection models for federated learning: An intrusion

detection case study,” *Applied Sciences*, vol. 8, no. 12, p. 2663, 2018.

[11] L. Tan, H. Xiao, K. Yu, M. Aloqaily, and Y. Jararweh, “A blockchainempowered crowdsourcing system for 5g-enabled smart cities,” *Computer Standards & Interfaces*, vol. 76, p. 103517, 2021.

[12] L. Tseng, X. Yao, S. Otoum, M. Aloqaily, and Y. Jararweh, “Blockchainbased database in an iot environment: challenges, opportunities, and analysis,” *Cluster Computing*, vol. 23, no. 3, pp. 2151–2165, 2020.

[13] M. Signorini, M. Pontecorvi, W. Kanoun, and R. Di Pietro, “Bad: a blockchain anomaly detection solution,” *IEEE Access*, vol. 8, pp. 173 481–173 490, 2020.

PREDICTING AND DEFINING B2B SALES WITH MACHINE LEARNING TECHNIQUE

Ganta James Mydhili (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West
Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

The objectives of this project are two-fold: 1) to use statistical modeling techniques to help a Fortune 500 paper and packaging company codify what drives sales success and 2) to develop a model that can predict sales success with a reasonable degree of accuracy. The desired long-run result is to enable the company to improve both top-line revenue and bottom-line profits by increasing sales close rates, shortening sales cycles, and decreasing the cost of sales. The research team generated several models to predict win propensities for individual sales opportunities, choosing the model with the greatest predictive power and ability to generate insights to use as the backbone for a client tool. To accomplish this, the team leveraged structured and unstructured data from the company's Salesforce.com customer relationship management system. The team experimented with several techniques including binomial logit and various decision tree methods, including boosting with gradient boost and random forest. Individual attributes of customers, opportunities, and internal documentation methods that have the greatest influence on sales success were identified. The best model predicted win propensity with an accuracy of 80%, with precision and recall of 86% and 77%, respectively, which proved to be an improvement over current sales forecast accuracy.

1. INTRODUCTION

The paper and packaging company that provided the data for this research has a long history of sales expertise. This expertise is captured predominantly in the intuition of sales representatives, many of whom have worked in the industry for 20 years or more. Intuition is not easy to record and disseminate across an entire sales force, however, and thus one of the company's most valuable resources is inaccessible to the broader organization. As a result, the company tasked

this team with extracting the most important factors in driving sales success and modeling win propensities using data from their customer relationship management (CRM) system. Most prior work in this space has been performed by private companies, both those that have developed proprietary technologies for internal use and those that sell B2B services related to predictive sales modeling. As a result, research in the field is typically unavailable to the public. Some examples include Implitis a company recently acquired by Salesforce.com that focuses on data automation and predictive modeling and InsightSquared, which sells software that includes a capability to forecast sales outcomes. The academic work that does exist either is related to forecasting aggregate sales instead of scoring opportunity-level propensity, or is based on custom algorithms that fall outside the standard tools used by data scientists in industry. The earliest relevant publication dates only to 2015, in which a joint team from Chinese and US universities employed a two-dimensional Hawkes Process model on seller-lead interactions to score win propensity. Other relevant research has centered around applying highly accurate machine learning algorithms based on sales pipeline data to integrate the insights they produce into an organization's practices, and explaining the output of black-box machine learning models. Considering the lack of visibility into work predicting sales outcome propensity, this research serves to create an initial baseline of understanding on the subject. This project applies and compares several well-known methods for classifying and scoring propensities, a majority of which fall into the category of decision tree modeling.

2. LITERATURE SURVEY

1) On Machine Learning towards Predictive Sales Pipeline Analytics

Sales pipeline win-propensity prediction is fundamental to effective sales management. In contrast to using subjective human rating, we propose a modern machine learning paradigm to estimate the winpropensity of sales leads over time. A profile-specific two-dimensional Hawkes processes model is developed to capture the influence from seller's activities on their leads to the win outcome, coupled with lead's personalized profiles. It is motivated by two observations: i) sellers tend to frequently focus their selling activities and efforts on a few leads during a relatively short time. This is evidenced and reflected by their concentrated interactions with the pipeline, including login, browsing and updating the sales leads which are logged by the system; ii) the pending opportunity is prone to reach its win outcome shortly after such temporally

concentrated interactions. Our model is deployed and in continual use to a large, global, B2B multinational technology enterprise (Fortune 500) with a case study. Due to the generality and flexibility of the model, it also enjoys the potential applicability to other real-world problems.

2) Integration of machine learning insights into organizational learning: A case of B2B sales forecasting

AUTHORS: M. Bohaneca, M.K. Borstnarb, M. Robnik-Sikonja

Business-to-business (b2b) sales forecasting can be described as a decision-making process, which is based on past data (internal and external), formalized rules, subjective judgment, and tacit organizational knowledge. Its consequences are measured in profit and loss. The research focus of this paper is aimed to narrow the gap between planned and realized performance, introducing a novel approach based on machine learning techniques. Preliminary results of machine learning model performance are presented, with focus on distilled visualizations that create powerful, yet human comprehensible and actionable insights, enabling positive climate for reflection and contributing to continuous organizational learning.

3. Explaining machine learning models in sales pre

A complexity of business dynamics often forces decision-makers to make decisions based on subjective mental models, reflecting their experience. However, research has shown that companies perform better when they apply data-driven decision-making. This creates an incentive to introduce intelligent, data-based decision models, which are comprehensive and support the interactive evaluation of decision options necessary for the business environment. Recently, a new general explanation methodology has been proposed, which supports the explanation of state-of-the-art black-box prediction models. Uniform explanations are generated on the level of model/individual instance and support what-if analysis. We present a novel use of this methodology inside an intelligent system in a real-world case of business-to-business (B2B) sales forecasting, a complex task frequently done judgmentally. Users can validate their assumptions with the presented explanations and test their hypotheses using the presented what-if parallel graph representation. The results demonstrate effectiveness and usability of the methodology. A significant advantage of the presented method is the possibility to evaluate seller's actions and to outline general recommendations in sales strategy. This flexibility of the approach and easy-to-follow explanations are suitable for many

different applications. Our well-documented real-world case shows how to solve a decision support problem, namely that the best performing black-box models are inaccessible to human interaction and analysis. This could extend the use of the intelligent systems to areas where they were so far neglected due to their insistence on comprehensible models. A separation of the machine learning model selection from model explanation is another significant benefit for expert and intelligent systems. Explanations unconnected to a particular prediction model positively influence acceptance of new and complex models in the business environment through their easy assessment and switching

3. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

4. CONCLUSION

This research served as a first step in the development of a broader initiative for a Fortune 500 paper and packaging company to operationalize predictive modeling on sales success. As such, the challenges with any large company often include requiring the building of deep local knowledge of the data, in addition to corralling a large organization to assist with accurate data collection. Despite initial inconsistencies in the data, overall accuracy appeared promising and indicated further improvements could be made with better data quality and quantity, more featurerelated investigation and tuning, or perhaps different methods such as neural nets. The analysis also uncovered new insights into what is important regarding sales success. But new insights are often accompanied by new questions: For instance, what kinds of data need to be captured to improve the model's predictive capabilities? How does the culture need to change to improve data capture? This cascade is to be expected, as the broader project lends itself to being a heavily iterative process. There may appear to be a seemingly infinite pool of potential next steps to take in this case. With this in mind, there are a few the team would recommend as the most prudent to consider. Currently, the company could feasibly use the non-meta-variable model to attempt prediction on opportunities in progress for those divisions where accuracy is adequate. To better achieve the objective of predicting open opportunities, it would be prudent to capture and model how opportunity fields change over time, perhaps via periodic snapshots. This way, the company would be able to make predictions at different stages in the opportunity lifecycle. Another important application of these kinds of prediction models is to assist in determining where to invest sales time and resources for business planning optimization. Predictions from accurate models are also worth rolling up into aggregate sales forecasts and adjusting existing "bottom-up" methods. Before these applications would be addressed however, data ops resources would be required to perform a number of critical tasks: continue building and tuning the model for better accuracy, establish a cadence around maintaining the models and incorporating new kinds of information, and connecting with the other business units to understand strategic priorities for operationalization.

5. REFERENCES

1. Implicit (Sales Cloud by Salesforce.com). [Online]. Available: <https://www.salesforce.com/blog/2014/08/infographic-7-powerfulpredictors-closed-won-opportunity-gp.html>

2. Insight Squared. [Online]. Available: <https://www.insightsquared.com/features/sales-forecasting/>
3. J. Yan, C. Zhang, H. Zha, et al., "On Machine Learning towards Predictive Sales Pipeline Analytics." Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, pp. 1945-1951, 2015. [Online]. Available: <https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9444/9488> [Accessed: Mon. 24 Sept. 2018].
4. M. Bohaneca, M.K. Borstnarb, M. Robnik-Sikonja, "Integration of machine learning insights into organizational learning: A case of B2B sales forecasting." 28th Bled eConference, June 7-10, 2015. [Online]. Available: [https://domino.fov.unimb.si/proceedings.nsf/Proceedings/B12ECF2381AB59EEC1257E5B004B39B7/\\$File/2_Bohanec.pdf](https://domino.fov.unimb.si/proceedings.nsf/Proceedings/B12ECF2381AB59EEC1257E5B004B39B7/$File/2_Bohanec.pdf) [Accessed: Tue. 25 Sept. 2018].
5. M. Bohaneca, M.K. Borstnarb, M. Robnik-Sikonja, "Explaining machine learning models in sales predictions." Expert Systems with Applications, no. 71, pp. 416-428, 2017. [Online]. Available: <http://lkm.fri.uni-lj.si/rmarko/papers/Bohanec17-ESwA-preprint.pdf> [Accessed: Tue. 25 Sept. 2018].
6. Gephart, M.A., Marsick, V.J., Mark, E., VanBuren, M.E., Spiro, M.S.: Learning organizations come alive. Training Development 50(12), 36–41 (1996)
7. Nonaka, I., Takeuchi, H.: The Knowledge Creating Organization. Oxford University Press, New York (1995)
8. Senge, P.: The Fifth discipline: The Art & Practice of the Learning Organization. Doubleday Currency, New York (1990)
9. Bohanec, M., Kljajić Borštnar, M., Robnik-Šikonja, M.: Modeling attributes for forecasting B2B opportunities acquisition. In: Proceedings of 34th Conference of Organizational science development, Portorož, Slovenia (2015)
10. Ngai, E.W.T., Xiu, L., Chau, D.C.K.: Application of data mining techniques in CRM: a literature review and classification. Expert Syst. Appl. 36, 2592–2602 (2009)

11. Monat, J. P.: Industrial sales lead conversion modeling. *Market. Intell. Plan.* 29(2), 178–194(2011)
12. Rieg, R.: Do forecast improve over time? *Int. J. Account. Inform. Manage.* 18(3) (2010)
13. Alvarado-Valencia, J.A., Barrero, L.H.: Reliance, trust and heuristics in judgmental forecasting. *Comput. Hum. Behav.* 36, 102–113 (2014)
14. Maaß, D., Spruit, M., Waal, P.D.: Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques, pp. 1–17. *Decision Analytics*, Springer(2014)
15. Witten, I.H., Eibe, F., Hall, M.A.: *Data mining—practical machine learning tools and techniques*, 3rd edn. Elsevier (2011)

MACHINE LEARNING AND END-TO-END DEEP LEARNING FOR THE DETECTION OF CHRONIC HEART FAILURE FROM HEART SOUNDS

Gedala Pradeep (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT: Chronic heart failure (CHF) affects over 26 million of people worldwide, and its incidence is increasing by 2% annually. Despite the significant burden that CHF poses and despite the ubiquity of sensors in our lives, methods for automatically detecting CHF are surprisingly scarce, even in the research community. We present a method for CHF detection based on heart sounds. The method combines classic Machine-Learning (ML) and end-to-end Deep Learning (DL). The classic ML learns from expert features, and the DL learns from a spectro-temporal representation of the signal. The method was evaluated on recordings from 947 subjects from six publicly available datasets and one CHF dataset that was collected for this study. Using the same evaluation method as a recent PhysoNet challenge, the proposed method achieved a score of 89.3, which is 9.1 higher than the challenge’s baseline method. The method’s aggregated accuracy is 92.9% (error of 7.1%); while the experimental results are not directly comparable, this error rate is relatively close to the percentage of recordings labeled as “unknown” by experts (9.7%). Finally, we identified 15 expert features that are useful for building ML models to differentiate between CHF phases (i.e., in the decompensated phase during hospitalization and in the recompensated phase) with an accuracy of 93.2%. The proposed method shows promising results both for the distinction of recordings between healthy subjects and patients and for the detection of different CHF phases. This may lead to the easier identification of new CHF patients and the development of home-based CHF monitors for avoiding hospitalizations.

INDEX TERMS chronic heart failure, deep learning, heart sounds, machine learning, PCG

1.INTRODUCTION

Chronic heart failure (CHF) is a chronic, progressive condition underscored by the heart's inability to supply enough perfusion to target tissues and organs at the physiological filling pressures to meet their metabolic demands [1]. CHF has reached epidemic proportions in the population, as its incidence is increasing by 2% annually. In the developed world, CHF affects 1-2% of the total population and 10% of people older than 65 years. Currently, the diagnosis and treatment of CHF uses approximately 2% of the annual healthcare budget. In absolute terms, the USA spent approximately 35 billion USD to treat CHF in 2018 alone, and the costs are expected to double in the next 10 years [2]. Despite the progress in medical- and device-based treatment approaches in the last decades, the overall prognosis of CHF is still dismal, as 5-year survival rate of this population is only approximately 50%. In the typical clinical course of CHF, we observe alternating episodes of compensated phases, when the patient feels well and does not display symptoms and signs of fluid overload, and decompensated phases, when symptoms and signs of systemic fluid overload (such as breathlessness, orthopnea, peripheral edema, liver congestion, pulmonary edema) can easily be observed. During the latter episodes, patients often require hospital admission to receive treatment with intravenous medications (diuretics, inotropes) to achieve a successful negative fluid balance and return to the compensation state. Early detection of HF worsening would allow a treating physician to adjust the patient's medical management on an outpatient basis in a timely manner and thus avoid the need for a hospital admission. Currently, an experienced physician can detect the worsening of HF by examining the patient and by characteristic changes in the patient's heart failure biomarkers, which are determined from the patient's blood. Unfortunately, clinical worsening of a CHF patient likely means that we are already dealing with a fully developed CHF episode that will most likely require a hospital admission. Additionally, in some patients, characteristic changes in heart sounds can accompany heart failure worsening and can be heard using phonocardiography. An example of a phonocardiogram (PCG) recording of a healthy subject is presented in Fig. 1. In healthy subjects, 2 heart sounds are typically heard (called S1 and S2). S1 is caused by the closure of the mitral valve and ventricular wall in the early systole, S2 is caused by the closure of the aortic and pulmonary valves at the beginning of the diastole. Here, the interval between S1 and S2 is called systole, i.e., the contraction phase of the cardiac cycle, and the interval between

S2 and S1 is called diastole, i.e., the relaxation phase of the cardiac cycle. Additional heart sounds (such as S3 and S4) can be heard in certain cardiac conditions and are never regarded as normal. In the case of CHF (in the course of decompensation), we can often hear a third sound (S3) that typically appears 0.1-0.2 s after the second sound, i.e., S2.

Recently, it has been demonstrated that some physiological parameters, such as the occurrence of additional heart sounds or increased blood pressure in the pulmonary circulation, already start to appear several weeks before the CHF patient develops a clinically evident decompensation episode. This is also an important therapeutic window where outpatient-based treatment interventions can reverse CHF deterioration and return the patient to the compensated state without the need for a hospital admission.

In recent years, many studies have proposed Machine-Learning (ML) approaches for the automatic detection of different heart conditions using PCG signals recorded with a digital stethoscope [1]. Nevertheless, methods that explicitly focus on CHF detection are quite scarce. The typical ML pipeline for the detection of different heart conditions is as follows: segmentation of the signals by detecting the “typical” heart sounds (i.e., S1 and S2), denoising of the signals, extracting individual frequency-domain and time-domain features, and learning a feature-based ML model (e.g., using ML algorithms, such as Random Forest or Support Vector Machine - SVM) that is capable of classifying healthy vs. unhealthy sounds. Most of the features currently used are based on medical and audio/signal analysis knowledge.

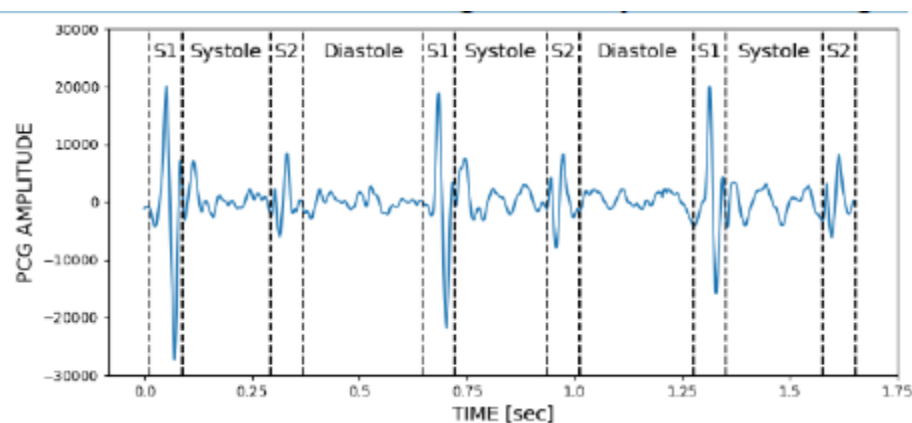


FIGURE 1. Example PCG recording with the PCG regions of interests (S1, S2, Systole and Diastole) marked.

However, a PCG recording that sounds unhealthy to one expert may sound healthy to another one; therefore, doctors never diagnose a CHF patient using only heart sounds, but rather use a holistic view of the patient instead (i.e., extensive medical history, blood pressure, laboratory tests, etc.). This uncertainty is one reason why 9.7% of the recordings in the recent PhysioNet cardiology challenge [3] were actually labeled as “unknown” by experts, while the rest of the recordings were labeled as healthy or unhealthy.

The recent advancements in Deep Learning (DL) suggest that end-to-end learning (i.e., ML models that learn directly from the raw data and no features are needed) can outperform the classic, feature-based ML. For example, DL has achieved breakthrough performance in tasks such as pattern recognition problems [4], image processing [5][6], natural language processing [7][8], speech and audio processing [9][10], and sensor data processing [11][12]. For CHF detection, a successful combination of classic ML and end-to-end DL can outperform each single approach [13]. The classic ML approach learns from a large body of expert-defined features, and the DL approach learns both from a time-domain (the raw PCG signal) representation of the signal and a temporal-domain representation (the spectrogram) of the signal. This approach was successful in our previous study of human activity recognition from smartphone sensor data [14].

In addition to distinguishing the CHF patients and healthy individuals, we focus on detecting the CHF state (compensated vs. decompensated) based on the analysis of heart sound recordings. Our work builds upon the initial studies, where we demonstrated that it is possible to distinguish between healthy individuals and patients in a decompensated CHF episode using a stack of machine-learning classifiers and expert features, showing promising results on a limited dataset [15]. We expand upon this approach using a considerably larger patient dataset, including six additional PhysioNet datasets, and an improved ML method that uses end-to-end DL. Furthermore, we investigate the differences in the heart sounds during the transition between the decompensated and recompensated states of CHF, with the aim of developing personalized monitoring models. Early detection of the worsening of CHF has the potential to reduce hospitalizations due to the worsening of the condition, which both improves the quality of life of patients and decreases the financial and logistic burden on the patient and the health system.

2. RELATED WORK

A typical ML pipeline includes segmentation of the signals, denoising of the signals, extracting individual frequency-domain and time-domain features, and learning ML capable of classifying healthy vs. unhealthy sounds [1][3]. Regarding the segmentation process, Schmidt et al. [16] developed an algorithm (later improved by Springer et al. [17]) that segments the signals into the following four stages: S1, S2, systole, and diastole. The algorithm extracts a variety of features that are then used to train a duration-dependent hidden semi-Markov model to segment the PCG. To ease the segmentation process, some researchers apply denoising techniques to remove environmental sounds and the noises caused by the human body itself. The next phase in the ML pipeline is the feature extraction, as the features are the basis for a successful classification. Most researchers focus mainly on time, frequency, and statistical features. The widely used features are as follows: heart rate, duration of S1, S2, SYS or DIA, total power of the PCG signal, zero crossing-rate, Mel-frequency Cepstral Coefficients, Wavelet Transform, Linear Predictive Coefficients, and Shannon entropy [18][19].

The final phase in the ML pipeline is learning and evaluation of the ML models. The most systematic comparison of the ML models was performed via the PhysioNet challenge [1]. The challenge aimed to encourage the development of algorithms to classify heart sound recordings collected from a variety of clinical or nonclinical (such as in-home visits) environments. More details about the challenge dataset can be found in section 2.2 of the *PhysioNet datasets*. During the challenge, the ML models were ranked using an average of the weighted-sensitivity and the weighted-specificity scores achieved by the models. The weights were used as a normalization factor for the noisy recordings in the data. The best score of 86.0 was achieved by Potes et al. [20] using a method that was based on an ensemble of classifiers combining the outputs of an AdaBoost classifier and a Convolutional Neural Network (CNN). The second-best result, which was 85.9, was achieved by Zabihi et al. [21] using an ensemble of feature-based feedforward neural networks. Similarly, Kay and Agarwal [22] used a fully connected, neural network and achieved a score of 85.2. In fourth place, Bobillo [23] achieved a score of 84.5 using a tensor technique. Homsy et al. [18], who achieved a score of 84.5, introduced an approach using a nested ensemble of algorithms that includes Random Forest, LogitBoost and a Cost-Sensitive Classifier. A probabilistic approach based on logical rules and a probability assessment was

proposed by Plesinger et al. [19], which achieved a score of 84.1. Rubin et al. [24], who achieved a score of 84.0, used CNNs on MFCC heat maps.

Although the challenge datasets present a great opportunity to compare methods for classifying heart sounds, unfortunately, the challenge did not specifically include recordings from CHF patients, and second, it does not provide full access to the challenge test datasets. Thus, we cannot make any CHF-detection comparison. However, we used the publicly available challenge datasets¹ for evaluation and compared the results to the challenge baseline method [3].

With respect to the related work, our approach differs in the following aspects. (i) Most of the approaches from the PhysioNet challenge used the algorithm developed by Schmid et al. [16] for the initial segmentation of the PCG signals, which is based on the detection of the typical heart sounds. However, in noisy environments, the detection of these sounds may be even more challenging than the classification itself. For that reason, our method does not use such a strict segmentation, but rather an overlapping-sliding window technique in combination with a *segment-based* classifier. We present the analysis of the method's performance with respect to the segment (window) size. (ii) The proposed end-to-end DL architecture learns both from the temporal representation of the signal and the spectral representation of the signal, whereas most of the approaches in the related works use end-to-end learning in one of the domains only (either spectral or temporal [14]). (iii) We used the PhysioNet Challenge datasets to evaluate our approach and to provide a comparison with the challenge baseline method; additionally, we used our own dataset, which, in addition to including the typical healthy vs. patient labels, is also labeled for the specific CHF phase, i.e., compensated (when the patient feels well) and decompensated (when the patient does not feel well) for some of the patients. This allowed us to extend the study beyond the typical healthy vs. patient analysis and to explore personalized models for detecting the different CHF phases. To the best of our knowledge, this is the first computer-science study developing CHF detection models.

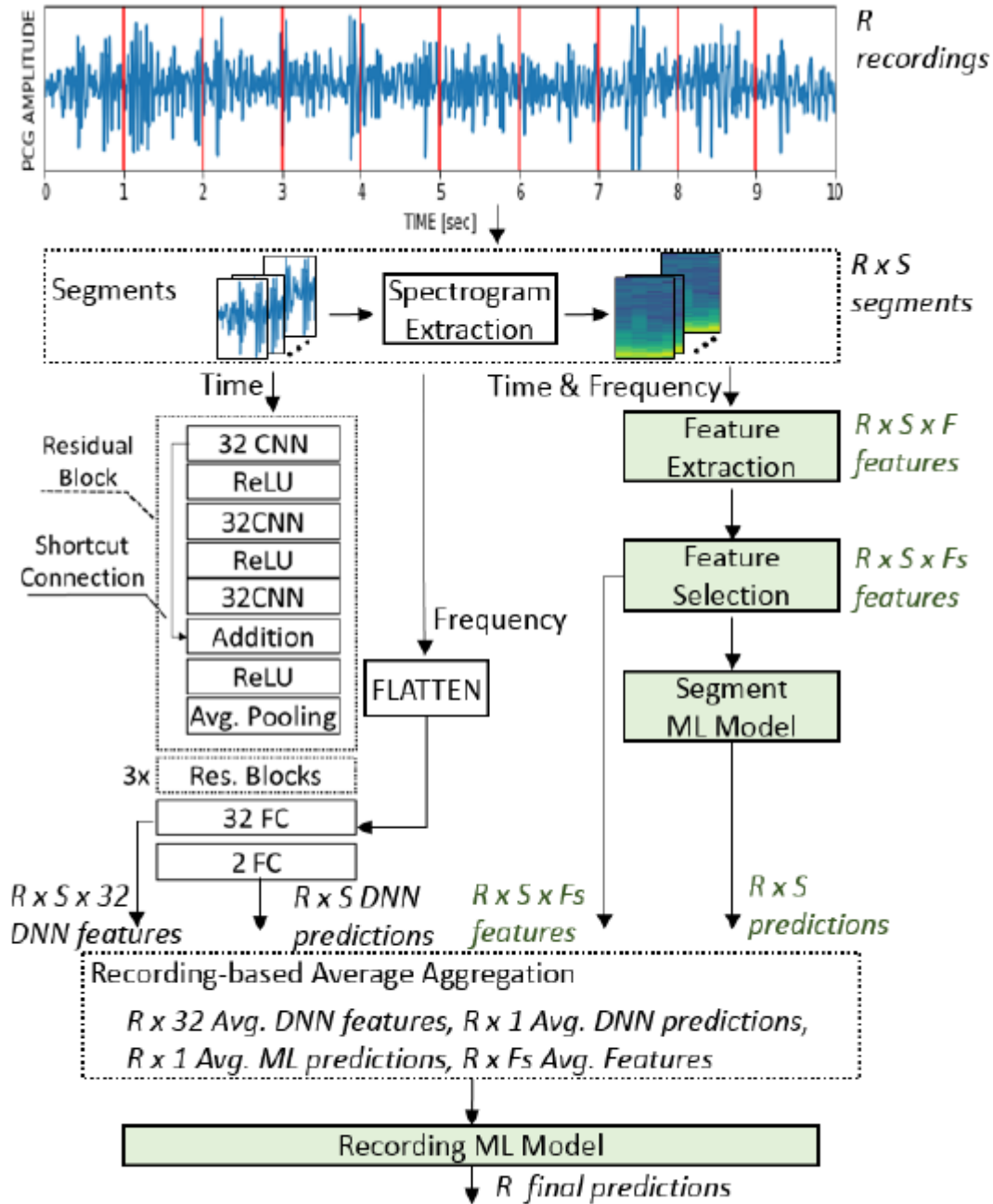


FIGURE 2. Proposed method. End-to-end DL (uncolored squares on the left). Classic ML (colored squares).

3. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

4. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5. CONCLUSIONS

In this paper, we presented a novel method for CHF detection from PCG audio recordings. The method combines classic ML and end-to-end DL. The classical ML learns from a large body of expert-defined features and the DL learns both from the time-domain (i.e., the raw PCG signal) representation of the signal and the spectral representation of the signal. We evaluated the method on our own dataset for CHF detection and additionally on six publicly available PhysioNet datasets used for the recent PhysioNet Cardiology Challenge. The challenge datasets

allowed us to extensively evaluate the performance of the method on similar domains. The evaluation results on all the datasets showed that, compared to the challenge baseline methods, our method achieves the best performance (see the *PhysioNet experiments* section). The facts that most of these datasets are labeled for different types of heart-related conditions and that the PCG audio is recorded from a different body position in most of the datasets (e.g., aortic area, pulmonic area, tricuspid area, and mitral area) strongly indicate that the proposed method is quite robust and that it is useful for detecting different types of heart-sound classification problems and not just for CHF detection, as long as domain-specific labeled data are provided.

Finally, we extended the study beyond the typical healthy vs. patient classification and explored personalized models for detecting different CHF phases, i.e., the recompensated phase (i.e., when the patient feels well) and the decompensated phase (i.e., when the patient needs medical attention). We identified 15 features that have different distributions depending on the phase. By using just two of these features, we were able to build a simple and transparent decision tree classifier (see Fig. 3) that is capable of distinguishing between the recompensated and the decompensated phases with an accuracy of 93.2%, calculated using a LOSO evaluation. While we are aware that there is a risk of overfitting in these final experiments, especially since the dataset contains only 44 samples, we believe that these results are very encouraging and represent a solid base for further development of personalized models. To the best of our knowledge, this is the first study to address such a problem.

6. REFERENCES

- [1] M. Gjoreski et al., “Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers,” in *2017 International Conference on Intelligent Environments (IE)*. IEEE, 2017, pp. 14-19.
- [2] J. Voigt et al., “A reevaluation of the costs of heart failure and its implications for allocation of health resources in the United States,” *Clinical cardiology*, vol. 37, no. 5, pp. 312-321, 2014.
- [3] G. D. Clifford et al., “Classification of normal/abnormal heart sound recordings: the PhysioNet/Computing in Cardiology Challenge 2016,” in *2016 Computing in Cardiology Conference (CinC)*. IEEE, 2016, pp. 609-612.

- [4] X. Jiang, Y. Pang, X. Li, and J. Pan, "Speed up deep neural network based pedestrian detection by sharing features across multi-scale models," *Neurocomputing*, vol. 185, pp. 163-170, 2016.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097-1105.
- [6] C. Szegedy et al., "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.
- [7] T. Young et al., "Recent trends in deep learning based natural language processing," *IEEE Computational intelligence magazine*, vol. 13, no. 3, pp. 55-75, 2018.
- [8] Y. Bengio et al., "A neural probabilistic language model," *Journal of machine learning research*, vol. 3, no. Feb, pp. 1137–1155, 2003.
- [9] S. Amiriparian et al., "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Interspeech*, 2018, pp. 2334-2338.
- [10] S. Amiriparian et al., "Bag-of-Deep-Features: Noise-robust deep feature representations for audio analysis," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1-7.
- [11] M. Gjoreski et al. "Deep affect recognition from R-R intervals," in *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 2017, pp. 754-762.
- [12] H. P. Martinez, Y. Bengio, G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational intelligence magazine*, vol. 8, no. 2, pp. 20-33, 2013.
- [13] M. Gams, "Weak intelligence: Through the principle and paradox of multiple knowledge," *Advances in computation: Theory and practice*, Volume 6, Nova science publishers, inc., NY, ISBN 1-56072-898-1, pp. 245, 2001.

[14] M. Gjoreski et al., “Classical and deep learning methods for recognizing human activities and modes of transportation with smartphone sensors,” in *Information Fusion*. 2019, under revision.

[15] M. Gjoreski et al., “Toward early detection and monitoring of chronic heart failure using heart sounds,” *Intelligent Environments*, pp. 336-343, 2019.



SCHOOL ENTERPRISE CORPORATION ON PYTHON DATA ANALYSIS TEACHING

Gorla Sai (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

To meet the needs of enterprises for data analysis talents, school-enterprise cooperative course Python Data Analysis introduces the basic theories and methods for data analysis by using the Python programming language. The teaching content is organized around practical cases designed according to the actual demand of the enterprise. Exploratory teaching methods are conducted to cultivate the communication, collaboration, critical thinking and creativity ability of students. Students fully participate in the teaching process by discussing, analyzing and programming the cases. The implementation scheme, organization of exploratory teaching and the design of teaching cases are introduced in this paper.

1. INTRODUCTION

Python Data Analysis is a school-enterprise cooperative course supported by the Ministry of Education of China. School-enterprise cooperation program aims to deepen integration between industry and education by building the courses that meet the needs of enterprises. With the development of mobile Internet, Internet of Things and artificial intelligence, more and more disciplines are on the basis of computation, such as computational physics, computational chemistry, and computational biology and so on. Data analysis technique, that is to discover useful information, suggest conclusions, and support decision-making, has become a basic supporting technique of computation. Python Data Analysis course introduces the basic theories and methods for data acquisition, processing, modeling and analysis by using Python programming language. This course is suitable for anyone who wants to perform data analysis. Python is a concise and

efficient programming language preferred as the entry language for programming. Python provides a rich third-party library for data analysis that enables learners to focus on problem solving. After learning this course, students will gain the basic data analysis ability as well as computational thinking for research and practice in all scientific fields they engage in. Reform is conducted on both the teaching content and teaching method. The teaching content is organized around practical cases designed according to the actual demand of enterprises. Exploratory teaching methods are conducted to cultivate the communication, collaboration, critical thinking and creativity ability of students. Students fully participate in the teaching process by discussing, analyzing and solving cases. The rest of this paper is organized as follows. We begin with the introduction of data analysis courses in universities in Section 2. The design of teaching cases is proposed in Section 3, and the



implementation scheme and organization of exploratory teaching are presented.

2. LITERATURE SURVEY

1) A research framework of smart education

The development of new technologies enables learners to learn more effectively, efficiently, flexibly and comfortably. Learners utilize smart devices to access digital resources through wireless networks and to immerse in both personalized and seamless learning. Smart education, a concept that describes learning in the digital age, has gained increased attention. This paper discusses the definition of smart education and presents a conceptual framework. A four-tier framework of smart pedagogies and ten key features of smart learning environments are proposed for fostering smart learners who need master knowledge and skills of 21st century learning. The smart pedagogy framework includes class-based differentiated instruction, group-based collaborative learning, individual-based personalized learning and mass-based generative learning. Furthermore, a technological architecture of smart education, which emphasizes the role of smart computing, is proposed. The tri-tier architecture and key functions are all presented. Finally, challenges of smart education are discussed.

School-Enterprise Cooperation on Python Data Analysis Teaching

To meet the needs of enterprises for data analysis talents, school-enterprise cooperative course Python Data Analysis introduces the basic theories and methods for data analysis by using the Python

programming language. The teaching content is organized around practical cases designed according to the actual demand of the enterprise. Exploratory teaching methods are conducted to cultivate the communication, collaboration, critical thinking and creativity ability of students. Students fully participate in the teaching process by discussing, analyzing and programming the cases. The implementation scheme, organization of exploratory teaching and the design of teaching cases are introduced in this paper. Index Terms— School-enterprise cooperation, Python, data.

3. EXISTING SYSTEM

The potential of artificial intelligence technology is undoubtedly magnificent. As the most widely used technology possessing the highest theoretical research value in artificial intelligence. Early machine learning courses are mainly set up for postgraduate students with majors of Computer and Artificial Intelligence. With the advent of the artificial intelligence and big data age, it is necessary to set up machine learning courses in undergraduates.

DISADVANTAGES OF EXISTING SYSTEM:

- The ability to collect, store, manage and process data has been difficult in existing methods.
- The stage of artificial intelligence is also defined as a discipline about knowledge, namely the technology about how to acquire and express the knowledge and convert it into practical applications



4. PROPOSED SYSTEM

School-enterprise cooperation program aims to deepen integration between industry and education by building the courses that meet the needs of enterprises.

Data analysis technique, that is to discover useful information, suggest conclusions, and support decision-making, has become a basic supporting technique of computation. Python Data Analysis course introduces the basic theories and methods for data acquisition, processing, modeling and analysis by using Python programming language. This course is suitable for anyone who wants to perform data analysis. Python is a concise and efficient programming language preferred as the entry language for programming.

ADVANTAGES OF PROPOSED SYSTEM:

After learning this course, students will gain the basic data analysis ability as well as computational thinking for research and practice in all scientific fields they engage in.

Ability to make faster, more informed business decisions, backed up by facts.

Deeper understanding of customer requirements which, in turn, builds better business relationships.

Improved flexibility and greater capability in order to react to change - both within the business and the market.

5. CONCLUSION

In this paper, we showed our exploration and reform on the school-enterprise course Python Data Analysis. We analyzed the requirement of enterprises and determined the teaching content accordingly. The

teaching content was further organized around the cases designed from simple to deep. To cultivate creativity and scientific exploratory ability, student centered teaching was performed. Students learned by trial, exploratory and verification of cases. The benefit of case based exploratory teaching of data analysis using Python is that students will focus on how to use computational thinking in practical data analysis rather than be overwhelmed by the details of theory and programming.

6. REFERENCES

- [1] Introduction to computer science and programming using python. <https://www.edx.org/course/introduction-computer-science-mitx-6-00-1x5#!>
- [2] Introduction to computational thinking and data science. <https://www.edx.org/course/introduction-computational-thinking-datamitx-6-00-2x-2>.
- [3] Play with data in python. <https://www.icourse163.org/course/nju1001571005>.
- [4] Python data analysis and presentation. <https://www.icourse163.org/course/bit-1001870002>
- [5] Shuaiguo Wang. Smart teaching tools in the context of mobile internet and big data. Modern Educational Technology, 5:26–32, 2017.

PRIVACY-PRESERVING SOCIAL MEDIA DATA PUBLISHING FOR PERSONALIZED RANKING-BASED RECOMMENDATION

Gorla Sirisha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Personalized recommendation is crucial to help users find pertinent information. It often relies on a large collection of user data, in particular users' online activity (e.g., tagging/rating/checking-in) on social media, to mine user preference. However, releasing such user activity data makes users vulnerable to inference attacks, as private data (e.g., gender) can often be inferred from the users' activity data. In this paper, we proposed PrivRank, a customizable and continuous privacy-preserving social media data publishing framework protecting users against inference attacks while enabling personalized ranking-based recommendations. Its key idea is to continuously obfuscate user activity data such that the privacy leakage of user-specified private data is minimized under a given data distortion budget, which bounds the ranking loss incurred from the data obfuscation process in order to preserve the utility of the data for enabling recommendations.

1. INTRODUCTION

DEVELOPING effective recommendation engines is critical in the era of Big Data in order to provide pertinent information to the users. To deliver high-quality and personalized recommendations, online services such as ecommerce applications typically rely on a large collection of user data, particularly user activity data on social media, such as tagging/rating records, comments, check-ins, or other types of user activity data. In practice, many users are willing to release the data (or data streams) about their online activities on social media to a service provider in exchange for getting high-quality personalized recommendations. In this paper, we refer to such user activity data as public data. However, they often consider part of the

data from their social media profile as private, such as gender, income level, political view, or social contacts. In the following, we refer to those data as private data. Although users may refuse to release private data, the inherent correlation between public and private data often causes serious privacy leakage. For example, one's political affiliation can be inferred from her rating of TV shows [1]; one's gender can be inferred from her activities on location-based social networks [2]. These studies show that private data often suffers from inference attacks [3], where an adversary analyzes a user's public data to illegitimately gain knowledge about her private data. It is thus crucial to protect user private data when releasing public data to recommendation engines.

To tackle this problem, privacy-preserving data publishing has been widely studied [4]. Its basic idea is to provide protection on the private data by distorting the public data before its publication, at the expense of a loss of utility of the public data in the latter processing stages. For the use case of recommendation engines, utility refers to the personalization performance based on the distorted public data, i.e., whether the recommendation engines can accurately predict the individual's preference based on the obfuscated data. There is an intrinsic trade-off between privacy and personalization.

On one hand, more distortion of public data leads to better privacy protection, as it makes it harder for adversaries to infer private data. On the other hand, it also incurs a higher loss in utility, as highly distorted public data prevents recommendation engines from accurately predicting users' real preferences.

To apply privacy-preserving data publishing techniques in the case of social media based recommendation, one immediate strategy is to obfuscate user public data on the user side before being sent to social media. However, such an approach is unrealistic as it hinders key benefits for users. In real-world use cases, social media provides users with a social sharing platform, where they can interact with their friends by intentionally sharing their comments/ratings on items, blogs, photos, videos, or even their real-time locations. For example, when a user watched a good movie and wants to share her high rating on it with her friends, she does not want the rating to be obfuscated in any sense. As it is inappropriate to obfuscate user public data before being sent to social media, an alternative solution is to protect user privacy when releasing their public data from social media to any other third-party services. Specifically, many third-party services for

social media require access to user activity data (or data streams) in order to provide them with data, these services may require optional access to users' profiles. While some privacy-conscious users want to keep certain data from their profiles (e.g., gender) as private, other non privacy-conscious users may not care about the same type of private data and choose to release them. Subsequently, an adversary could illegitimately infer the private data of the privacy-conscious users, by learning the correlation between the public and the private data from the non privacy-conscious users. Therefore, it is indispensable to provide privacy protection when releasing user public data from social media.

2. EXISTING SYSTEM

To protect user privacy when publishing user data, the current practice mainly relies on policies or user agreements, e.g., on the use and storage of the published data. However, this approach cannot guarantee that the users' sensitive information is actually protected from a malicious attacker. Therefore, to provide effective privacy protection when releasing user data, privacy-preserving data publishing has been widely studied. Its key idea is to obfuscate user data such that published data remains useful for some application scenarios while the individual's privacy is preserved. According to the attacks considered, existing work can be classified into two categories. The first category is based on heuristic techniques to protect ad-hoc defined user privacy. Specific solutions mainly tackle the privacy threat when attackers are able to link the data owner's identity to a record, or an attribute in the published data. The second category is theory-based and focuses on the uninformative principle, i.e., on the fact that the published data should provide attackers with as little additional information as possible beyond background knowledge.

Disadvantages

- Privacy is less.
- Performance is low.

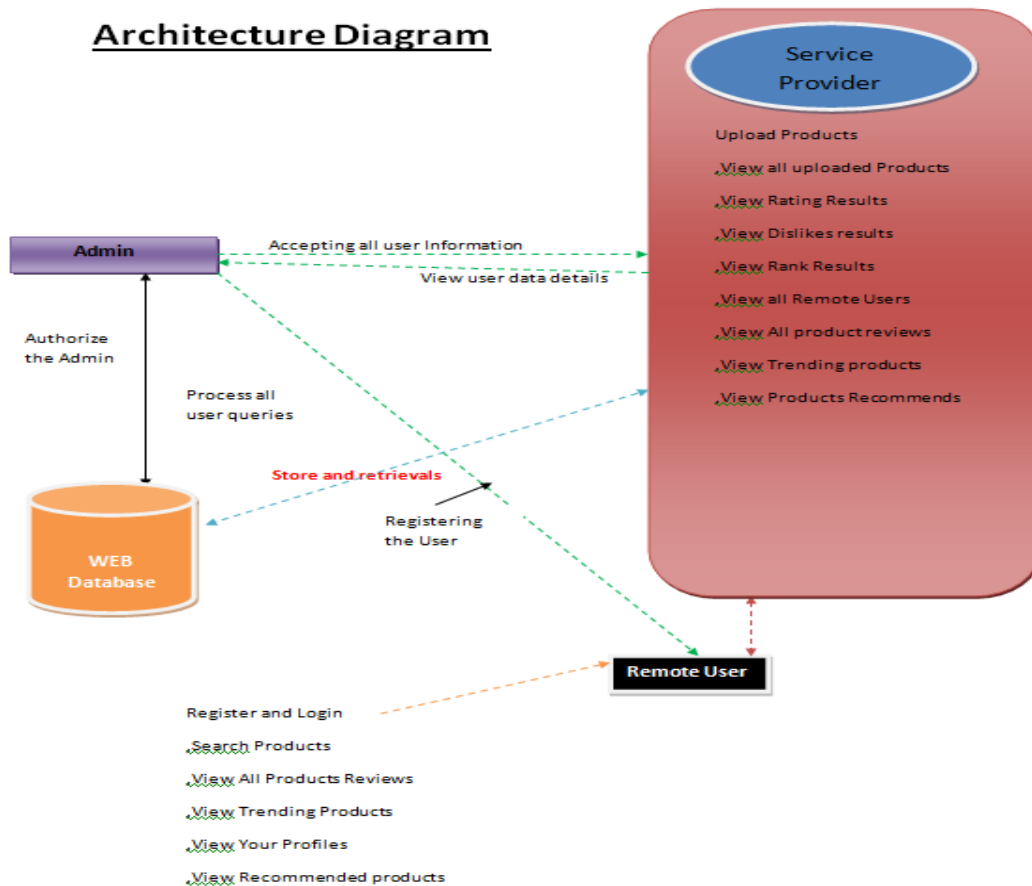
3. PROPOSED SYSTEM

The system proposes PrivRank, a customizable and continuous privacy preserving data publishing framework protect users against inference attacks while enabling personalized

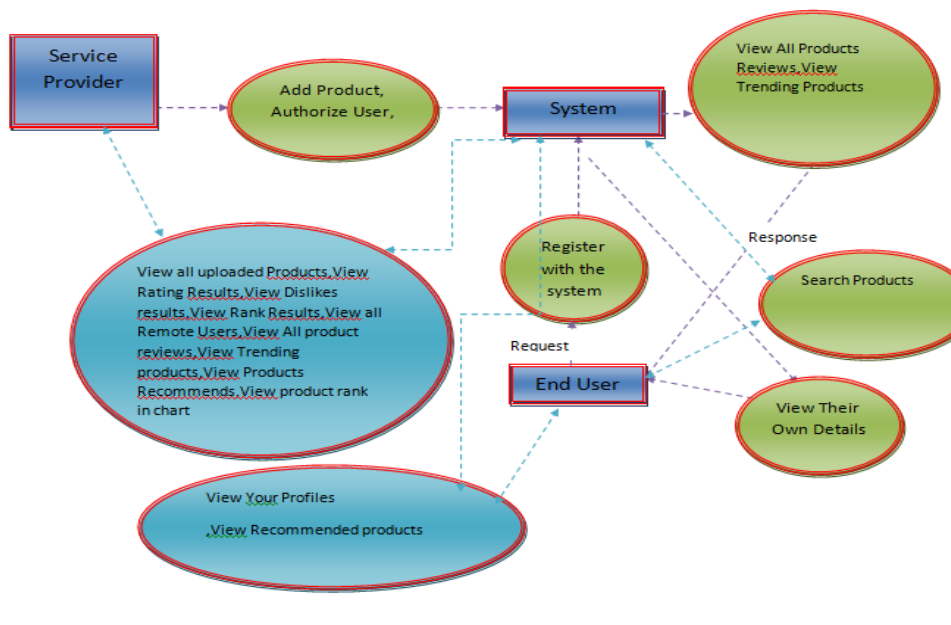
ranking based recommendation. It provides continuous protection of user-specified private data against inference attacks by obfuscating both the historical and streaming user activity data before releasing them, while still preserving the utility of the published data for enabling personalized ranking based recommendation by efficiently limiting the pair wise ranking loss incurred from data obfuscation.

Advantages

- Privacy is more.
- Performance is better.



➤ Data Flow Diagram :



4. CONCLUSION

This paper introduced PrivRank, a customizable and continuous privacy-preserving social media data publishing framework. It continuously protects user-specified data against inference attacks by releasing obfuscated user activity data, while still ensuring the utility of the released data to power personalized ranking-based recommendations. To provide customized protection, the optimal data obfuscation is learned such that the privacy leakage of user-specified private data is minimized; to provide continuous privacy protection, we consider both the historical and online activity data publishing; to ensure the data utility for enabling ranking-based recommendation, we bound the ranking loss incurred from the data obfuscation process using the Kendall-rank distance. We showed through extensive experiments that PrivRank can provide an efficient and effective protection of private data, while still preserving the utility of the published data for different ranking-based recommendation use cases.

In the future, we plan to extend our framework by considering the data types with continuous values rather than discretized values, and explore further data utility beyond personalized recommendation.

5. REFERENCES

- [1] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," in Proc. of GlobalSIP. IEEE, 2013.
- [2] D. Yang, D. Zhang, Q. Bingqing, and P. Cudre-Mauroux, "Privcheck: Privacy-preserving check-in data publishing for personalized location based services," in Proc. of UbiComp'16. ACM, 2016.
- [3] C. Li, H. Shirani-Mehr, and X. Yang, "Protecting individual information against inference attacks in data publishing," in Advances in Databases: Concepts, Systems and Applications. Springer, 2007, pp. 422–433.
- [4] B. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computer Survey, vol. 42, no. 4, p. 14, 2010.
- [5] I. A. Junglas, N. A. Johnson, and C. Spitzmuller, "Personality traits and concern for privacy: an empirical study in the context of location-based services," European Journal of Information Systems, vol. 17, no. 4, pp. 387–402, 2008.
- [6] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in Proc. of RecSys'10. ACM, 2010, pp. 39–46.
- [7] N. Li, R. Jin, and Z.-H. Zhou, "Top rank optimization in linear time," in Advances in neural information processing systems, 2014, pp. 1502–1510.
- [8] M. G. Kendall, "Rank correlation methods." 1948.
- [9] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.
- [10] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," IEEE Transactions on Information Forensics and Security, vol. 8, no. 6, pp. 838–852, 2013.

SPAMMER DETECTION AND FAKE USER IDENTIFICATION ON SOCIAL NETWORKS

Gottumukkala Bhargavi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT: Social networking sites engage millions of users around the world. The users' interactions with these social sites, such as Twitter and Facebook have a tremendous impact and occasionally undesirable repercussions for daily life. The prominent social networking sites have turned into a target platform for the spammers to disperse a huge amount of irrelevant and deleterious information. Twitter, for example, has become one of the most extravagantly used platforms of all times and therefore allows an unreasonable amount of spam. Fake users send undesired tweets to users to promote services or websites that not only affect legitimate users but also disrupt resource consumption. Moreover, the possibility of expanding invalid information to users through fake identities has increased those results in the unrolling of harmful content. Recently, the detection of spammers and identification of fake users on Twitter has become a common area of research in contemporary online social Networks (OSNs). In this paper, we perform a review of techniques used for detecting spammers on Twitter. Moreover, a taxonomy of the Twitter spam detection approaches is presented that classifies the techniques based on their ability to detect: (i) fake content, (ii) spam based on URL, (iii) spam in trending topics, and (iv) fake users. The presented techniques are also compared based on various features, such as user features, content features, graph features, structure features, and time features. We are hopeful that the presented study will be a useful resource for researchers to find the highlights of recent developments in Twitter spam detection on a single platform.

1. INTRODUCTION

It has become quite unpretentious to obtain any kind of information from any source across the world by using the Internet. The increased demand of social sites permits users to collect abundant amount of information and data about users. Huge volumes of data available on these sites also draw the attention of fake users [1]. Twitter has rapidly become an online source for acquiring real-time information about users. Twitter is an Online Social Network (OSN) where

users can share anything and everything, such as news, opinions, and even their moods. Several arguments can be held over different topics, such as politics, current affairs, and important events. When a user tweets something, it is instantly conveyed to his/her followers, allowing them to outspread the received information at a much broader level [2]. With the evolution of OSNs, the need to study and analyze users' behaviors in online social platforms has intensity. Many people who do not have much information regarding the OSNs can easily be tricked by the fraudsters. There is also a demand to combat and place a control on the people who use OSNs only for advertisements and thus spam other people's accounts. Recently, the detection of spam in social networking sites attracted the attention of researchers. Spam detection is a difficult task in maintaining the security of social networks.

It is essential to recognize spams in the OSN sites to save users from various kinds of malicious attacks and to preserve their security and privacy. These hazardous maneuvers adopted by spammers cause massive destruction of the community in the real world. Twitter spammers have various objectives, such as spreading invalid information, fake news, rumors, and spontaneous messages. Spammers achieve their malicious objectives through advertisements and several other means where they support different mailing lists and subsequently dispatch spam messages randomly to broadcast their interests. These activities cause disturbance to the original users who are known as non-spammers. In addition, it also decreases the reputation of the OSN platforms. Therefore, it is essential to design a scheme to spot spammers so that corrective efforts can be taken to counter their malicious activities [3].

Several research works have been carried out in the domain of Twitter spam detection. To encompass the existing state-of-the-art, a few surveys have also been carried out on fake user identification from Twitter. Tingmin *et al.* [4] provide a survey of new methods and techniques to identify Twitter spam detection. The above survey presents a comparative study of the current approaches. On the other hand, the authors in [5] conducted a survey on different behaviors exhibited by spammers on Twitter social network. The study also provides a literature review that recognizes the existence of spammers on Twitter social network. Despite all the existing studies, there is still a gap in the existing literature. Therefore, to bridge the gap, we review state-of-the-art in the spammer detection and fake user identification on Twitter. Moreover, this survey presents a taxonomy of the Twitter spam detection approaches and attempts to offer a detailed description of recent developments in the domain.

The aim of this paper is to identify different approaches of spam detection on Twitter and to present a taxonomy by classifying these approaches into several categories. For classification, we have identified four means of reporting spammers that can be helpful in identifying fake identities of users. Spammers can be identified based on: (i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. Table 1 provides a comparison of existing techniques and helps users to recognize the significance and effectiveness of the proposed methodologies in addition to providing a comparison of their goals and results. Table 2 compares different features that are used for identifying spam on Twitter. We anticipate that this survey will help readers find diverse information on spammer detection techniques at a single point.

2. EXISTING SYSTEM

Shen *et al.* [29] investigated issues of detecting spammers on Twitter. The proposed method combines characteristics withdrawal from text content and information of social networks. The authors used matrix factorization to determine the underline feature matrix or the tweets and then came up with a social regularization with interaction coefficient to teach the factorization of the underline matrix. Subsequently, the authors combined knowledge with social regularization and factorization matrix processes, and performed experiments on the real-world Twitter dataset, i.e., UDI Twitter dataset.

Washha *et al.* [31] described the Hidden Markov Model for filtering the spam related to recent time. The method supports the accessible and obtainable information in the tweet object to recognize spam tweets and the tweets that are handled previously related to the same topic.

Jeong *et al.* [17] analyzed the follow spam on Twitter as an alternative of dispersion of provoking public messages, spammers follow authorized users, and followed by authorized users. Categorization techniques were proposed that are used for the detection of follow spammers. The focus of the social relation is cascaded and formulated into two mechanism, i.e., social status filtering and trade significance profile filtering, where each of which uses two-hop sub networks that are centered at each other. Assemble techniques and cascading filtering are also proposed for combining the properties of both trade significance profile and social status. To check whether a user is fake or not, a two-hop social network for each user is focused to gather social information from social networks.

Meda *et al.* [21] presented a technique that utilizes a sampling of non-uniform features inside a machine learning system by the adaptation of random forest algorithm to recognize spammer insiders. The proposed framework focuses on the random forest and non-uniform feature sampling techniques. The random forest is a learning algorithm for the categorization and regression that works by assembling several decision trees at preparation time and selecting the one with the majority votes by individual trees. The scheme integrates bootstrap aggregating technique with the un-planned selection of features.

Disadvantages

- There is no filtering system based on a preprocessing schedule and on Naïve Bayes algorithm to discard the tweets containing inaccurate information,
- Less security due No URL Based Spam Detection.

3. PROPOSED SYSTEM

In the proposed system, the system elaborates a classification of spammer detection techniques. The system shows the proposed taxonomy for identification of spammers on Twitter. The proposed taxonomy is categorized into four main classes, namely, (i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. Each category of identification methods relies on a specific model, technique, and detection algorithm.

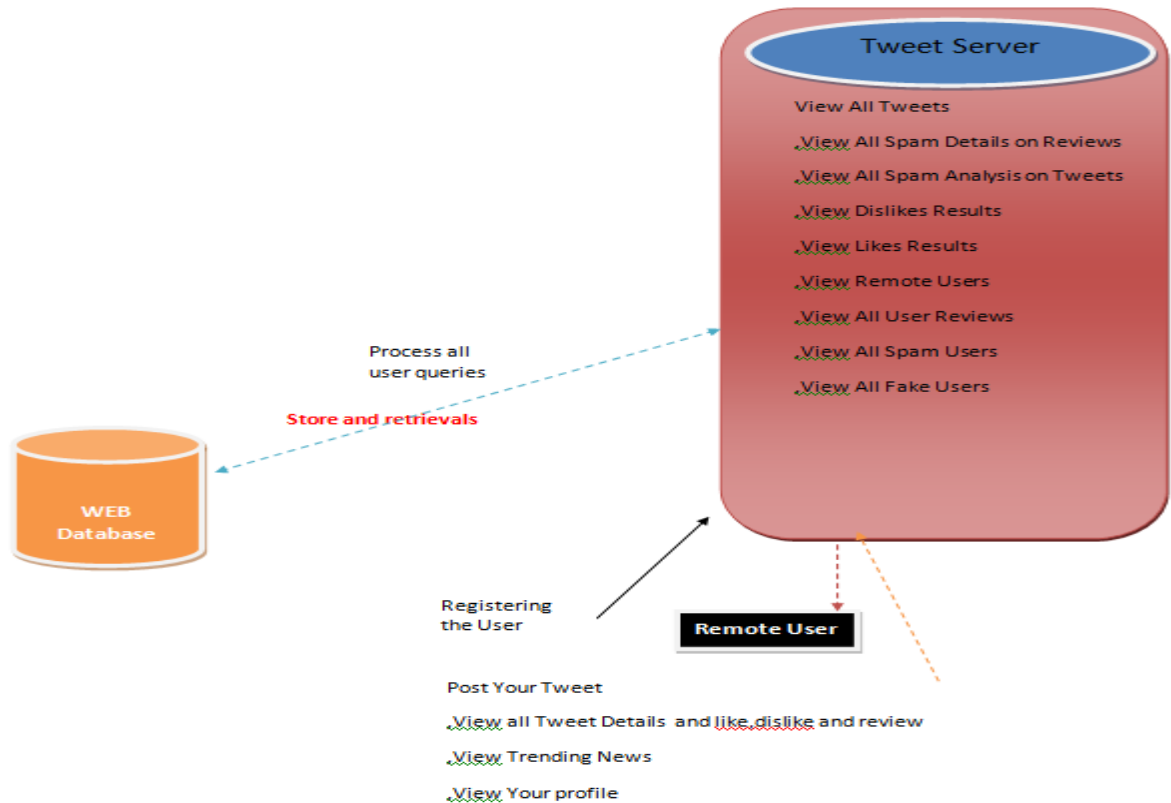
The first category (fake content) includes various techniques, such as regression prediction model, malware alerting system, and Lfun scheme approach. In the second category (URL based spam detection), the spammer is identified in URL through different machine learning algorithms. The third category (spam in trending topics) is identified through Naïve Bayes classifier and language model divergence. The last category (fake user identification) is based on detecting fake users through hybrid techniques.

Advantages

- The average numbers of verified accounts that were either spam or non-spam and (ii) the number of followers of the user accounts.
- The fake content propagation was identified through the metrics that include: (i) social reputation, (ii) global engagement, (iii) topic engagement, (iv) likability, and (v) credibility. After that, the authors utilized regression prediction model to ensure

the overall impact of people who spread the fake content at that time and also to predict the fake content growth in future.

Architecture Diagram



4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY

◆ SOCIAL FEASIBILITY

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSION

In this paper, we performed a review of techniques used for detecting spammers on Twitter. In addition, we also presented a taxonomy of Twitter spam detection approaches and categorized them as fake content detection, URL based spam detection, spam detection in trending topics, and fake user detection techniques. We also compared the presented techniques based on several features, such as user features, content features, graph features, structure features, and time features. Moreover, the techniques were also compared in terms of their specified goals and datasets used. It is anticipated that the presented review will help researchers find the information on state-of-the-art Twitter spam detection techniques in a consolidated form.

Despite the development of efficient and effective approaches for the spam detection and fake user identification on Twitter [34], there are still certain open areas that require considerable attention by the researchers. The issues are briefly highlighted as under: False news identification on social media networks is an issue that needs to be explored because of the serious repercussions of such news at individual as well as collective level [25]. Another associated topic that is worth investigating is the identification of rumor sources on social media. Although a few studies based on statistical methods have already been conducted to detect the sources of rumors, more sophisticated approaches, e.g., social network based approaches, can be applied because of their proven effectiveness.

7. REFERENCES

- [1] B. Erçahin, Ö. Akta³, D. Kiliñç, and C. Akyol, "Twitter fake account detection," in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, pp. 388392.

- [2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on Twitter," in Proc. Collaboration, Electron. Messaging, Anti- Abuse Spam Conf. (CEAS), vol. 6, Jul. 2010, p. 12.
- [3] S. Gharge, and M. Chavan, "An integrated approach for malicious tweets detection using NLP," in Proc. Int. Conf. Inventive Commun. Comput. Technol. (ICICCT), Mar. 2017, pp. 435438.
- [4] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Comput. Secur.*, vol. 76, pp. 265284, Jul. 2018.
- [5] S. J. Soman, "A survey on behaviors exhibited by spammers in popular social media networks," in Proc. Int. Conf. Circuit, Power Comput. Tech- nol. (ICCPCT), Mar. 2016, pp. 16.
- [6] A. Gupta, H. Lamba, and P. Kumaraguru, "1.00 per RT #BostonMarathon # prayforboston: Analyzing fake content on Twitter," in Proc. eCrime Researchers Summit (eCRS), 2013, pp. 112.
- [7] F. Concone, A. De Paola, G. Lo Re, and M. Morana, "Twitter analysis for real-time malware discovery," in Proc. AEIT Int. Annu. Conf., Sep. 2017, pp. 16.
- [8] N. Eshraqi, M. Jalali, and M. H. Moattar, "Detecting spam tweets in Twitter using a data stream clustering algorithm," in Proc. Int. Congr. Technol., Commun. Knowl. (ICTCK), Nov. 2015, pp. 347351.
- [9] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted Twitter spam," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 4, pp. 914925, Apr. 2017.
- [10] C. Buntain and J. Golbeck, "Automatically identifying fake news in popular Twitter threads," in Proc. IEEE Int. Conf. Smart Cloud (SmartCloud), Nov. 2017, pp. 208215.
- [11] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian, "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 6576, Sep. 2015.
- [12] G. Stafford and L. L. Yu, "An evaluation of the effect of spam on Twitter trending topics," in Proc. Int. Conf. Social Comput., Sep. 2013, pp. 373378.
- [13] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for Twitter," in Proc. 14th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST), Jan. 2017, pp. 466471.



PREDICTING URBAN WATER QUALITY WITH UBIQUITOUS DATA - A DATA DRIVEN APPROACH

Grandhi Harsha Nagu (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K.R.Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT—Urban water quality is of great importance to our daily lives. Prediction of urban water quality help control water pollution and protect human health. However, predicting the urban water quality is a challenging task since the water quality varies in urban spaces non-linearly and depends on multiple factors, such as meteorology, water usage patterns, and land uses. In this work, we forecast the water quality of a station over the next few hours from a data-driven perspective, using the water quality data and water hydraulic data reported by existing monitor stations and a variety of data sources we observed in the city, such as meteorology, pipe networks, structure of road networks, and point of interests (POIs). First, we identify the influential factors that affect the urban water quality via extensive experiments. Second, we present a multi-task multi-view learning method to fuse those multiple datasets from different domains into an unified learning model. We evaluate our method with real-world datasets, and the extensive experiments verify the advantages of our method over other baselines and demonstrate the effectiveness of our approach.

Index Terms—Urban computing; data mining; urban water quality prediction; multi-view learning; multi-task learning; big data.

1. INTRODUCTION

Urban water is a vital resource that affects various aspects of human, health and urban lives. People living in major cities are increasingly concerned about the urban water quality, calling for technology that can monitor and predict the water quality in real time throughout the city. Urban water quality, which serves as “a powerful environmental determinant” and “a foundation for the prevention and control of waterborne diseases” [1], refers to the physical, chemical and biological characteristics of a water body, and several chemical indexes (such as residual chlorine, turbidity and pH) can be used as effective measurements for the water quality in current urban water distribution systems [2]. With the increasing demand for water quality information, several water quality monitoring stations have been deployed throughout the city’s water distribution system to provide the real-time water quality reports in a city. Figure 1 illustrates the water quality monitor stations that have been deployed in Shenzhen, China. Besides water quality monitoring, predicting the urban water quality plays an essential role in many urban aquatic projects, such as informing waterworks’ decision making (e.g., pre-adjustment of chlorine from the waterworks), affecting governments’ policy making (e.g., issuing pollution alerts or performing a pollution control), and providing maintenance suggestions (e.g., suggestions for replacements of certain pipelines). Predicting urban water quality, however, is



very challenging due to the following reasons. First, urban water quality varies by locations non-linearly and depends on multiple factors, such as meteorology, water usage patterns, land use, and urban structures. As depicted in Figure 1, the water quality indexes (RC) reported by the three stations demonstrate different patterns. Existing hydraulic model-based approaches try to model water quality from physical and chemical perspective, but such hydraulic model can hardly capture all of those complex factors. Moreover, the parameters in model are hard to get, which makes it difficult to extend to other water distribution systems. Second, as all the stations are connected through the pipeline system, the water quality among different stations are mutually correlated by several complex factors, such as attributes in pipe networks and distribution of POIs. Traditional hydraulic model-based approaches build hydraulic model for each station and ignore their spatial correlations, and thus their performance is far from satisfactory. Hence, besides identifying the influential factors, how to efficiently characterize and incorporate such relatedness poses another challenge. Fortunately, in the era of big data [3] [4] [5], unprecedented data in urban areas (e.g., meteorology, POIs, and road networks) can provide complementary information to help predict the urban water quality. For example, temperature can be an indicator of water quality, with higher temperature indicating better water quality. The possible reason is that the water consumption tends to grow when temperature is high since most people may choose to take a shower, and the increased water consumption is one major cause that prevents the water quality's deterioration in the distribution systems. To benefit from the unprecedented data in urban areas, in this paper, we predict the water quality of a station through a data-driven perspective using a variety of data sets, including water quality data, hydraulic data, meteorology data, pipe networks data, road networks data, and POIs. First, we perform extensive experiments and data analytics between the water quality and multiple potential factors, and identify the most influential ones that have an effect on the urban water quality. Second, we present a novel spatio-temporal multi-task multi-view learning (stMTMV) framework to fuse the heterogeneous data from multiple domains and jointly capture each station's local information as well as their global information into a unified learning model [6].

2. OVERVIEW

Preliminary Definition 1 (Water Quality): Urban water quality refers to the physical, chemical and biological characteristics of a water body [2]. In current urban water distribution systems, several monitoring stations are deployed in the distribution systems to report three important quality indexes, i.e., residual chlorine, turbidity and pH, in real time. The three indexes can be used as effective measurements for the water quality in current urban water distribution systems [1]. In this paper, we consider Residual Chlorine (RC) as the water quality index since it “inactivates the bacteria and some viruses that cause diarrheal disease”, and “can protect the water from recontamination during storage” [7], which is widely employed as the major water quality index in the environmental science [1] [2]. **Definition 2 (Water Hydraulics):** Water hydraulics describe the hydraulic characteristics of the water in urban water distribution systems.



It consists of two major indexes: flow and pressure, and can be obtained through the deployed flow and pressure sensors. Definition 3 (Pipe Network): Urban water is distributed through the pipe network, which is an underground network of interconnecting mains or pipes. A pipe network PN comprises of a set of pipeline segments p that connect between each other in the format of a graph. Each pipeline segment p an undirected edge having two terminal nodes, and has several attributes, including length $p.len$, diameter $p.d$, and age $p.age$. Definition 4 (Road Network): A road network RN comprises of a set of road segments r , connecting each other in the format of a graph. Each road segment r is a directed edge having two terminal nodes, a series of intermediate points between the two terminals, a length of $r.len$.

3. EXISTING SYSTEM

Several studies in the environmental science have been tried to analyze the water quality problems via data-driven based approaches, and those studies covers a range of topics, from the physical process analysis in the river basin, to the analysis of concurrent input and output time series [64] [65]. The approaches adopted in these studies include instance-based learning models (e.g., kNN) as well as neural network models (e.g., ANN). In general, those data-driven approaches in the environmental science can fall into the following three major categories: Instance-based Learning models (IBL), Artificial Neural Network models (ANN) and Support Vector Machine models (SVM).

Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to test model [66]. As a typical example of IBL, k-Nearest Neighbors

(k-NN) is widely used due to its simplicity and incredibly good performance in practice.

For example, the work introduced by Karlsson et al. [67] addressed the classical rainfall-runoff forecasting problem by k-NN algorithm, and demonstrated promising results. Toth et al. [68] used k-NN to predict the rainfall depths from the history data, and showed the persistent outperformance of k-NN over other time series prediction methods.

As another example, Ostfeld et al. [69] developed a hybrid genetic k-Nearest Neighbor algorithm to calibrate the two-dimensional surface quantity and water quality model. Artificial Neural Network (ANN) is a network inspired by biological neural networks (in particular the human brain), which consists of multiple layers of nodes (neurons) in a directed graph with each layer fully connected to the next one [65]. Neural networks have been widely employed to solve a wide variety of tasks, and can achieve good results. For instance, Moradkhani et al. [70] proposed an hourly streamflow forecasting method based on a radial-basis function (RBF) network and demonstrated its advantages over other numerical prediction methods. Also, the work introduced by Kalin [44] predicted the water quality indexes in watersheds through ANN.

Support Vector Machines (SVMs) are typical supervised learning models that analyze data used for classification and regression [71].



In aquatic studies, it was also extended to solving prediction problems [64]. For instance, Liong et al. [72] addressed the issue of flood forecasting using Support Vector Regression (SVR) which is an extension of SVM. Another work by Xiang et al. [73] utilized a LS-SVM model to deal with the water quality prediction problem in Liuxi River in Guangzhou.

However, none of these approaches is applied into urban scenarios, which is quite different from our applications. Moreover, those existing approaches process the data from a single source, and can hardly integrate the data from different sources. Thus, their applications in the urban scenarios are restricted.

Disadvantages

The system is implemented only Multi-task Multi-view Learning Approaches.

Instance-based learning models (IBL) is a family of learning algorithms that model a decision problem with instances or examples of training data that are deemed important to the model.

4. PROPOSED SYSTEM

_ Data-driven Perspective: We present a novel data-driven approach to co-predict the future water quality among different stations with data from multiple domains. Additionally, the approach is not restricted to urban water quality prediction, but also can be applied to other multi-locations based coprediction problem in many other urban applications.

_ Influential Factor Identification: We identify spatially-related (such as POIs, pipe networks, and road networks) and temporally-related features (e.g., time of day, meteorology and water hydraulics), contributing to not only our application but also the general problem of water quality prediction.

_ Unified Learning Model: We present a novel spatio-temporal multi-view multi-task learning framework (stMTMV) to integrate multiple sources of spatio-temporal urban data, which provides a general framework of combining heterogeneous spatio-temporal properties for prediction, and can also be applied to other spatio-temporal based applications.

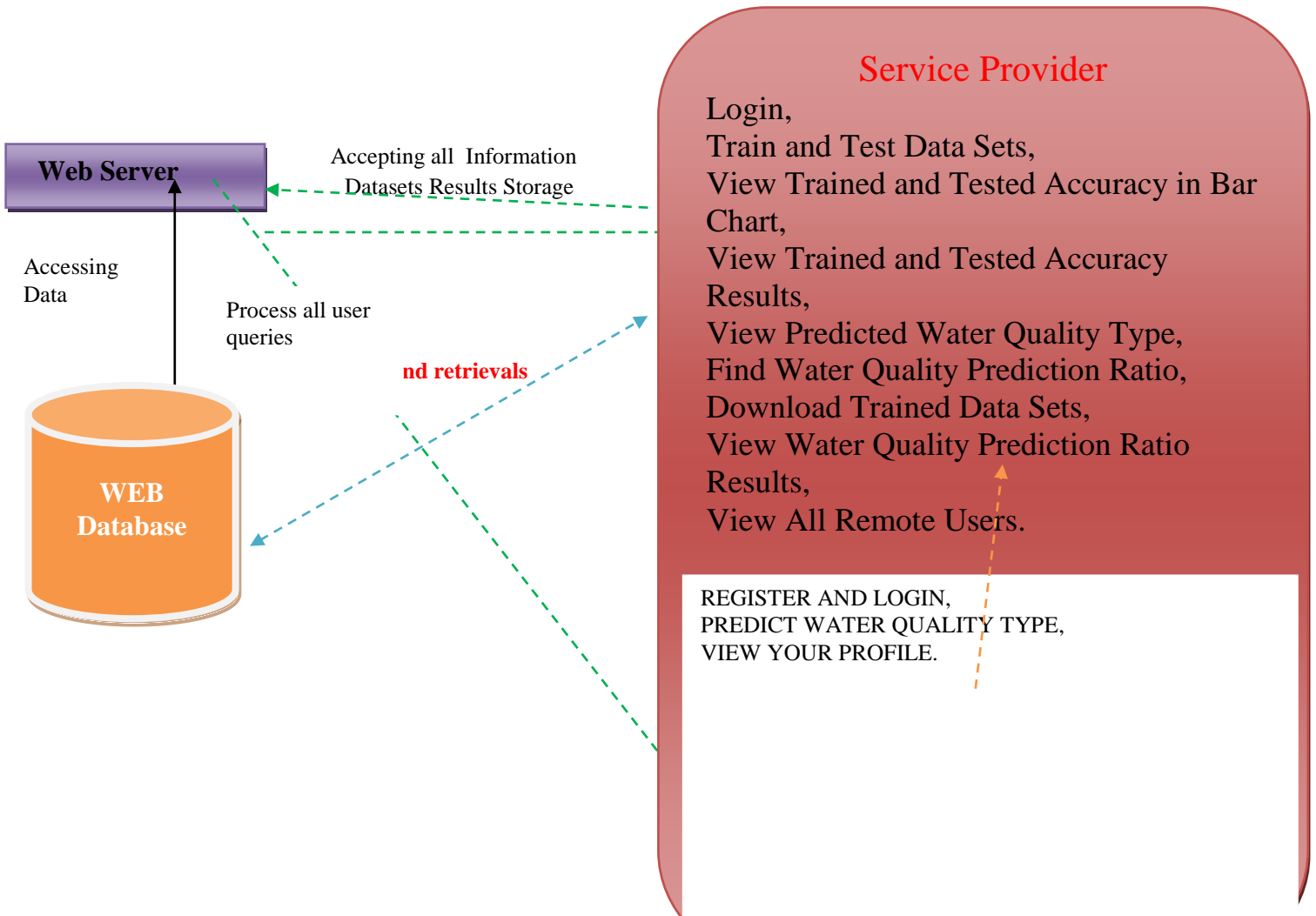
_ Real evaluation: We evaluate our method by extensive experiments that use real-world datasets in Shenzhen, China. The results demonstrate the advantages of our method beyond other baselines, such as ARMA, Kalman filter, and ANN, and reveal interesting discoveries that can bring social good to urban life.

Advantages

1) Water quality data: We collect water quality data every five minutes from 15 water quality monitoring stations in Shenzhen City. It comprises residual chlorine (RC), turbidity (TU) and pH. In this paper, we only use RC as the index for water quality, since RC is the most important and effective measurement for water quality in current urban water distribution system.

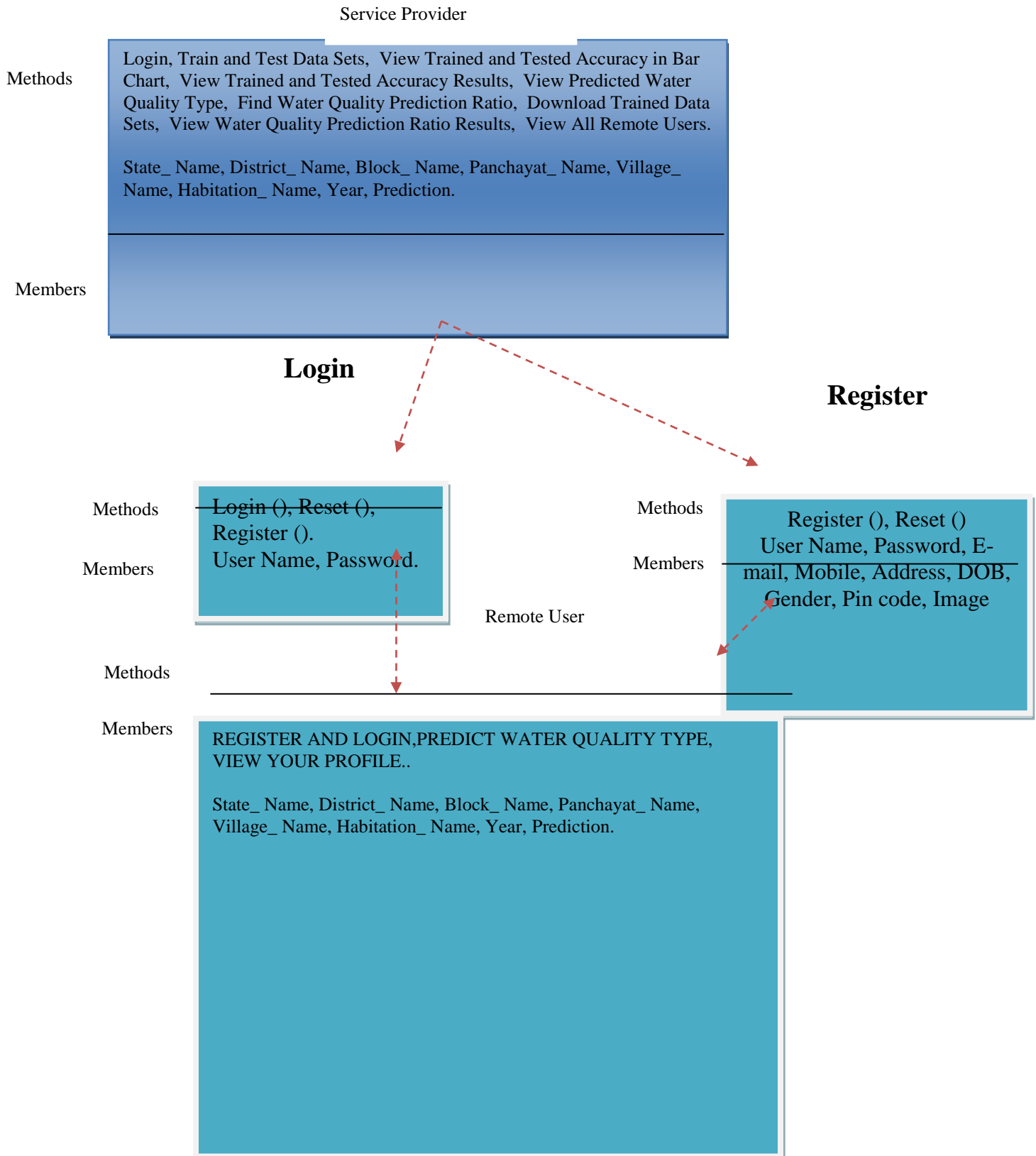
2) Hydraulic data: Hydraulic data consists of flow and pressure, which are collected every five minutes from 13 flow sites and 14 pressure sites, respectively.

Architecture Diagram



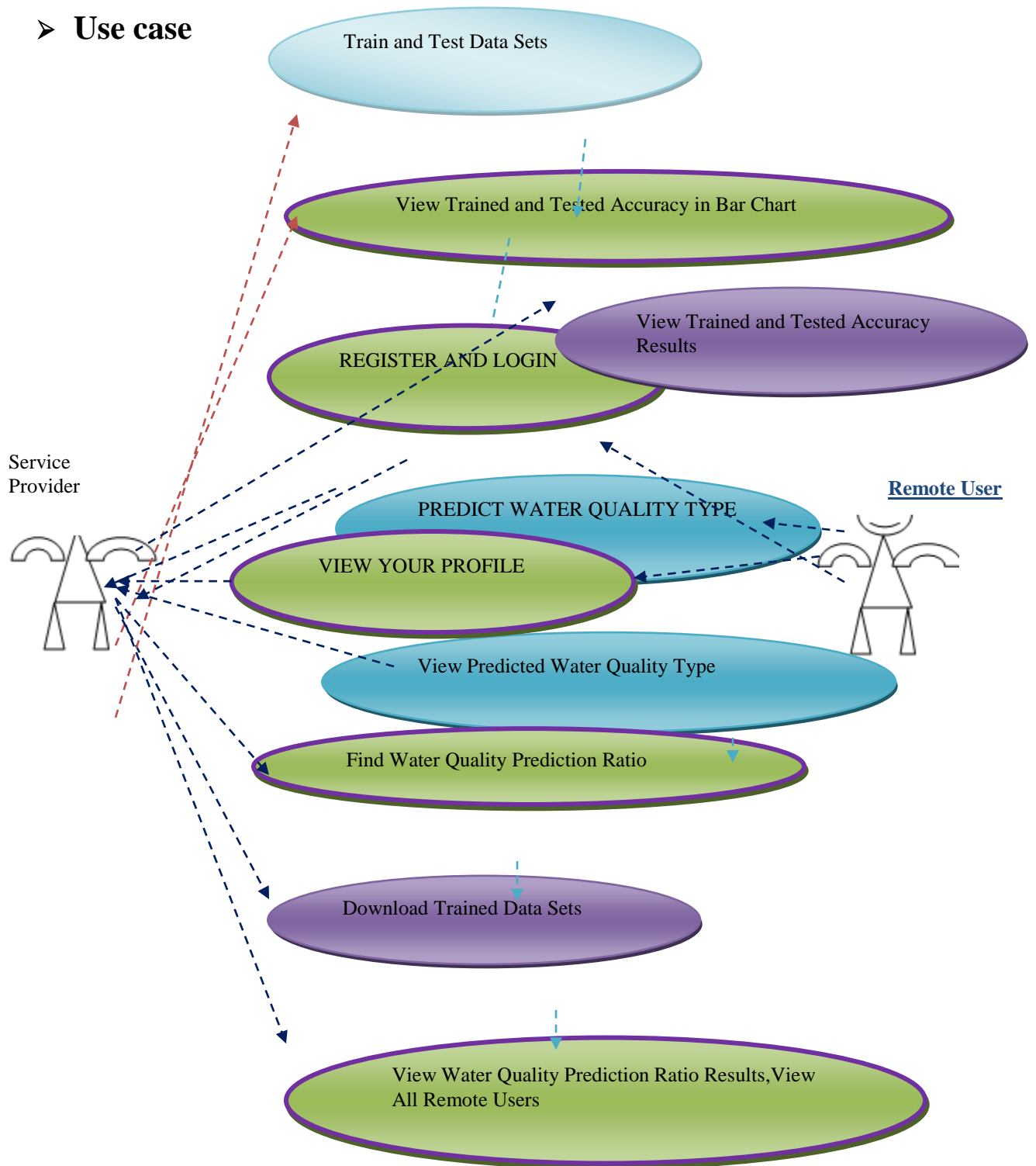


➤ Class Diagram :





➤ Use case





5. CONCLUSION

This paper presents a novel data-driven approach to forecast the water quality of a station by fusing multiple sources of urban data. We evaluate our approach based on Shenzhen's water quality and various urban data. The experimental results demonstrate the effectiveness and efficiency of our approach. Specifically, our approach outperforms the traditional RC decay model [2] and other classical time series predictive models (ARMA, Kalman) in terms of RMSE metric. Meanwhile, as our approach consists of two components, each of the components demonstrates its effectiveness through extensive experiments and analysis. In particular, the first component is the influential factors identification, which explores the factors that affect the urban water quality via extensive experiments and analysis in Section 3 and 4. The second one is a spatiotemporal multi-view multi-task learning (STMTMV) framework that consists of multi-view learning and multi-task learning. The experiments have shown that STMTMV has a predictive accuracy of around 85% for forecasting next 1-4 hours, which outperforms the single-task methods (LR) by approximately 11% and the single-view methods (t-view and s-view) by approximately 11% and 12%, respectively. The code has been released at: <https://www.microsoft.com/en-us/research/publication/urbanwater-quality-prediction-based-multi-task-multi-view-learning-2/> In future, we plan to deal with the water quality inference problems in the urban water distribution systems through a limited number of water quality monitor stations.

6. REFERENCES

- [1] W. H. Organization, Guidelines for drinking-water quality, 2004, vol. 3.
- [2] L. A. Rossman, R. M. Clark, and W. M. Grayman, "Modeling chlorine residuals in drinking-water distribution systems," *Journal of environmental engineering*, vol. 120, no. 4, pp. 803–820, 1994.
- [3] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Transactions on Big Data*, vol. 1, no. 1, pp. 16–34, 2015.
- [4] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 38:1–38:55, 2014.
- [5] Y. Zheng, H. Zhang, and Y. Yu, "Detecting collective anomalies from multiple spatio-temporal datasets across different domains," 2015.
- [6] Y. Liu, Y. Zheng, Y. Liang, S. Liu, and D. S. Rosenblum, "Urban water quality prediction based on multi-task multi-view learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.
- [7] H. Cohen, "Free chlorine testing," <http://www.cdc.gov/safewater/chlorineresidual-testing.html>, 2014, accessed on 5 August 2016.
- [8] B. D. Barkdoll and H. Didigam, "Effect of user demand on water quality and hydraulics of distribution systems," in *Proceedings of the World Water and Environmental Resources Congress*, 2003.



- [9] P. Castro and M. Neves, "Chlorine decay in water distribution systems case study–lousada network," *Electronic Journal of Environmental, Agricultural and Food Chemistry*, vol. 2, no. 2, pp. 261–266, 2003.
- [10] L. W. Mays, *Water distribution system handbook*, 1999.
- [11] L. A. Rossman and P. F. Boulos, "Numerical methods for modeling water quality in distribution systems: A comparison," *Journal of Water Resources planning and management*, vol. 122, no. 2, pp. 137–146, 1996.
- [12] W. M. Grayman, R. M. Clark, and R. M. Males, "Modeling distributionsystem water quality: dynamic approach," *Journal of Water Resources Planning and Management*, vol. 114, no. 3, pp. 295–312, 1988.
- [13] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003, pp. 2–11.
- [14] G. Luo, K. Yi, S.-W. Cheng, Z. Li, W. Fan, C. He, and Y. Mu, "Piecewise linear approximation of streaming time series data with max-error guarantees," in *Proceedings of the IEEE International Conference on Data Engineering*, 2015, pp. 173–184.
- [15] E. O. Brigham and E. O. Brigham, *The fast Fourier transform*. PrenticeHall Englewood Cliffs, NJ, 1974, vol. 7.

MACHINE LEARNING FOR FAST AND RELIABLE SOURCE- LOCATION ESTIMATION IN EARTHQUAKE EARLY WARNING

Grandhi Nagaswarnadurga Divyasri (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

We develop a random forest (RF) model for rapid earthquake location with an aim to assist earthquake early warning (EEW) systems in fast decision making. This system exploits P-wave arrival times at the first five stations recording an earthquake and computes their respective arrival time differences relative to a reference station (i.e., the first recording station). These differential P-wave arrival times and station locations are classified in the RF model to estimate the epicentral location. We train and test the proposed algorithm with an earthquake catalog from Japan. The RF model predicts the earthquake locations with a high accuracy, achieving a Mean Absolute Error (MAE) of 2.88 km. As importantly, the proposed RF model can learn from a limited amount of data (i.e., 10% of the dataset) and much fewer (i.e., three) recording stations and still achieve satisfactory results (MAE<5 km). The algorithm is accurate, generalizable, and rapidly responding, thereby offering a powerful new tool for fast and reliable source-location prediction in EEW.

1. INTRODUCTION

EARTHQUAKE hypocenter localization is essential in the field of seismology and plays a critical role in a variety of seismological applications such as tomography, source characterization, and hazard assessment. This underscores the importance of developing robust earthquake monitoring systems for accurately determining the event origin times and hypocenter locations. In addition, the rapid and reliable characterization of ongoing earthquakes is a crucial, yet challenging, task for developing seismic hazard mitigation tools like earthquake early

warning (EEW) systems [1]. While classical methods have been widely adopted to design EEW systems, challenges remain to pinpoint hypocenter locations in real-time largely due to limited information in the early stage of earthquakes. Among various key aspects of EEW, timeliness is a crucial consideration and additional efforts are required to further improve the hypocenter location estimates with minimum data from 1) the first few seconds after the P-wave arrival and 2) the first few seismograph stations that are triggered by the ground shaking.

The localization problem can be resolved using a sequence of detected waves (arrival times) and locations of seismograph stations that are triggered by ground shaking. Among various network architectures, the recurrent neural network (RNN) is capable of precisely extracting information from a sequence of input data, which is ideal for handling a group of seismic stations that are triggered sequentially following the propagation paths of seismic waves. This method has been investigated to improve the performance of real-time earthquake detection [2] and classification of source characteristics. Other machine learning based strategies have also been proposed for earthquake monitoring. Comparisons between traditional machine learning methods, including the nearest neighbor, decision tree, and the support vector machine, have also been made for the earthquake detection problem [3]. However, a common issue in the aforementioned machine learning based frameworks is that the selection of input features often requires expert knowledge, which may affect the accuracy of these methods. Convolution neural networks-based clustering methods have been used to regionalize earthquake epicenters [4] or predict their precise hypocenter locations [5]. In the latter case, three-component waveforms from multiple stations are exploited to train the model for swarm event localization.

In this study, we propose a RF-based method to locate earthquakes using the differential P-wave arrival times and station locations (Figure 1). The proposed algorithm only relies on P wave arrival times detected at the first few stations. Its prompt response to earthquake first arrivals is critical for rapidly disseminating EEW alerts. Our strategy implicitly considers the influence of the velocity structures by incorporating the source-station locations into the RF model. We evaluate the proposed algorithm using an extensive seismic catalog from Japan. Our test results show that the RF model is capable of determining the locations of earthquakes

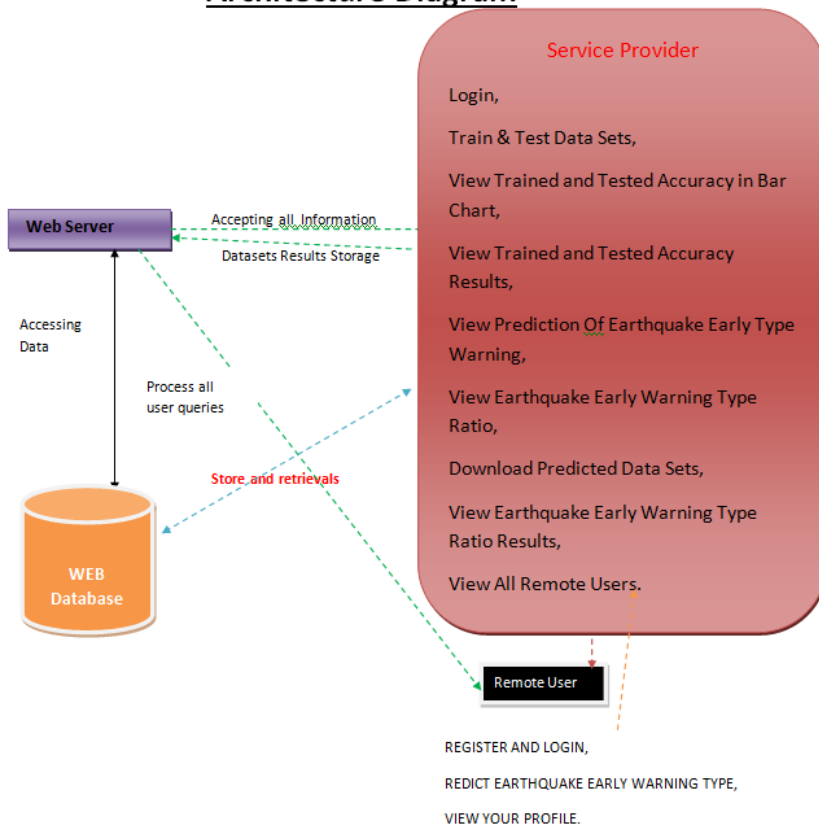
accurately with minimal information, which sheds new light on developing efficient machine learning.

2. EXISTING SYSTEM

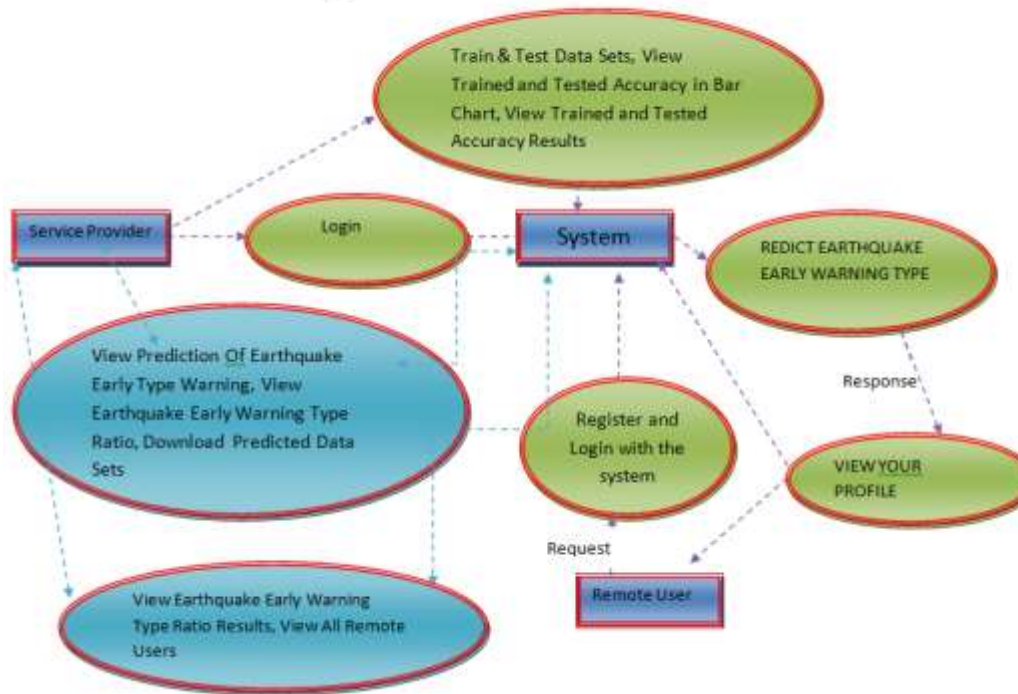
Earthquake early warning (EEW) systems are required to report earthquake locations and magnitudes as quickly as possible before the damaging S wave arrival to mitigate seismic hazards. Deep learning techniques provide potential for extracting earthquake source information from full seismic waveforms instead of seismic phase picks.

We developed a novel deep learning EEW system that utilizes fully convolutional networks to simultaneously detect earthquakes and estimate their source parameters from continuous seismic waveform streams. The system determines earthquake location and magnitude as soon as very few stations receive earthquake signals and evolutionarily improves the solutions by receiving continuous data. We apply the system to the 2016 M 6.0 Central Apennines, Italy Earthquake and its first-week aftershocks. Earthquake locations and magnitudes can be reliably determined as early as 4 s after the earliest P phase, with mean error ranges of 8.5–4.7 km and 0.33–0.27, respectively.

Architecture Diagram



➤ Data Flow Diagram :



3. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

5. CONCLUSION

We use the P-wave arrival time differences and the location of the seismic stations to locate the earthquake in a real-time way. Random forest (RF) has been proposed to perform this regression problem, where the difference latitude and longitude between the earthquake and the seismic stations are considered as the RF output. The Japanese seismic area is used as a case of study, which demonstrates very successful performance and indicates its immediate applicability. We extract all the events having at least five P-wave arrival times from nearby seismic stations. Then, we split the extracted events into training and testing datasets to construct a machine learning model. In addition, the proposed method has the ability to use only three seismic stations and 10% of the available dataset for training, still with encouraging performance, indicating the flexibility of the proposed algorithm in real-time earthquake monitoring in more challenging areas. Despite the sparse distribution of many networks around the world, which makes the random forest method difficult to train an effective model, one can use numerous synthetic datasets to compensate for the shortage of ray paths in a target area due to insufficient catalog and station distribution.

6. REFERENCES

- [1] Q. Kong, R. M. Allen, L. Schreier, and Y.-W. Kwon, “Myshake: A smartphone seismic network for earthquake early warning and beyond,” *Science advances*, vol. 2, no. 2, p. e1501055, 2016.
- [2] T.-L. Chin, K.-Y. Chen, D.-Y. Chen, and D.-E. Lin, “Intelligent real-time earthquake detection by recurrent neural networks,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5440–5449, 2020.

- [3] T.-L. Chin, C.-Y. Huang, S.-H. Shen, Y.-C. Tsai, Y. H. Hu, and Y.-M. Wu, “Learn to detect: Improving the accuracy of earthquake detection,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8867–8878, 2019.
- [4] O. M. Saad, A. G. Hafez, and M. S. Soliman, “Deep learning approach for earthquake parameters classification in earthquake early warning system,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2020.
- [5] X. Zhang, J. Zhang, C. Yuan, S. Liu, Z. Chen, and W. Li, “Locating induced earthquakes with a network of seismic stations in oklahoma via a deep learning method,” *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.
- [6] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [7] S. M. Mousavi, W. L. Ellsworth, W. Zhu, L. Y. Chuang, and G. C. Beroza, “Earthquake transformeran attentive deep-learning model for simultaneous earthquake detection and phase picking,” *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.
- [8] S. M. Mousavi and G. C. Beroza, “A Machine- Learning Approach for Earthquake Magnitude Estimation,” *Geophysical Research Letters*, vol. 47, no. 1, p. e2019GL085976, 2020

FADOHS: FRAMEWORK FOR DETECTION AND INTEGRATION OF UNSTRUCTURED DATA OF HATE SPEECH ON FACEBOOK USING SENTIMENT AND EMOTION ANALYSIS

Grandhi Pavani Seshalakshmi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

Hate speech is a form of expression that assaults a person or a community based on race, origin, religion, sexual orientation, or other attributes. Although it can be expressed in multiple ways, both online and offline, the increasing popularity of social media has exponentially increased both its use and severity. Therefore, the aim of this research is to locate and analyze the unstructured data of selected social mediaposts that intend to spread hate in the comment sections. To address this issue, we propose a novel framework called FADOHS, which combines data analysis and natural language processing strategies, to sensitize all social media providers to the pervasiveness of hate on social media. Specifically, we use sentiment and emotion analysis algorithms to analyze recent posts and comments on these pages. Posts suspected of containing dehumanizing words will be processed before fed to the clustering algorithm for further evaluation. According to the experimental results, the proposed FADOHS framework is able to surpass the state-of-the-art approach in terms of precision, recall, and F1 scores by approximately 10%.

1. INTRODUCTION

Mark Zuckerberg, CEO of Face book, once commented ``Hate speech and racism do not have a place on Face book" [1]. Although Face book has adopted various artificial intelligence (AI) techniques to prevent hate speech on its platform, some issues persist. For example, as stated by the company when they published statistics on the crackdown on hate speech, ``For hate speech,

our technology still does not work well, so it needs to be checked by our review teams. We removed 2.5 million pieces of hate speech in Q1 2018, where 38% of which was "tagged by our technology" [2]. The most persistent question in this pursuit is very difficult to address using AI alone: *What is hate speech?* This question draws continuous discussion, which brings various definitions of hate speech; for example, "Hate speech is public expressions which spread, incite, promote or justify hatred, discrimination or hostility toward a specific group" [3] and "We define hate speech as a direct attack on people based on what we call protected characteristics – race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability" [4].

Face book admitted that the problem comes from the fact that AI is not yet sufficiently advanced to determine whether someone is promoting hate or simply describing an experience [5]. Sara Chinnasamy and Norain Abdul Manaf mentioned that hate speech can also be promoted in subtle ways, such as discussing controversial topics to elicit hate comments [6]. Anat Ben-David and Ariadna Matamoros- Fernandez argued that despite the efforts of Face book, hate comments abound. The authors remarked that many users express their subliminal hatred through vicious messages or commentaries. These posts, which are not uncovered by Face book algorithms, are common on the platform. The authors also concluded that overt hate speech and covert discriminatory practices remain pervasive on Face book, despite policies and efforts designed to combat them [7]. Having defined hate speech, we can establish a framework for its investigation. The authors of "Hate Me, Hate Me Not: Hate Speech on Face book" [8] have proposed several classification methods to distinguish among different types of hate speech. More specifically, they leverage morpho-syntactic features, sentiment polarity, and word-embedded lexicons to design and implement two classifiers for Italian. Their framework utilizes support vector machines (SVMs) and long short-term memory (LSTM) networks. This study was premised on the concept outlined in Del Vigna *et al.*'s study and our understanding of hate speech. We investigated preliminary ways to unearth hate speech on the Face book platform, especially covert speech in the comment section of posts discussing controversial topics.

2. EXISTING SYSTEM

Ben-David and Matamoros-Fernandez's related study on overt hatred and covert disrespectful practices on the Internet [7] is based on network and multimodal analyses. It studies information

and images found on social media, and it retrieves data from various Facebook pages with content associated with hate speech using tools such as Netvizz [12]. In our proposed framework, we examined the dataset [13] using a Facebook graph application programming interface (API) and emotional analysis [14].

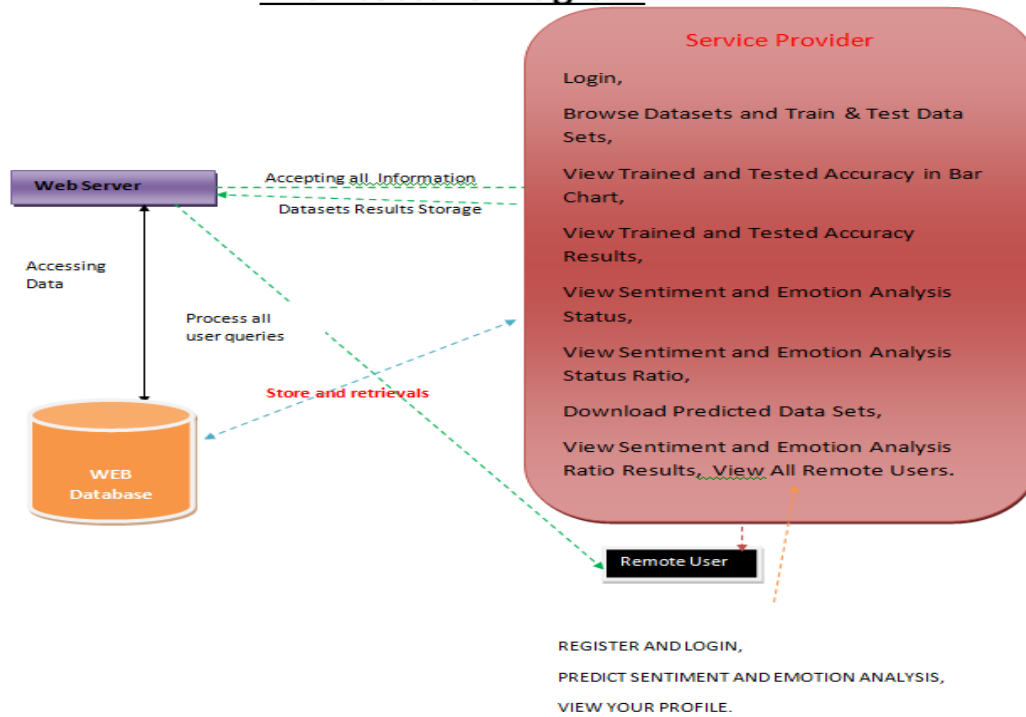
In another related study, the authors used the valence-aware Dictionary and sEntiment Reasoner (VADER) tool as a simple rule-based model for general sentiment analysis [15]. The VADER tool uses both qualitative and quantitative methods to analyze and validate the data [16]_[28]. Subsequently, data validation is attuned to sentiments using microblog_like information. We also used the VADER tool for SA (sentiment analysis). However, unlike the research in [15], we incorporated the JAMMIN tool to perform emotional analysis experiments and, specifically, track posts with negative comments. The research [8] created a typology of abhorrence based on different "loathing levels." The author's utilized morphogrammatical highlights, notion extremity, and word-installed dictionaries to plan and actualize two classifiers for the Italian language. Furthermore, they used the SVM and LSTM [8]. However, our approach is designed to uncover hate discourse on Facebook, particularly the "unmistakable" manifestations of hatred posted as remarks on divisive topics (e.g., immigration, religion, and race). The study in [29] highlights future issues that Facebook and Twitter would face in identifying hate speech on their respective platforms. Their tool was crowd-sourcing. Although their framework has not been fully evaluated, their study includes a quality-of-service (QoS) assessment for platform providers. Thus, they developed an intuitive tool for cracking down on hate speech. However, although their tool can identify information that does not adhere to QoS policies, we believe such a format is inefficient because of the use of Python programming tools [29]. In this study, we implemented a procedure to filter hate-filled posts and comments on social media. The related research describes the concept of "platform racism" - an emerging form of racial prejudice emanating from social media pages [30]. Hatred itself is a form of discrimination, depending on the culture associated with a particular group. Annotation of the study that suggested a possible algorithm for compiling the contents was provided. Although the experiment revealed important trends relevant to our specific research question, we focused on Facebook as a platform and utilized both data extraction and experimental setup, according to the seed pages. Online trends leading to offline consequences [31] examined the link between social media and hate crimes using Facebook data. Interestingly, it was concluded that social media is often used as the

propagation mechanism of hate [31]. Although such a study is important, we believe that the data used by researchers could be strengthened by analyses such as ours, in which social media analytics are used to target and identify negative comments posted on certain hate-promoting Facebook pages [13].

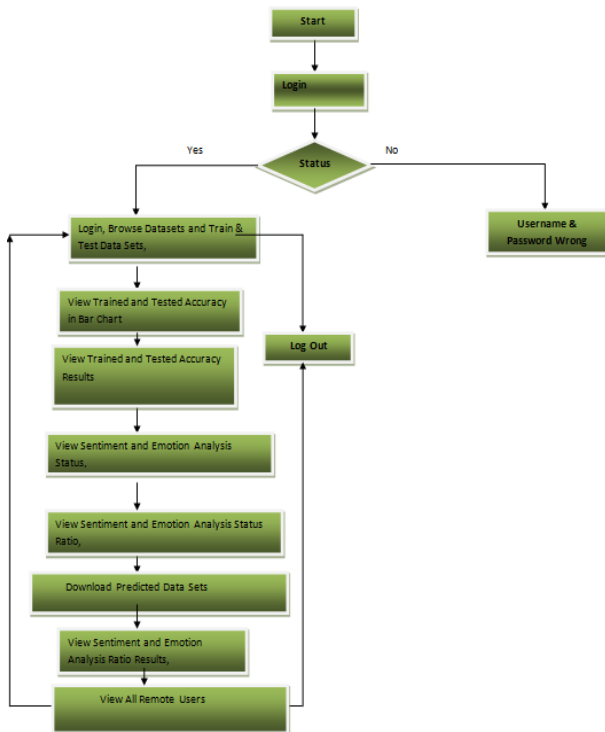
3. PROPOSED SYSTEM

In contrast, in this linked study [32], the authors examined the most effective methods for detecting hate speech in written text. Based on this survey and our framework, we conducted various tests to compare the accuracy of the three best methods. The authors of [33] proposed an optimization approach based on meta heuristic searching. The ant lion optimization (ALO) and moth flame optimization (MF) algorithms were designed for the HSD problem. This is the first attempt to use optimization algorithms as solution-search strategies for automatic HSDs. An efficient representation scheme and a flexible fitness function were designed for this purpose. However, the FADHOS approach not only identifies unstructured data from Facebook allegedly promoting hate speech, such as commonly discussed topics, but also identifies and integrates them by topic in clusters using sentiment and emotion analysis. In 2019, OpenAI released generative pre-trained transformer 2 (GPT-2) models [34]_[45]. These were built using transformer decoder blocks. We tested the quality of our dataset by investigating the best dataset of our framework (moderate level of hate speech dataset), in which we performed several experiments using the Nobel model of OpenAI - the GPT-2 model [34]. The major goal of this experiment was to determine the level at which our hate speech dataset can enhance the performance of the GPT-2 model. The literature review in this section provides the background for this study as well as a potential motivation for further investigation. The primary commitment of our research is to accurately and efficiently locate social media pages that discuss sensitive topics and establish a reliable system that categorizes posts and incorporates unstructured information with frequently discussed themes that intentionally or unintentionally spread hate discourse.

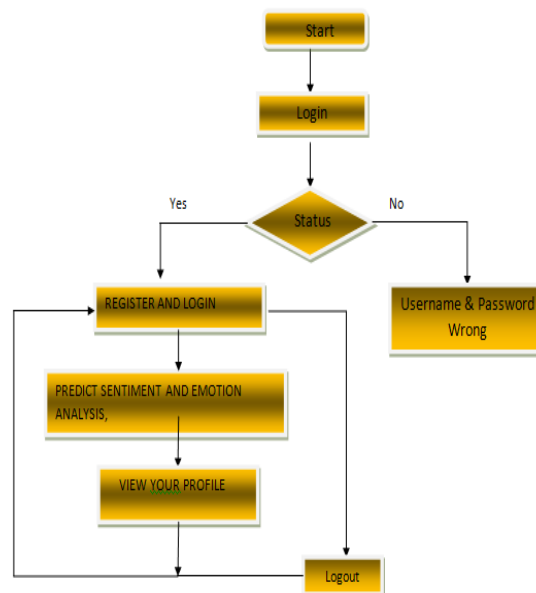
Architecture Diagram



➤ Flow Chart : Service Provider



➤ Flow Chart : Remote User



4. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSION

In this study, we propose FADOHS, which identifies and integrates unstructured data from Face book pages that allegedly promote hate speech, such that the usual topics discussed can be identified. Initially, this issue was challenging because non-personal pages and accounts on Face book tend to avoid using overtly explicit terms in posts to avoid being removed from the platform or to avoid criticism. Nevertheless, many pages still manage to stir negative emotions and appear to promote hate speech among their followers by discussing controversial topics

while keeping their vocabulary fairly innocuous. The proposed framework provides a novel solution to cluster posts and comments, detecting highly discussed topics that generate hate speech and identifying hate speech.

FADOHS combines graph analysis, dictionaries, sentiment/emotion analysis as well as clustering methods to cluster and analyze posts that may contain hate speech. To properly address the hate speech issue, we commence our analysis with a small set of pages that are known to discuss sensitive topics potentially eliciting hate comments. From this analysis, we can build three levels of direct social graphs and identify prominent pages using graph analysis. Using predetermined dictionaries, sentiment, and emotion analysis, we isolate posts containing a certain level of negativity in the comments. The output leads us to confidently conclude that unstructured data could be identified and integrated from hate speech-promoting pages. The next essential phase aims at categorizing these data, which is achieved by applying the K-means clustering algorithm, followed by testing varying configurations to unearth intuitively groups of topics. We then manually analyze the posts that fall within each group and assign a manual label to each cluster. By Comparing the manual label to the cluster centroids, we can conclude that both variables are matched, thereby confirming the effectiveness of our approach.

Our experiments demonstrate that several pages allegedly promoting hate speech and related topics are identified from a small set of seeds. This work represents a clear example of taking unstructured data such as Face book posts, and applying a framework for meaningful analysis. According to the experimental results, the proposed FADOHS framework is able to surpass the state-of-the-art approach in terms of precision, recall, and F1 scores by approximately 10%.

In future studies, we plan to further utilize our framework not only on comments but also their replies, in an attempt to accurately identify individuals who are suspected of promoting hate speech. Long-term benefits can be extremely valuable because this may be able to detect cyber bullies and cyber terrorists. We would also like to perform a more in-depth examination of the emotion filtering and clustering findings to identify the most dependable setup for optimizing outcomes.

7. REFERENCES

- [1] *Zuckerberg Refugee Crisis: Hate Speech Has, Place Facebook*, Street Guardian, Honolulu, HI, USA, 2010.
- [2] Fortune. (2018). *Facebook Removed 2.5 Million Pieces Hate Speech 1st Quarter*. Accessed: Jul. 16, 2018. [Online]. Available: <https://fortune.com/2018/05/15/facebook-hate-speech-removals/>.
- [3] ILGA. (2018). *Hate Crime & Hate Speech*. Accessed: May 6, 2018. [Online]. Available: <https://www.ilga-europe.org/what-we-do/ouradvocacy-work/hate-crime-hate-speech>
- [4] Facebook. (2020). *Community Standards Home*. Accessed: May 11, 2018. [Online]. Available: <https://www.facebook.com/communitystandards/>.
- [5] CNBC. (2020). *Facebook's Artificial Intelligence Still Has Trouble Finding Hate Speech_But it Finds a Lot of Nudity*. Accessed: May 11, 2018. [Online]. Available: <https://www.cnbc.com/2018/05/15/facebookartificial-intelligence-still-finds-it-hard-to-identify-hate-speech.html>
- [6] S. Chinnasamy and N. A. Manaf, "Social media as political hatred mode in Ts 2018 general election," in *SHS Web Conf.*, vol. 53, 2018, p. 2005.
- [7] A. Matamoros-Fernández and J. Farkas, "Racism, hate speech, and social media: A systematic review and critique," *Telev. New Media*, vol. 22, no. 2, pp. 205_224, Feb. 2021.
- [8] F. Del Vigna, A. Cimino, F. Dell-Torletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in *Proc. 1st Italian Conf. Cybersecur. (ITASEC)*, Venice, Italy, 2017, pp. 86_95.
- [9] M. Ahmed, R. Seraj, and S. M. S. Islam, "The K-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020.

- [10] A. Moubayed, M. Injadat, A. Shami, and H. Lut_yya, ``Student engagement level in an e-Learning environment: Clustering using K-means," *Amer. J. Distance Educ.*, vol. 34, no. 2, pp. 137_156, Apr. 2020.
- [11] Z. Lv, T. Liu, J. A. Benediktsson, and H. Du, ``Novel land cover change detection method based on K-means clustering and adaptive majority voting using bitemporal remote sensing images," *IEEE Access*, vol. 7, pp. 34425_34437, 2019.
- [12] D.Kucukusta, M. Perelygina, andW. S. Lam, ``CSR communication strategies and stakeholder engagement of upscale hotels in social media," *Int. J. Contemp. Hospitality Manage.*, vol. 31, no. 5, pp. 2129_2148, May 2019.
- [13] A. Rodriguez, C. Argueta, and Y.-L. Chen, ``Automatic detection of hate speech on Facebook using sentiment and emotion analysis," in *Proc. Int. Conf. Artif. Intell. Inf. Commun. (ICAIIIC)*, Feb. 2019, pp. 169_174.



A SYSTEMATIC REVIEW OF PREDICTING ELECTIONS BASED ON SOCIAL MEDIA DATA

Gudimetla Charan Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

The way politicians communicate with the electorate and run electoral campaigns was reshaped by the emergence and popularization of contemporary social media (SM), such as Facebook, Twitter, and Instagram social networks (SNs). Due to the inherent capabilities of SM, such as the large amount of available data accessed in real time, a new research subject has emerged, focusing on using the SM data to predict election outcomes. Despite many studies conducted in the last decade, results are very controversial and many times challenged. In this context, this article aims to investigate and summarize how research on predicting elections based on the SM data has evolved since its beginning, to outline the state of both the art and the practice, and to identify research opportunities within this field. In terms of method, we performed a systematic literature review analyzing the quantity and quality of publications, the electoral context of studies, the main approaches to and characteristics of the successful studies, as well as their main strengths and challenges and compared our results with previous reviews. We identified and analyzed 83 relevant studies, and the challenges were identified in many areas such as process, sampling, modeling, performance evaluation, and scientific rigor. Main findings include the low success of the most-used approach, namely volume and sentiment analysis on Twitter, and the better results with new approaches, such as regression methods trained with traditional polls. Finally, a vision of future research on integrating advances in process definitions, modeling, and evaluation is also discussed, pointing out, among others, the need for better investigating the application of state-of-the-art machine learning approaches.

1. INTRODUCTION

SOCIAL media (SM) has played a central role in politics and elections throughout this decade. We have entered a new era mediated by SM in which politicians conduct permanent campaigns without geographic or time constraints, and additional information about them can be obtained not only by the press but also directly from their profiles on social networks (SNs) and through other people sharing and amplifying their voices

on SM. In this new scenario, SM is used extensively in electoral campaigns [1], and an online campaign's success can even decide elections. In practice recent examples of SM engagement and electoral success include the 2016 U.S. residential election, when Donald Trump focused his campaign on free-media marketing [2], and the 2018 Brazilian presidential election, when the candidate with more SM



engagement but little exposition on traditional media was elected [3].

Moreover, in some way, it is possible to measure how a politician's message is spreading over SM and try to estimate how much attention a candidate is receiving or how many people are talking about a candidate. Thus, considering a large amount of data available in real time and the low cost of their acquisition, combined with the advances of techniques for processing them, a new research subject has emerged, focusing on using the SM data to predict election outcomes.

Only 2 years after Twitter and Facebook's launch for the general public, studies to predict elections based on the SM data started to be published: Tilton [4] can be considered a preliminary study focused on student elections, published in 2008. In addition, two studies published in 2010 at the same forum, Tumasjan *et al.* [5] and O'Connor *et al.* [6], are considered seminal studies regarding predicting political elections based on SM. The former presented an approach based on the volume counting of posts on Twitter (tweets), and the latter was based on the sentiment extracted from those tweets.

2. EXISTING SYSTEM

In 2013, Kalampokis *et al.* [29] presented a systematic review aiming to understand the predictive power of SM, not only in the electoral context. By analyzing 52 studies, 11 regarding election predictions, they identified that main approaches were based on volume, sentiment, and user profiling.

In addition, the use of predictive analysis using linear regression was identified, but

not on the studies related to the political context. In addition, they verified that 40% of studies that had used sentiment-related variables challenged SM predictive power, i.e., was not successful, and this number increased to 65% in the case of lexicon-based approaches.

Finally, they emphasized the lack of predictive analytics evaluation and controversial results of electoral predicting studies. In the same year, Gayo-Avello [30] presented a study that we consider the first review specifically on predicting elections with SM, focused on Twitter. By analyzing ten previous studies from 2010 to 2013, he concluded that "the presumed predictive power regarding electoral prediction has been somewhat exaggerated." Moreover, as in [29], he identified volume and sentiment analysis as main approaches and the need to use more up-to-date methods for sentiment analysis. In addition, he expanded the list of challenges, such as the dependency of arbitrary decisions made by researchers regarding keywords, parties, candidates and selection of the data collection period, and problems related to Twitter, such as demographic and self-selection bias, and bias related to spam, misleading propaganda, and astroturfing. He ended the study pointing out that

regression models may be a future direction. In 2015, studies from Prada [31] and O'Leary [32] presented in general lines the main approaches for predicting using Twitter in many different domains, and briefly described a few studies related to election predictions (2 and 11 studies,



respectively). In 2018, Kwak and Cho [33] presented the results of a survey including 69 papers that supported the argument that SM can be used in understanding political agenda, rather than in election forecast.

Ultimately, most recent studies [34], [35] presented limited nonsystematic surveys, both analyzing 13 papers, adding some arguments to the original review from Gayo-Avello [30]. Koli *et al.* [34] argued that prediction using Twitter can have better results in developed countries, due to a higher literacy rate and internet access, than in developing countries. In addition, Bilal *et al.* [35] considered the challenges of sentiment analysis in languages other than English. Despite these new arguments, recent studies fail to identify novel approaches as well as approaches using SM other than Twitter and Facebook.

Disadvantages

- 1) Data uncover is the main weakness in the existing system.
- 2) The system doesn't have a techniques to test and train for large scale data sets.

3. PROPOSED SYSTEM

- ❖ The proposed system aims at identifying the electoral contexts being studied, such as the year and country in which the election took place and the type of election. This question is intended to ascertain whether the studies are best suited or paying attention to any particular electoral context.
- ❖ The objective of this proposed system is to identify the main approaches used, their main characteristics, how they are

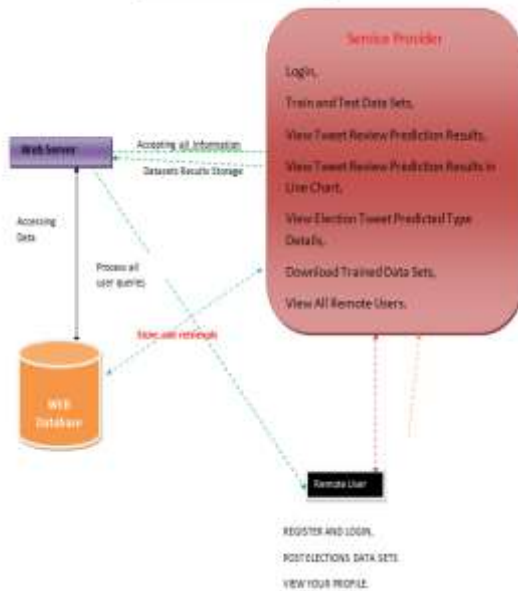
modeled and applied to predict elections, and what are the metrics used to assess their performance.

- ❖ The objective of this proposed system is to identify the main characteristics of allegedly successful studies in order to identify in which specific contexts, which approaches, and which factors yield effective results.
- ❖ After studying the context, approaches, and characteristics of successful studies, the answer to this question aims to summarize the main perceived strengths, weaknesses, challenges, and opportunities in this new research area to guide future research.

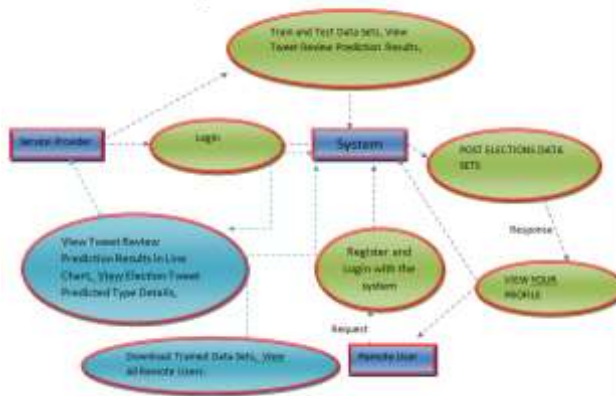
Advantages

- Unique studies include approaches based on prediction market, cluster detection, centrality score, statistical physics of complex networks, and analysis of groups of supporters, solely or in combination with previously described approaches.
- The system performed statistical tests on results to verify whether they were statistically significant.

Architecture Diagram



Data Flow Diagram



5. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

6. CONCLUSION

This study collected more than 500 articles, 90 of which were focused on predicting elections based on SM data, investigating, and summarizing how this new research field has evolved since 2008. Among these studies, 83 are primary studies aiming at predicting elections and seven are surveys or reviews of past studies.

The results show that the number of publications in this area is increasing and research is spread across 28 countries from all continents. Nevertheless,



there cannot yet be found any prominent researchers, research groups, or clusters performing sustainable research in the area. In addition, there was no identification of a common well-known forum for publication on this subject, and results are spread across many forums.

Moreover, as main challenges, we identified issues in four areas. Regarding processes, we highlight the lack of well defined, replicable, and generalizable processes, and lack of prediction capabilities during the campaign. In sampling, issues are mainly related to the fact that SNs and Twitter data do not represent representative samples, and studies were performed with many arbitrary data collection choices. Regarding modeling, we found difficulties crossing data from multiple networks, the high susceptibility to volume manipulation, the lack of use of state-of-the-art ML techniques and technical modeling weaknesses. And considering performance evaluation and scientific rigor of studies, the lack of statistical analysis of results and of meaningful comparison with related works are also main issues.

Finally, the study presented the authors' point of view on the future directions of predicting elections using SM data in three axes: process definitions, model definitions and sampling, and study evaluation. As main directions, we highlight the need for repeatable processes based on well-known methodologies, for example, CRISP-DM or SEMMA; the use of state-of-the-art methods for regression based on machine learning that can combine data from multiple SNs, such as ANN; and

the use of statistical tests for results evaluation, such as Wilcoxon signedrank test and others.

REFERENCES

- [1] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *J. Inf. Technol. Politics*, vol. 13, no. 1, pp. 72–91, Jan. 2016.
- [2] P. L. Francia, "Free media and Twitter in the 2016 presidential election: The unconventional campaign of Donald Trump," *Social Sci. Comput. Rev.*, vol. 36, no. 4, pp. 440–455, Aug. 2018.
- [3] K. Brito, N. Paula, M. Fernandes, and S. Meira, "Social media and presidential campaigns—preliminary results of the 2018 Brazilian presidential election," in *Proc. 20th Annu. Int. Conf. Digit. Government Res.*, Jun. 2019, pp. 332–341.
- [4] S. Tilton, "Virtual polling data: A social network analysis on a student government election," *Webology*, vol. 5, no. 4, pp. 1–8, 2008.
- [5] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: What 140 characters reveal about political sentiment," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 1–8.
- [6] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. 4th Int. AAAI Conf. Weblogs Social Media*, 2010, pp. 1–8.
- [7] E. Sang and J. Bos, "Predicting the 2011 Dutch senate election results with Twitter," in *Proc. Workshop Semantic Anal. Social Media*, 2012, pp. 53–60.



[8] A. Ceron, L. Curini, S. M. Iacus, and G. Porro, “Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens’ political preferences with an application to Italy and France,” *New Media Soc.*, vol. 16, no. 2, pp. 340–358, Mar. 2014.

[9] K. Singhal, B. Agrawal, and N. Mittal, “Modeling Indian general elections: Sentiment analysis of political Twitter data,” in *Information Systems Design and Intelligent Applications* (Advances in Intelligent Systems and Computing). New Delhi, India: Springer, 2015.

[10] N. Dwi Prasetyo and C. Hauff, “Twitter-based election prediction in the developing world,” in *Proc. 26th ACM Conf. Hypertext Social Media (HT)*, 2015, pp. 149–158.

A DEEP LEARNING APPROACH FOR ROBUST DETECTION OF BOTS IN TWITTER USING TRANSFORMERS .

Gurram Karimunisa (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

During the last decades, the volume of multimedia content posted in social networks has grown exponentially and such information is immediately propagated and consumed by a significant number of users. In this scenario, the disruption of fake news providers and bot accounts for spreading propaganda information as well as sensitive content throughout the network has fostered applied research to automatically measure the reliability of social networks accounts via Artificial Intelligence (AI). In this paper, we present a multilingual approach for addressing the bot identification task in Twitter via Deep learning (DL) approaches to support end-users when checking the credibility of a certain Twitter account. To do so, several experiments were conducted using state-of-the-art Multilingual Language Models to generate an encoding of the text-based features of the user account that are later on concatenated with the rest of the metadata to build a potential input vector on top of a Dense Network denoted as Bot-DenseNet. Consequently, this paper assesses the language constraint from previous studies where the encoding of the user account only considered either the metadata information or the metadata information together with some basic semantic text features. Moreover, the Bot-DenseNet produces a low dimensional representation of the user account which can be used for any application within the Information Retrieval (IR) framework.

1.INTRODUCTION

In recent years, social media platforms such as Twitter or Face book have gained a large level of both popularity and influence among millions of users due to the benefits of publishing, propagating and exchanging large volumes of multimedia content along the network. Therefore, these platforms allow users to establish a digital community as remarked in [22], which has made possible not only to discover and embrace new relationships but to maintain and boost existing ones.

On the other hand, due to both the great influence these platforms have on the lifestyle of people and its evolving as a potential communication tool, they have exponentially promoted its attraction for marketing and commercial purposes by analysing the behaviour and opinion of users in different topics or events such as political elections. Consequently, numerous research studies have been fostered in the social media field with different purposes including sentiment analysis [35], traffic control [48], or consumer behaviour mining [4].

However, the considerable growth of social media platforms has also provoked the desire of altering people's opinion in certain topics by spreading propaganda or bias information. Many of these controlling procedures are carried out by Bots which are widely described in numerous investigations [31], [32], [40] such as automatic systems which are capable of generating and spreading multimedia content throughout the network without the supervision of a human being. Furthermore, with the disruptive growth of Artificial Intelligence (AI) algorithms, the identification of bots or nonreliable sources has become a crucial challenge to be investigated. It raised many studies and publications with the goal of building robust automatic systems to improve the quality of experience of consumers in such platforms by reducing their privacy risks as well as increasing the trustworthiness on the platform itself at the same time. Therefore, this paper aims to contribute to the state-of-the-art in this field by proposing a novel method for automatically

2.EXISTING SYSTEM

In [11], authors annotated more than 8000 accounts and proposed a classifier which achieved a considerable level of accuracy for such set of samples. Additionally, [38] presented a model for Twitter bot detection based on a large number of metadata from the account to perform the classification. More recently, several scientific studies have incorporated more annotated samples to support this research such as

[27], [44], [45] including some procedures for achieving better level of accuracy by strategically selecting a subset of training samples that better generalize the problem. In [22], a language-agnostic approach is employed to identify potential features to distinguish between human and bot accounts. The model is then trained and validated using over 8000 samples distributed in an unbalanced fashion and its performance reaches an accuracy of 98%.

Moreover, authors in [32] proposed a 2D Convolutional Network model based on user-generated contents for detecting bots from human accounts including its gender (male, female account) covering both Spanish and English languages. A similar goal is explored by authors in [40], where both Word and Character N-Grams are employed as main features to perform the classification. A

different manner of addressing the problem was recently proposed by authors in [5], where novel altmetrics data to investigate social networks are analysed and they are used to train a Graph Convolutional Network (GCN) which reaches over 70% of accuracy in this task. On the other hand, authors in [34] presented a novel one-class classifier to enhance Twitter bot detection without any requiring previous information about them.

Disadvantages

- 1) .The system doesn't have A MULTILINGUAL APPROACH FOR USER ACCOUNT ENCODING VIA TRANSFORMERS.
- 2). The system couldn't implement to detect the following (i) Level of activity, (ii) Level of popularity, (iii) Profile information.

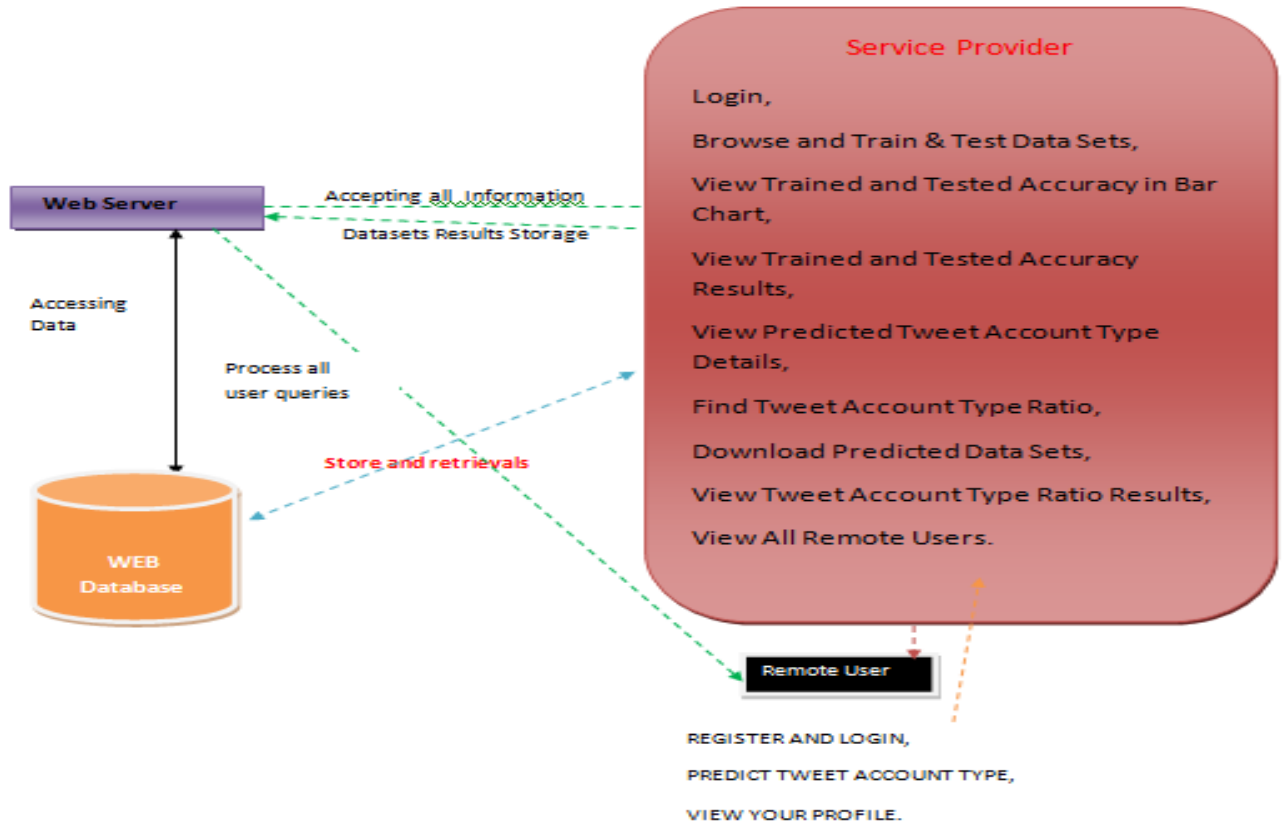
3.PROPOSED SYSTEM

Present a multilingual approach for addressing the bot identification task in Twitter via Deep learning (DL) approaches to support end-users when checking the credibility of a certain Twitter account. To do so, several experiments were conducted using state-of-the-art Multilingual Language Models to generate an encoding of the text-based features of the user account that are later on concatenated with the rest of the metadata to build a potential input vector on top of a Dense Network denoted as Bot-DenseNet. Consequently, this paper assesses the language constraint from previous studies where the encoding of the user account only considered either the metadata information or the metadata information together with some basic semantic text features. Moreover, the Bot-DenseNet produces a low-dimensional representation of the user account which can be used for any application within the Information Retrieval (IR) framework.

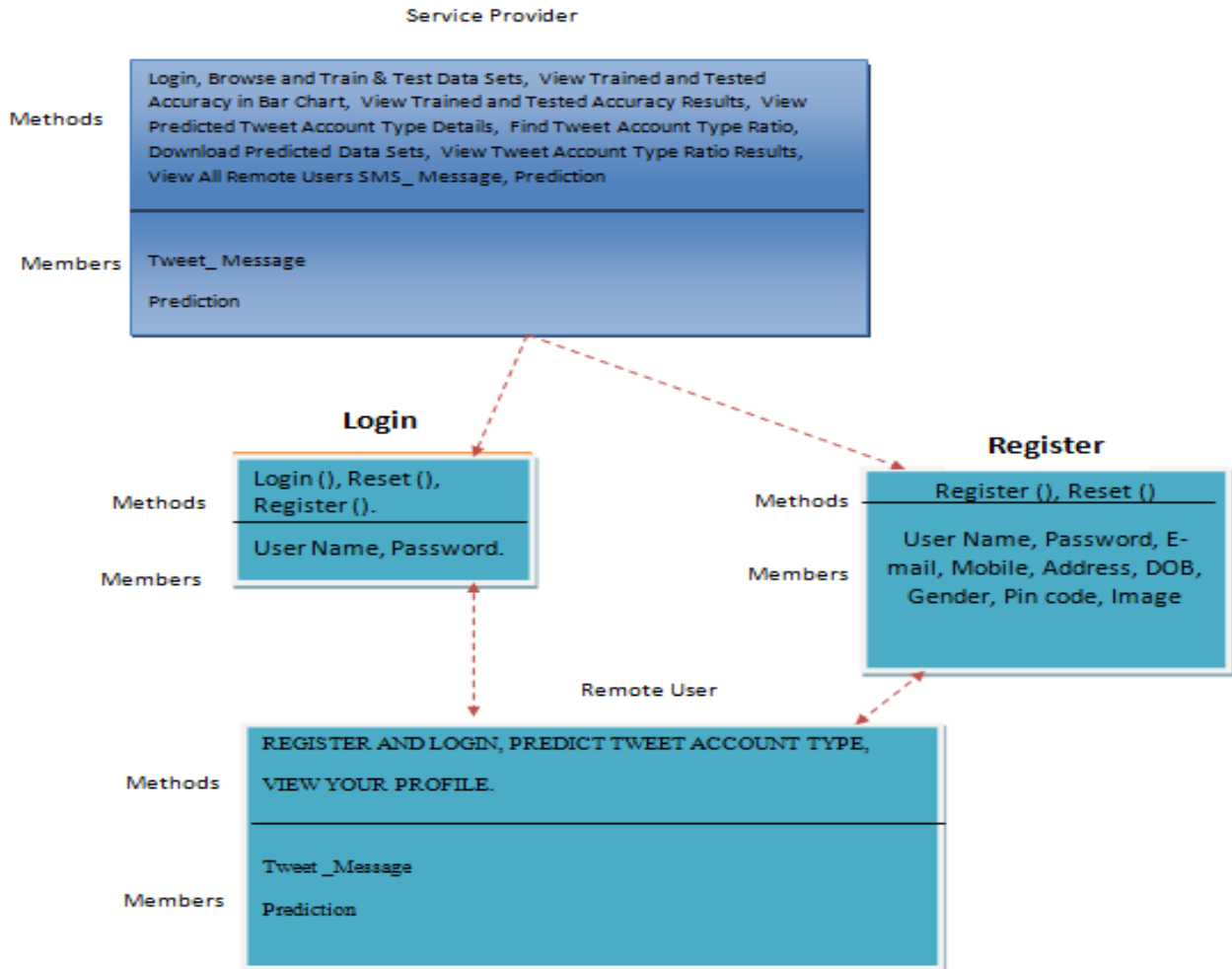
Advantages

- (i) a preprocessing stage where a multilingual input vector of the user account is generated
- (ii) a final decision system for identifying whether the account has a normal or abnormal behavior according to existence patterns in the input vector generated during the first stage.

Architecture Diagram



➤ **Class Diagram :**



4.SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5.CONCLUSION

In this paper, a robust solution for detecting Bots in Twitter accounts has been described. In particular, this study has taken advantage of Transfer learning techniques via powerful state-of-the-art NLP models such as Transformers to extract compact multilingual representations of the text-based features associated with user accounts. By doing so, several constraints presented in previous studies related to process text-based features to improve the input feature vector from multiple languages were mitigated.

Furthermore, by employing the text encodings along with additional metadata on top of a dense-based neural network, a final classifier named as Bot-DenseNet has been trained and validated using a large set of samples collected via the Twitter API. More specifically, several experiments were conducted using different combinations of Word Embeddings, document embeddings (Pooling and LSTMs) and Transformers to obtain a single vector regarding the text-based features of the user account. Subsequently, a detailed comparison of the performance of the proposed classifier when using these approaches of Language Models as part of the input has been presented to investigate which input vector provides the best result in terms of performance simplicity in the generation of decision boundaries and feasibility.

In particular, the comparison of these experiments suggested that the Bot-DenseNet achieves the most adequate trade-off between performance and feasibility when using the so called Roberta Transformer as part of the input feature vector.

Consequently, this paper provides two main contributions to the scientific community including a DL model for automatically detecting bots as well as a robust manner of representing any Twitter account as a low-dimensional feature vector throughout an intermediate layer of the aforementioned model. Moreover, this compact representation of the Twitter account can be used as a baseline for recommender or search engines, similarity analysis or any other application related with social media mining.

Finally, this study also remarks the outstanding performance of novel Transformers in downstream NLP tasks as the one presented, by providing a more robust input vector which leads the final classifier model to be more capable of extracting relevant low-level features from it. As Future work, the latest Transformers such as the GPT-3 [17] and T5 [33] will be considered for generating the input vector of the proposed DL model in order to compare the performance with the work described on this paper. Moreover, novel approaches such the one described by authors in [24] to automatically generate non-parametric Two-Sample tests based on the so-called Maximum Mean Discrepancy (MMD) [18], will be considered once all the user embeddings are generated, to find discrepancies and similarities between the distributions of both bots and non-bots embeddings.

6.REFERENCES

- [1] Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics.
- [2] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 54–59, 2019.
- [3] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, 2018.
- [4] Ahmed Sulaiman M Alharbi and Elise de Doncker. Twitter sentiment analysis with a deep neural network: An enhanced approach using user behavioral information. Cognitive Systems Research, 54:50–61, 2019.
- [5] Naif Radi Aljohani, Ayman Fayoumi, and Saeed-Ul Hassan. Bot prediction on social networks of twitter in altmetrics using deep graph convolutional networks. Soft Computing, pages 1–12, 2020.
- [6] Monika Arora and Vineet Kansal. Character level embedding with deep convolutional neural network for text normalization of unstructured data for twitter sentiment analysis. Social Network Analysis and Mining, 9(1):12, 2019.

- [7] Alessandro Balestrucci, Rocco De Nicola, Omar Inverso, and Catia Trubiani. Identification of credulous users on twitter. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pages 2096–2103, 2019.
- [8] Amlaan Bhoi and Sandeep Joshi. Various approaches to aspect-based sentiment analysis. arXiv preprint arXiv:1805.01984, 2018.
- [9] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Transactions on Dependable and Secure Computing, 9(6):811–824, 2012.
- [10] Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. Recurrent batch normalization. arXiv preprint arXiv:1603.09025, 2016.

CAMPUS PLACEMENTS PREDICTION & ANALYSIS USING MACHINE LEARNING

Hema Latha Guntipalli (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

K. R. Rajeswari, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT Placement of students is one of the most important objectives of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with. That is why all the institutions, arduously, strive to strengthen their placement department so as to improve their institution on a whole. Any assistance in this particular area will have a positive impact on an institution's ability to place its students. This will always be helpful to both the students, as well as the institution. In this study, the objective is to analyse previous year's student's data and use it to predict the placement chance of the current students. This model is proposed with an algorithm to predict the same. Data pertaining to the study were collected from the same institution for which the placement prediction is done and also suitable data pre-processing methods were applied. This proposed model is also compared with other traditional classification algorithms such as Decision tree and Random forest with respect to accuracy, precision and recall. From the results obtained it is found that the proposed algorithm performs significantly better in comparison with the other algorithms mentioned.

1. INTRODUCTION

Placements are considered to be very important for each and every college. The basic success of the college is measured by the campus placement of the students. Every student takes admission to the colleges by seeing the percentage of placements in the college. Hence, in this regard the approach is about the prediction and analyses for the placement necessity in the colleges that helps to build the colleges as well as students to improve their placements [1].

In Placement Prediction system predicts the probability of a undergrad students getting placed in a company by applying classification algorithms such as Decision tree and Random forest. The main objective of this model is to predict whether the student he/she gets placed or not in campus

recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous years data of the students.

From the above mentioned machine learning models Supervised learning is used in this paper.

2. EXISTING SYSTEM

Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar used Logistic regression technique on their college placement dataset which got 83.33% of accuracy [2]. Jai Ruby, Dr. K. David used ID3, J48, REP Tree, NB Tree, MLP, Decision Table Classification techniques on the placement dataset collected from their college. The results had shown that ID3 predicted well among them with an accuracy of 82.1% [3]. Ankita A Nichat, Dr. Anjali B Raut used C4.5 classification technique on the placement dataset which was collected from their college which got 80% of accuracy [4].

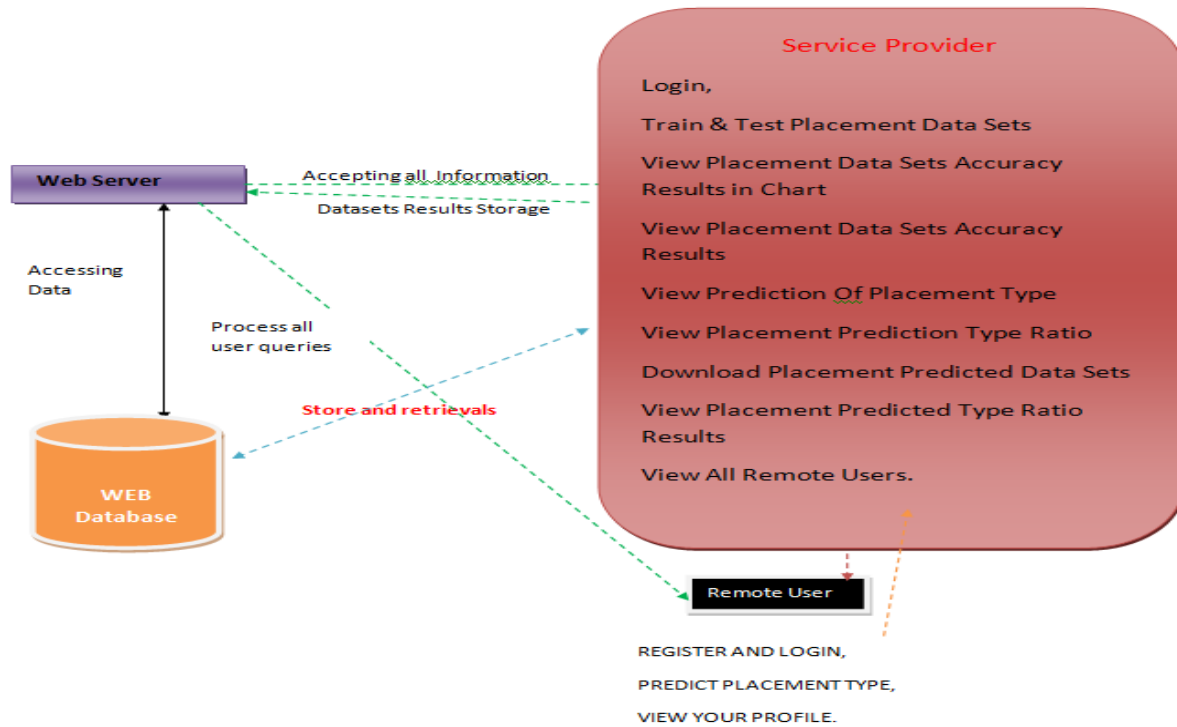
NaiveBayes, OneR classification techniques on the data gathered from their high school. The results had shown that J48 and Simple Cart predicted well among them with an accuracy of 79.61% [5]. Ajay Kumar Pal and Saurabh Pal collected the data for the study and analysis of the student's educational performance basically for training and placement. The authors used different classification algorithm and used WEKA data mining tool [6]. They concluded that naive Bayes classification model is the better algorithm based on the placement data with found accuracy of 86.15% and overall time taken to build the model is at 0 sec. As compared with others Naive Bayes classifier had lowest average error i.e. 0.28.

Ravi Tiwari and Awadhesh Kumar Sharma built the prediction model to improve the placement of the students [7]. They used WEKA as the data mining tool to build the model using random tree algorithm. They also used ID3, Bayes Net, RBF network, J48, algorithms on the student data set. They resolved that the RT (Random Tree) algorithm is more accurate with 73% for the classification/prediction of the model. The accuracy using ID3 and J48 is 71%. Bayes Net is 70%

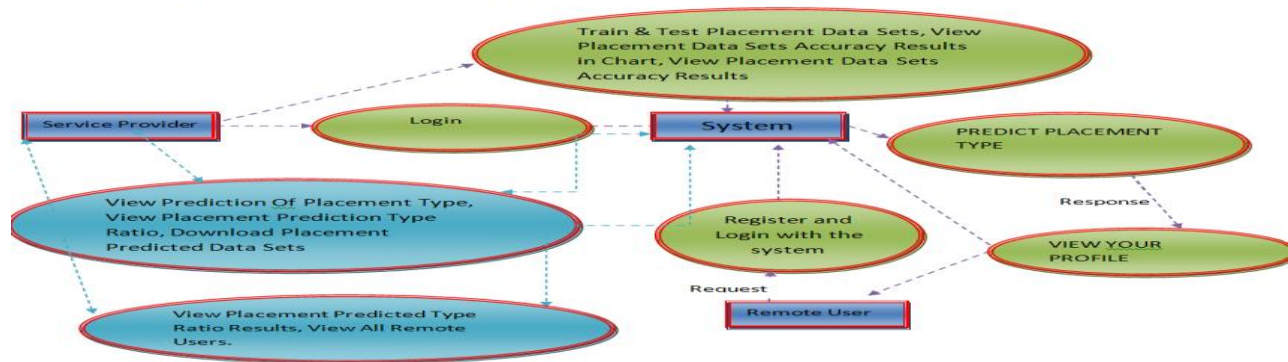
Proposed System

In Placement Prediction system predicts the probability of a undergrad students getting placed in a company by applying classification algorithms such as Decision tree and Random forest. The main objective of this model is to predict whether the student he/she gets placed or not in campus recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous years data of the students.

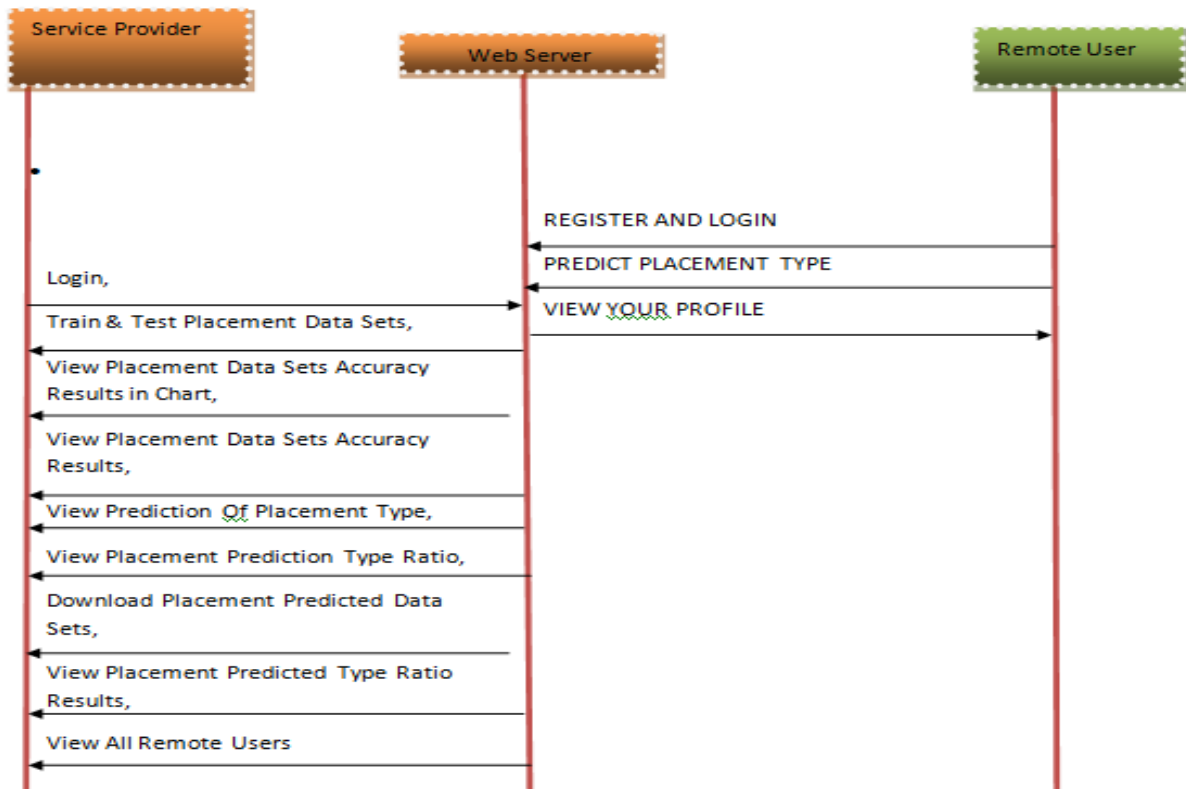
Architecture Diagram



Data Flow Diagram :



➤ Sequence Diagram



3. SYSTEM STUDY

3.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5. CONCLUSION

The campus placement activity is incredibly a lot of vital as institution point of view as well as student point of view. In this regard to improve the student's performance, a work has been analyzed and predicted using the classification algorithms Decision Tree and the Random forest

algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for Decision tree is 84% and for the Random Forest is 86%. Hence, from the above said analysis and prediction it's better if the Random Forest algorithm is used to predict the placement results.

6. REFERENCES

- [1]. Mangasuli Sheetal B, Prof. Savita Bakare “Prediction of Campus Placement Using Data Mining Algorithm- Fuzzy logic and K nearest neighbour” International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016 .
- [2]. Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar “PPS-Placement prediction system using logistic regression” IEEE international conference on MOOC,innovation and Technology in Education(MITE), December 2014.
- [3]. Jai Ruby, Dr. K. David “Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study” International Journal for Research in Applied Science & Engineering Technology (IJRASET) Vol. 2,Issue 11,November 2014.
- [4]. Ankita A Nichat, Dr.Anjali B Raut “Predicting and Analysis of Student Performance Using Decision Tree Technique” International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2017.
- [5]. Oktariani Nurul Pratiwi “Predicting Student Placement Class using Data Mining” IEEE International Conference 2013.
- [6]. Ajay Kumar Pal and Saurabh Pal, “Classification Model of Prediction for Placement of Students”, I. J. Modern Education and Computer Science, 2013, 11, 49-56
- [7]. Ravi Tiwari and Awadhesh Kumar Sharma, “A Data Mining Model to Improve Placement”, International Journal of Computer Applications (0975 – 8887) Volume 120 – No.12, June 2015
- [8]. Ms.sonal patil, Mr.Mayur Agrawal, Ms.Vijaya R. Baviskar “Efficient Processing of Decision Tree using ID3 & improved C4.5 Algorithm”, International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1956-1961

HEART DISEASE PREDICTION USING BIO INSPIRED ALGORITHMS

Himabindu Manchiraju (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

Heart related diseases or cardiovascular diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. So, there is a need of reliable, accurate and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. Many researchers, in recent times, have been using several machine learning techniques to help the health care industry and the professionals in the diagnosis of heart related diseases. This paper presents a survey of various models based on such algorithms and techniques and analyze their performance. Models based on supervised learning algorithms such as Support Vector Machines (SVM), K-Nearest Neighbour (KNN), NaïveBayes, Decision Trees (DT), Random Forest (RF) and ensemble models are found very popular among the researchers.

1. INTRODUCTION

According to a report by McKinsey [1], 50% of Americans have one or more chronic diseases, and 80% of American medical care fee is spent on chronic disease treatment. With the improvement of living standards, the incidence of chronic disease is increasing. The United States has spent an average of 2.7 trillion USD annually on chronic disease treatment. This amount comprises 18% of the entire annual GDP of the United States. The healthcare problem of chronic diseases is also very important in many other countries. In China, chronic diseases are the main cause of death, according to a Chinese report on nutrition and chronic diseases in 2015, 86.6% of deaths are caused by chronic diseases. Therefore, it is essential to perform risk

assessments for chronic dis- eases. With the growth in medical data [2], collecting elec- tronic health records (EHR) is increasingly convenient [3]. Besides, [4] rst presented a bio-inspired high-performance heterogeneous vehicular telematics paradigm, such that the collection of mobile users' health-related real-time big data can be achieved with the deployment of advanced hetero- geneous vehicular networks. Chen et al. proposed a healthcare system using smart clothing for sustainable health monitoring. Qiu et al. [8] had thoroughly studied the het- erogeneous systems and achieved the best results for cost minimization on tree and simple path cases for heteroge- neous systems. Patients' statistical information, test results and disease history are recorded in the EHR, enabling us to identify potential data-centric



solutions to reduce the costs of medical case studies. Wang et al. [9] proposed an efficient low estimating algorithm for the telehealth cloud system and designed a data coherence protocol for the PHR(Personal Health Record)-based distributed system. Bates et al. [10] proposed six applications of big data in the field of health-care. Qiu et al. [11] proposed an optimal big data sharing algorithm to handle the complicated data set in telehealth with cloud techniques. One of the applications is to identify high-risk patients which can be utilized to reduce medical cost since high-risk patients often require expensive healthcare. Moreover, in their paper proposing health-care cyber-physical system [12], it innovatively brought forward the concept of prediction-based healthcare applications, including health risk assessment. Prediction using traditional disease risk models usually involves a machine learning algorithm (e.g., logistic regression and regression analysis, etc.), and especially a supervised learning algorithm by the use of training data with labels to train the model [13], [14]. In the test set, patients can be classified into groups of either high-risk or low-risk. These models are valuable in clinical situations and are widely studied [15], [16]. However, these schemes have the following characteristics and defects. The data set is typically small, for patients and diseases with specific conditions [17], the characteristics are selected through experience. However, these pre-selected characteristics may not satisfy the changes in the disease and its influencing factors.

With the development of big data analytics technology, more attention has been paid to disease prediction from the perspective of big data analysis, various researches have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of risk classification [18], [19], rather than the previously selected characteristics. However, those existing work mostly considered structured data. For unstructured data, for example, using convolutional neural network (CNN) to extract text characteristics automatically has already attracted wide attention and also achieved very good results [20], [21]. However, to the best of our knowledge, none of previous work handle Chinese medical text data by CNN. Furthermore, there is a large difference between diseases in different regions, primarily because of the diverse climate and living habits in the region. Thus, risk classification based on big data analysis, the following challenges remain: How should the missing data be addressed? How should the main chronic diseases in a certain region and the main characteristics of the disease in the region be determined? How can big data analysis technology be used to analyze the disease and create a better model?

To solve these problems, we combine the structured and unstructured data in healthcare field to assess the risk of disease. First, we used latent factor model to reconstruct the missing data from the medical records collected from a hospital in central China. Second, by using statistical

knowledge, we could determine the major chronic diseases in the region. Third, to handle structured data, we consult with hospital experts to extract useful features. For unstructured text data, we select the features automatically using CNN algorithm. Finally, we propose a novel CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm for structured and unstructured data. The disease risk model is obtained by the combination of structured and unstructured features. Through the experiment, we draw a conclusion that the performance of CNN-MDRP is better than other existing methods.

2. SYSTEM DESIGN

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

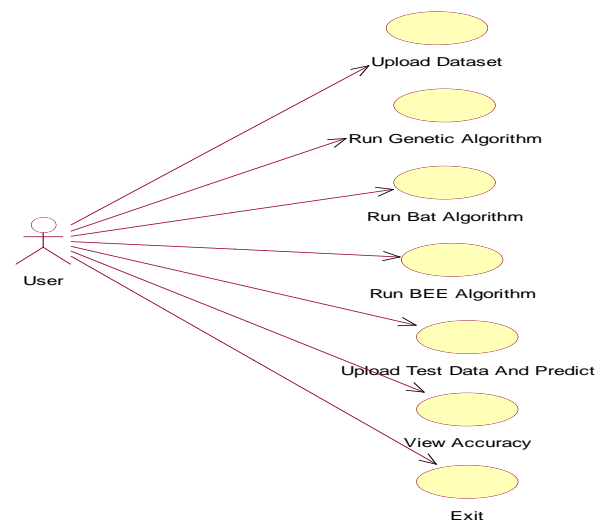
The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

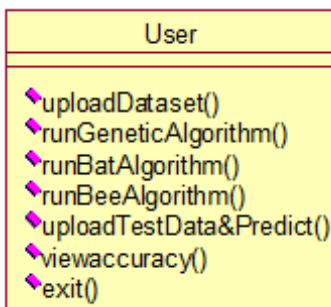
USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



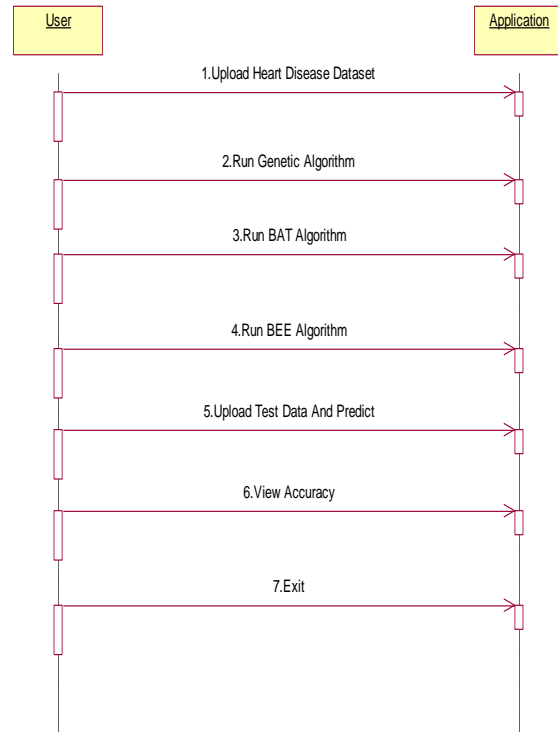
CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information

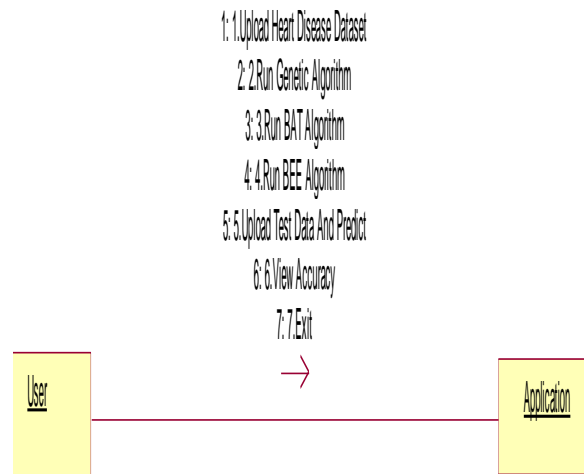


SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



Collaboration Diagram:



3. TEST RESULT

In Due Course, latest technology advancements will be taken into consideration. As part of technical build-up many components of the networking system will be generic in nature so that future projects can either use or



interact with this. The future holds a lot to offer to the development and refinement of this project.

In this project student want to detect heart disease from dataset using Bio Inspired 4 features optimizing algorithms such as Genetic Algorithm, Bat, Bee and ACO. Here ACO algorithm is design in python to solve Travelling Salesman Problem to find shortest path and it cannot be implemented with heart disease dataset, so I am implementing 3 algorithms called Genetic, Bat and Bee.

Bio inspired algorithms design to optimized features used in dataset for training classification algorithms to increase prediction accuracy, sometime some datasets may have irrelevant values inside dataset and those irrelevant attributes or values may degrade classification accuracy so using optimize algorithms we can reduce features (attribute values) from dataset. This optimize algorithms will be applied on dataset to check whether all values are related to dataset or not, if any attribute found unrelated then it will removed from dataset.

To implement this algorithms I am using Heart disease dataset which contains 14 attributes and 4 class labels where 0 refers to No heart Disease and 1 refers to stage1 disease and 2 and 3 refers stage 3 and 4 disease.

Below are some values from dataset to train algorithms

age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,class
63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0,0

67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0,2

67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0,1

37.0,1.0,3.0,130.0,250.0,0.0,0.0,187.0,0.0,3.5,3.0,0.0,3.0,0

First records contains dataset column names and remaining records are the values of dataset. In last column we have class values as 0, 2, 1 and 3 as disease stage.

Test dataset also contains record values but it will not have class labels and application will apply that test values on train dataset to predict it class labels. Some values from test dataset.

age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal

63.0,1.0,1.0,145.0,233.0,1.0,2.0,150.0,0.0,2.3,3.0,0.0,6.0

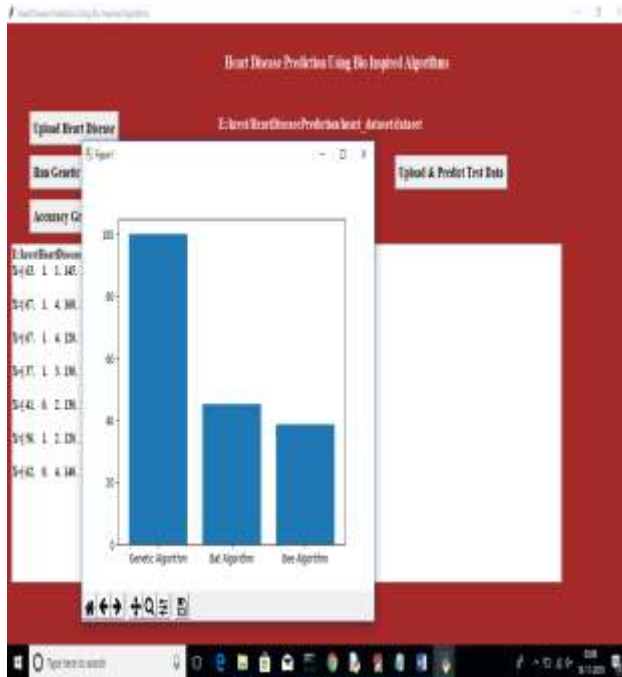
67.0,1.0,4.0,160.0,286.0,0.0,2.0,108.0,1.0,1.5,2.0,3.0,3.0

67.0,1.0,4.0,120.0,229.0,0.0,2.0,129.0,1.0,2.6,2.0,2.0,7.0

In above test dataset we can see there is no class name and application will predict it.

All this files are available inside 'heart_dataset' folder.

4. OUTPUT RESULT



In above graph x-axis represents Algorithm Name and y-axis represents accuracy of those algorithms

5. CONCLUSION

In this paper, we propose a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm.

6. REFERENCES

[1] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The 'Big Data' Revolution in Healthcare:

Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.

[2] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[3] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: Towards better research applications and clinical care," *Nature Rev. Genet.*, vol. 13, no. 6, pp. 395–405, 2012.

[4] D. Tian, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dynamic and self-adaptive network selection method for multimode communications in heterogeneous vehicular telematics," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3033–3049, Dec. 2015.

[5] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system," *IEEE Commun.*, vol. 55, no. 1, pp. 54–61, Jan. 2017.

[6] M. Chen, Y. Ma, J. Song, C. Lai, and B. Hu, "Smart clothing: Connecting human with clouds and big data for sustainable health monitoring," *ACM/Springer Mobile Netw. Appl.*, vol. 21, no. 5, pp. 825–845, 2016.

[7] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2017, doi: 10.1109/ACCESS.2016.2641480.

[8] M. Qiu and E. H.-M. Sha, "Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems," *ACM Trans. Design Autom. Electron. Syst.*, vol. 14, no. 2, p. 25, 2009.

[9] J. Wang, M. Qiu, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform," *J. Syst. Archit.*, vol. 72, pp. 69–79, Jan. 2017.

[10] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage



- high-risk and high-cost patients,” *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.
- [11] L. Qiu, K. Gai, and M. Qiu, “Optimal big data sharing approach for telehealth in cloud computing,” in *Proc. IEEE Int. Conf. Smart Cloud (SmartCloud)*, Nov. 2016, pp. 184–189.
- [12] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, “HealthCPS: Healthcare cyber-physical system assisted by cloud and big data,” *IEEE Syst. J.*, vol. 11, no. 1, pp. 88–95, Mar. 2017.
- [13] K. Lin, J. Luo, L. Hu, M. S. Hossain, and A. Ghoneim, “Localization based on social big data analysis in the vehicular networks,” *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2016.2641467.
- [14] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced fingerprinting and trajectory prediction for iot localization in smart buildings,” *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 3, pp. 1294–1307, Jul. 2016.
- [15] D. Oliver, F. Daly, F. C. Martin, and M. E. McMurdo, “Risk factors and risk assessment tools for falls in hospital in-patients: A systematic review,” *Age Ageing*, vol. 33, no. 2, pp. 122–130, 2004.
- [16] S. Marcoon, A. M. Chang, B. Lee, R. Salhi, and J. E. Hollander, “Heart score to further risk stratify patients with low TIMI scores,” *Critical Pathways Cardiol.*, vol. 12, no. 1, pp. 1–5, 2013.
- [17] S. Bandyopadhyay et al., “Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data,” *Data Mining Knowl. Discovery*, vol. 29, no. 4, pp. 1033–1069, 2015.
- [18] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, “A relative similarity based method for interactive patient risk prediction,” *Data Mining* 1691–1700.
- Knowl. Discovery*, vol. 29, no. 4, pp. 1070–1093, 2015.
- [19] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, “Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration,” *J. Biomed. Inform.*, vol. 53, pp. 220–228, Feb. 2015.
- [20] J. Wan et al., “A manufacturing big data solution for active preventive maintenance,” *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2017.2670505.
- [21] W. Yin and H. Schutze, “Convolutional neural network for paraphrase identification,” in *Proc. HLT-NAACL*, 2015, pp. 901–911.
- [22] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, “Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care,” in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 855–864.
- [23] S. Zhai, K.-H. Chang, R. Zhang, and Z. M. Zhang, “Deepintent: Learning attentions for online advertising with recurrent neural networks,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1295–1304.
- [24] K. Hwang and M. Chen, *Big Data Analytics for Cloud/IoT and Cognitive Computing*. Hoboken, NJ, USA: Wiley, 2017.
- [25] H. Chen, R. H. Chiang, and V. C. Storey, “Business intelligence and analytics: From big data to big impact,” *MIS Quart.*, vol. 36, no. 4, pp. 1165–1188, 2012.
- [26] S. Basu Roy et al., “Dynamic hierarchical classification for patient riskof-readmission,” in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp.

EARLY DETECTION OF CANCER USING AI

Irla Vineetha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract: Cancer is an aggressive disease with a low median survival rate. Ironically, the treatment process is long and very costly due to its high recurrence and mortality rates. Accurate early diagnosis and prognosis prediction of cancer are essential to enhance the patient's survival rate. Developments in statistics and computer engineering over the years have encouraged many scientists to apply computational methods such as multivariate statistical analysis to analyze the prognosis of the disease, and the accuracy of such analyses is significantly higher than that of empirical predictions. Furthermore, as artificial intelligence (AI), especially machine learning and deep learning, has found popular applications in clinical cancer research in recent years, cancer prediction performance has reached new heights. This article reviews the literature on the application of AI to cancer diagnosis and prognosis, and summarizes its advantages. We explore how AI assists cancer diagnosis and prognosis, specifically with regard to its unprecedented accuracy, which is even higher than that of general statistical applications in oncology. We also demonstrate ways in which these methods are advancing the field. Finally, opportunities and challenges in the clinical implementation of AI are discussed. Hence, this article provides a new perspective on how AI technology can help improve cancer diagnosis and prognosis, and continue improving human health in the future. **Keywords:** Cancer diagnosis; Prognosis prediction; Deep learning; Machine learning; Deep neural network

1 INTRODUCTION

In this project we are implementing Artificial Intelligence algorithm called as Neural Networks with various optimizer techniques such as ADAM, SGD and Gradient Descent Mini Batch to predict cancer disease. To train AI algorithms we have used images given by you and this images contains 3 different types of cancer or stages and below screen showing such cancer details. To implement this project we have designed following modules Upload Histopathological Images Dataset: using this module we will upload dataset to application

Preprocess Dataset: using this module we will read all images and then resize all images to equal size and then normalize pixel values. After processing we will split dataset into train and test

Train AI with ADAM: using this module we will feed Training Data to AI algorithm with optimizer as ADAM. After training we will apply test data in trained model to calculate prediction accuracy. Train AI with SGD: using this module we will feed Training Data to AI algorithm with optimizer as SGD. After training we will apply test data in trained model to calculate prediction accuracy

Train AI with MiniBatch: using this module we will feed Training Data to AI algorithm with optimizer as MiniBatch. After training we will apply test data in trained model to calculate prediction accuracy

Comparison Table: using this module we will plot all algorithm accuracy and show performance in tabular format

2. INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

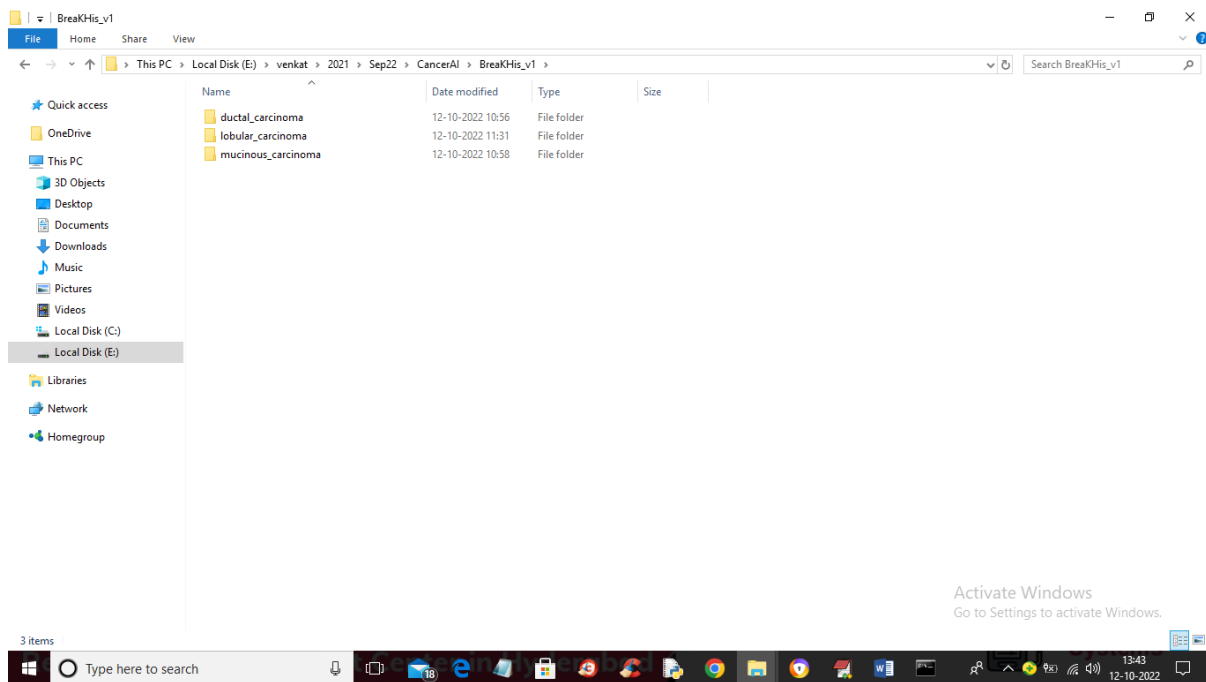
- Convey information about past activities, current status or projections of the

- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

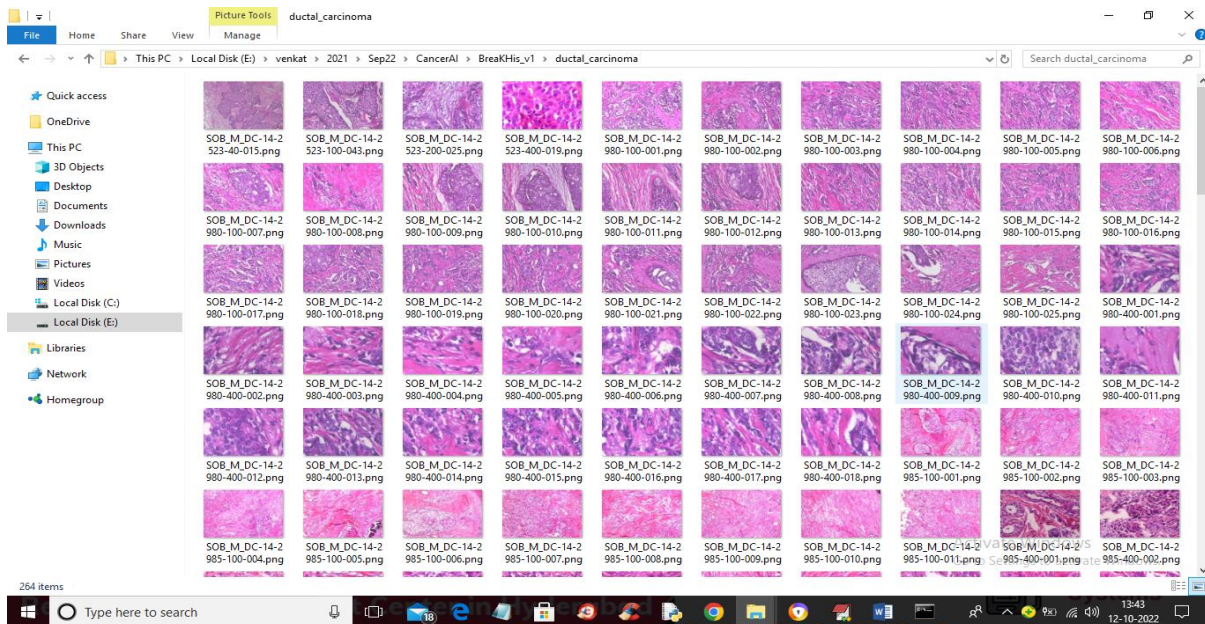
3. OUTPUT SCREENS

Early Detection of Cancer using AI

In this project we are implementing Artificial Intelligence algorithm called as Neural Networks with various optimizer techniques such as ADAM, SGD and Gradient Descent Mini Batch to predict cancer disease. To train AI algorithms we have used images given by you and this images contains 3 different types of cancer or stages and below screen showing such cancer details



In above screen you can see dataset contains 3 different types of cancers and just go inside any folder to view those images



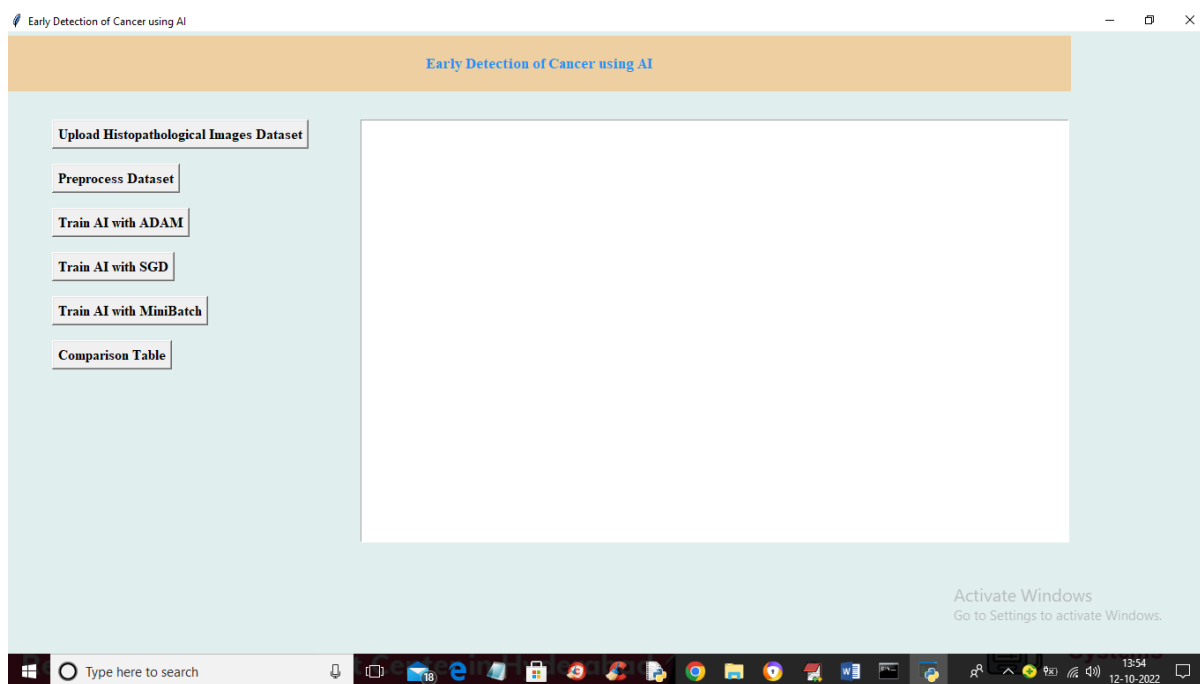
So by using above images we are training AI with 3 different optimizers.

To implement this project we have designed following modules

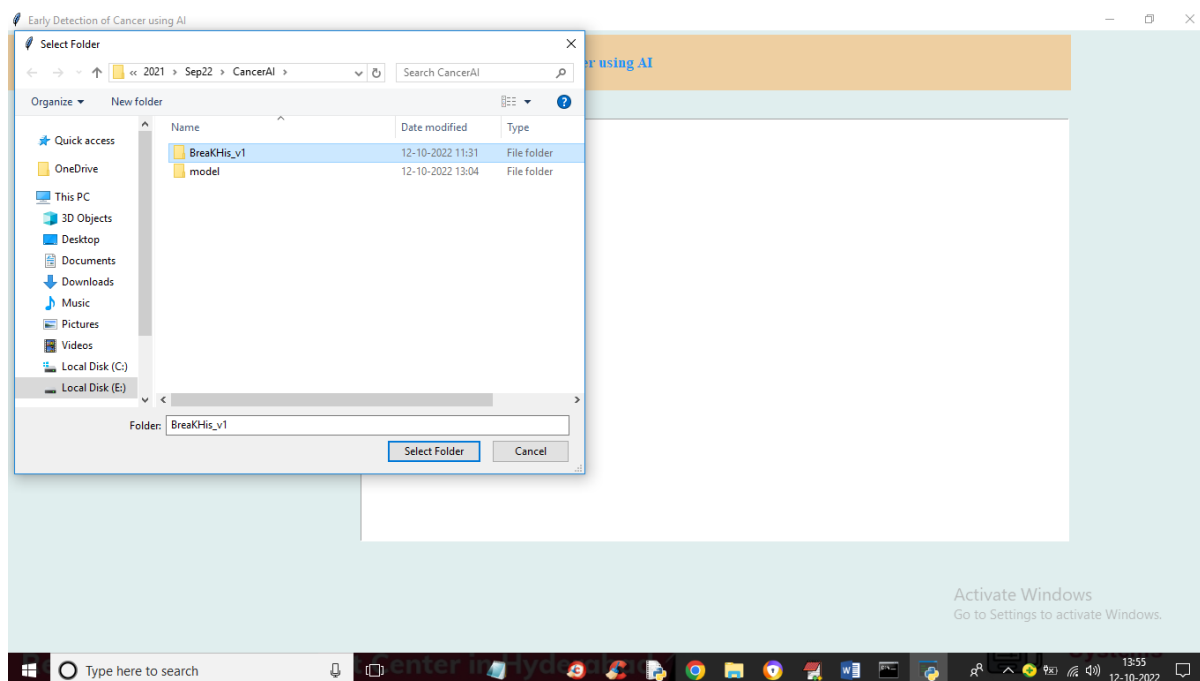
- 1) Upload Histopathological Images Dataset: using this module we will upload dataset to application
- 2) Preprocess Dataset: using this module we will read all images and then resize all images to equal size and then normalize pixel values. After processing we will split dataset into train and test
- 3) Train AI with ADAM: using this module we will feed Training Data to AI algorithm with optimizer as ADAM. After training we will apply test data in trained model to calculate prediction accuracy
- 4) Train AI with SGD: using this module we will feed Training Data to AI algorithm with optimizer as SGD. After training we will apply test data in trained model to calculate prediction accuracy
- 5) Train AI with MiniBatch: using this module we will feed Training Data to AI algorithm with optimizer as MiniBatch. After training we will apply test data in trained model to calculate prediction accuracy
- 6) Comparison Table: using this module we will plot all algorithm accuracy and show performance in tabular format

SCREEN SHOTS

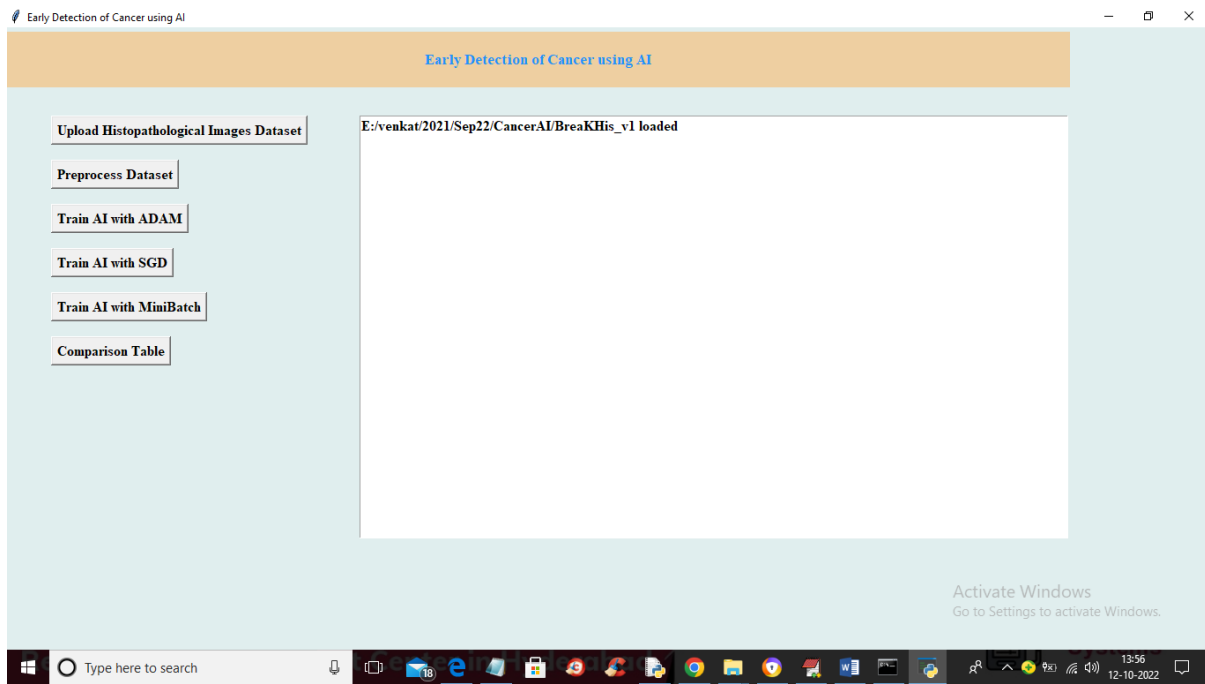
To run project double click on 'run.bat' file to get below screen



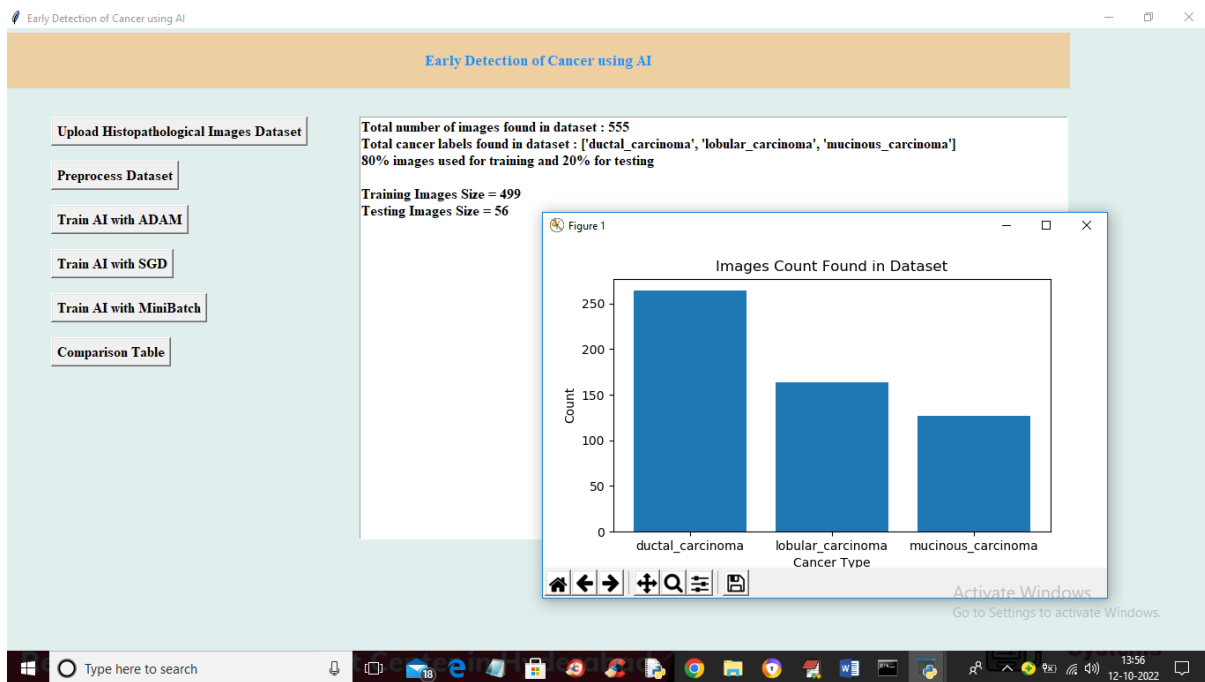
In above screen click on 'Upload Histopathological Images Dataset' button o upload dataset and get below output



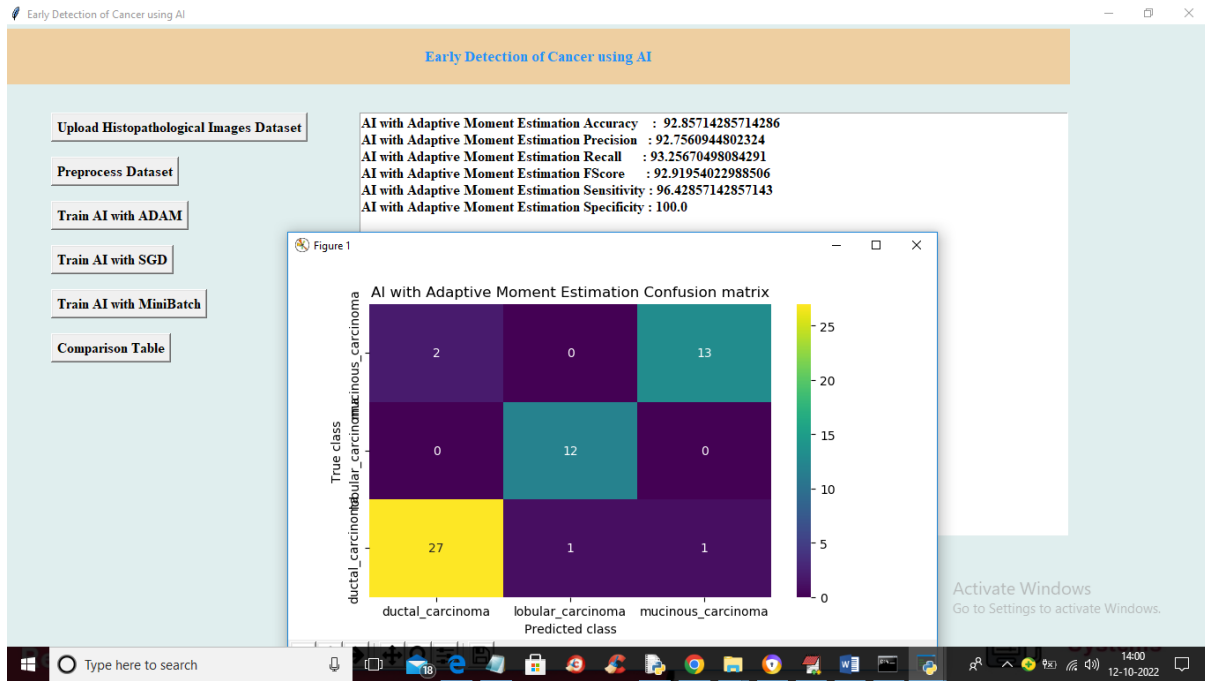
In above screen selecting and uploading entire 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below output



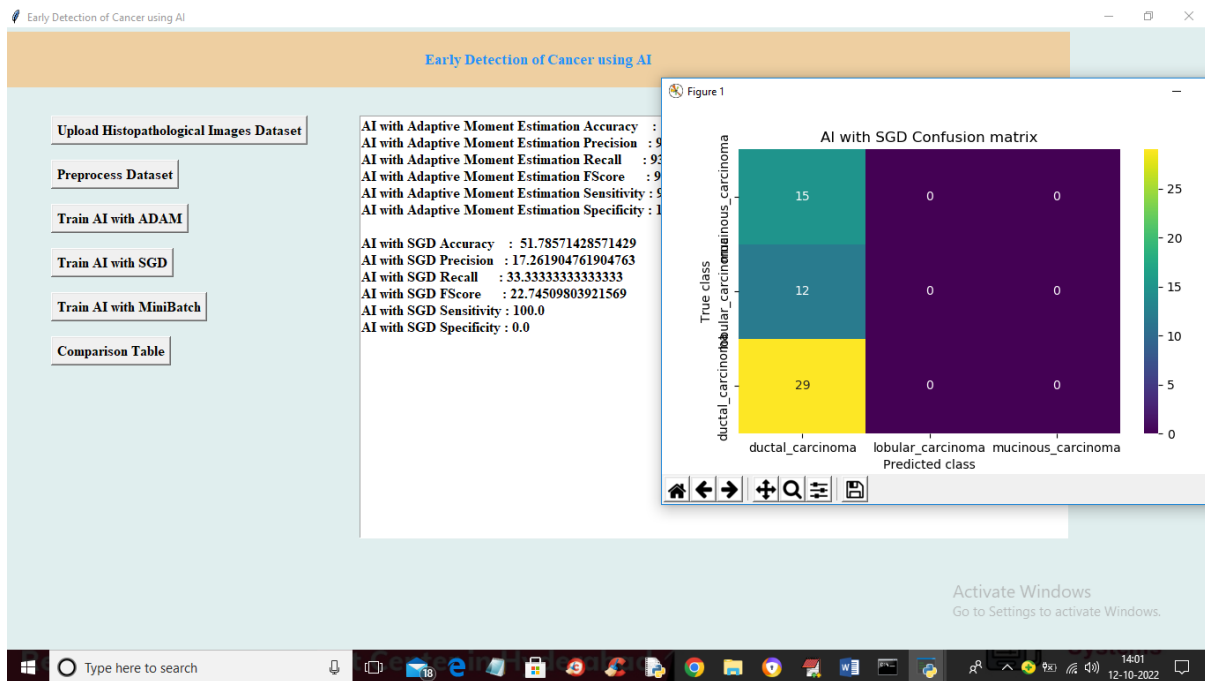
In above screen dataset loaded and now click on ‘Preprocess Dataset’ button to read and process images and get below output



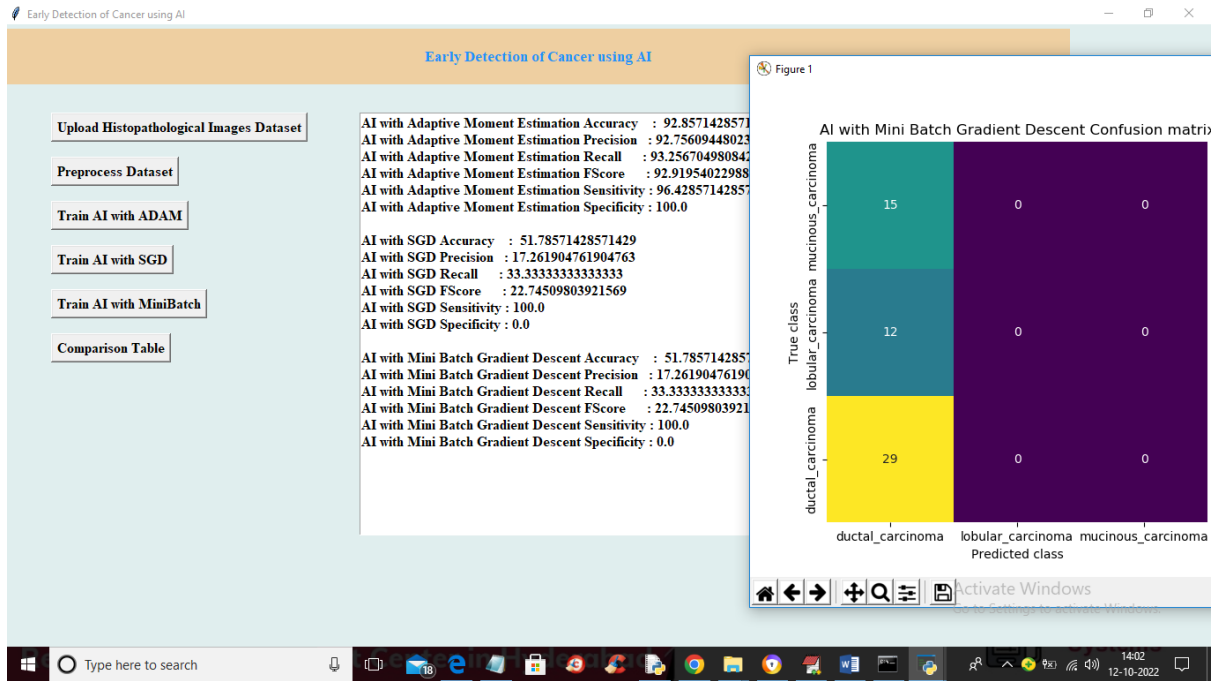
In above screen we can see dataset contains 555 images and then showing train and test data size and in graph x-axis represents ‘Cancer’ type and y-axis represents number of images of that cancer type and now close above image and then click on ‘Train AI with ADAM’ button to train AI and get below output



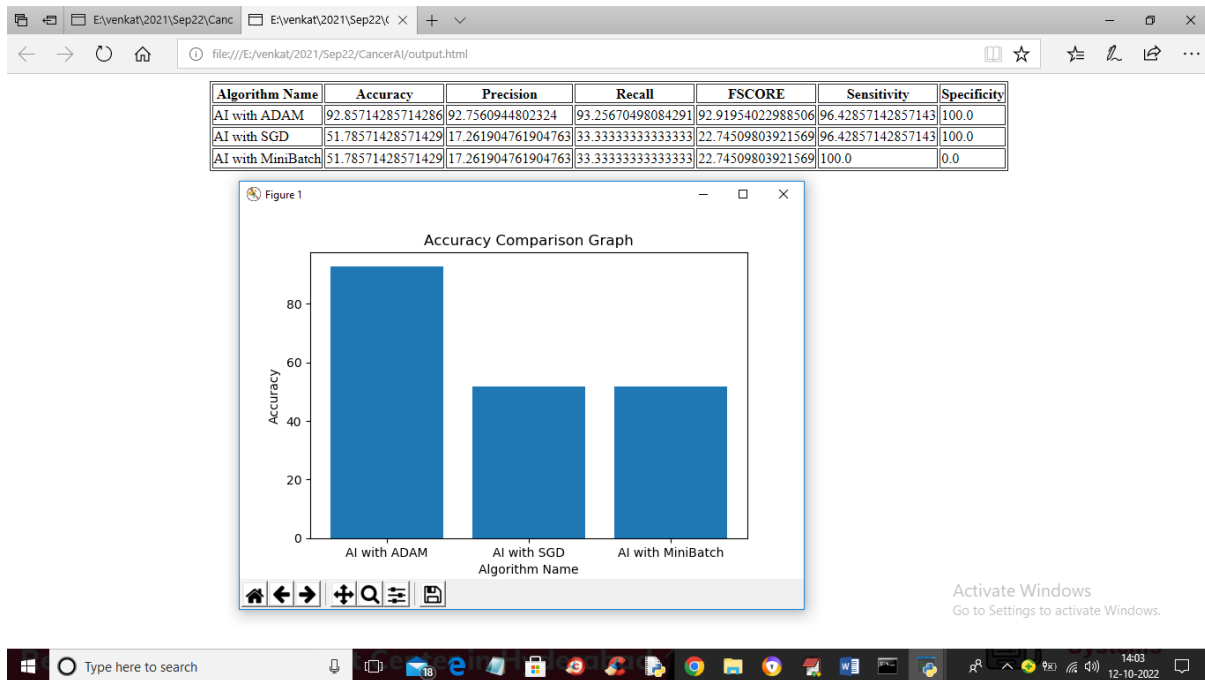
In above screen AI with ADAM got 92% accuracy and in confusion matrix graph x-axis represents Predicted Labels and y-axis represents TRUE labels and different colour boxes represents CORRECT Prediction count and same blue colour boxes represents incorrect prediction count. Now close above and then click on 'Train with SGD' button to train with SGD and get below output



In above screen AI with SGD we got 51% accuracy and now close above graph and then click on 'Train with MiniBatch' button to train AI and get below output



In above screen with Mini Batch also we got 51% accuracy and now close above graph and then click on ‘Comparison Table’ button to get below output



In above screen we can see all algorithm performance in tabular and graphical format and in all algorithms ‘AI with ADAM’ got high performance or accuracy

4. CONCLUSION

AI is slowly pervading all aspects of our lifestyle, especially medicine. The review presented in this paper shows that researchers are rapidly acquiring a much deeper understanding of the challenges and opportunities presented by AI as an intelligent information science in the field of cancer diagnosis and care. The potential of AI for various types of cancer prognosis and diagnosis is reported in this paper. But, the limit of review is that we did not include the genomics and radiomics data applied by AI to acquire clinical precise medicine. We expect that AI-based clinical cancer research will result in a paradigm shift in cancer treatment, thereby resulting in dramatic improvement in patient survival due to enhanced prediction rates. Thus, it is logical to expect that the challenges of cancer prognosis and diagnosis will be solved by advances in AI in the foreseeable future.

5. REFERENCE

- [1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA: A Cancer Journal for Clinicians, 69 (2019) 7-34.
- [2] Simmons CPL, McMillan DC, McWilliams K, Sande TA, Fearon KC, Tuck S, Fallon MT, Laird BJ: Prognostic Tools in Patients With Advanced Cancer: A Systematic Review. J Pain Symptom Manage 2017, 53(5):962-970 e910
- [3] S. Huang, Dang, Y., Li, F., Wei, W., Ma, Y., Qiao, S., & Wang, Q, Biological intensity-modulated radiotherapy plus neoadjuvant chemotherapy for multiple peritoneal metastases of ovarian cancer: A case report, Oncology Letters, (2015) 1239-1243.
- [4] S. Huang, Q. Zhao, Nanomedicine-combined immunotherapy for cancer, Current medicinal chemistry, (2019).
- [5] S. Huang, C.I. Fong, M. Xu, B.-n. Han, Z. Yuan, Q. Zhao, Nano-loaded natural killer cells as carriers of indocyanine green for synergetic cancer immunotherapy and phototherapy, Journal of Innovative Optical Health Sciences, 12 (2019) 1941002.
- [6] Z. Obermeyer, E.J. Emanuel, Predicting the Future - Big Data, Machine Learning, and Clinical Medicine, N Engl J Med, 375 (2016) 1216-1219.
- [7] K.P.E. Gillies R J, Hricak H., Radiomics Images Are More than Pictures, They Are Data, Radiology, 278 (2015) 563-577.

- [8] A. Allahyar, J. Ubels, J. de Ridder, A data-driven interactome of synergistic genes improves networkbased cancer outcome prediction, *PLoS Comput Biol*, 15 (2019) e1006657.
- [9] M.J. Mitchell, R.K. Jain, R. Langer, Engineering and physical sciences in oncology: challenges and opportunities, *Nat Rev Cancer*, 17 (2017) 659-675.
- [10] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H. Aerts, Artificial intelligence in radiology, *Nat Rev Cancer*, 18 (2018) 500-510.

PERSONALIZED TRAVEL PLANNING SYSTEM

Jaddu Srihari (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

Nowadays tourism transportation has become a hot topic of research, and the rapid development of Internet technology has overloaded information, which has made it impossible to provide services with different preferences for different users. Therefore, personalized tourism transportation has become the current mainstream trend. According to the different preferences of travelers for money and travel time, based on the analysis of mainstream tourism services, and combined with multi-source traffic data, this paper proposes a mathematical model for personalized travel planning. This paper proposes a two-stage spatiotemporal network solution algorithm. In the first stage, based on the set of travel attractions given by the traveler, the shortest path algorithm is used to plan an approximate optimal path that meets the traveler's preferences and to implement connection of multiple travel modes. The second stage is combined with the spatiotemporal network to achieve daily travel planning between multiple attractions. The two-stage spatiotemporal network algorithm is feasible for solving path planning problems, and can simplify route planning problems with time windows, which provides a useful reference for future personalized travel planning recommendations.

1. INTRODUCTION

1.1 Problem statement:

Choosing a tourist destination from the information that is available on the Internet and through other sources is one of the most complex tasks for tourists when planning travel, both before and during travel. Previous Travel Recommendation Systems (TRSs) have attempted to solve this problem. However, some of the technical aspects such as system accuracy and the practical aspects such as usability and satisfaction have been neglected..

1.1 Motivation:

To address this issue, it requires a full understanding of the tourists' decision-making and novel models for their information search process. This paper proposes a novel human-

centric TRS that recommends destinations to tourists in an unfamiliar city. It considers both technical and practical aspects using a real world data set we collected. The system is developed using a two-steps feature selection method to reduce number of inputs to the system and recommendations are provided by decision tree C4.5. The experimental results show that the proposed TRS can provide personalized recommendation on tourist destinations that satisfy the tourists.

1.2 Objective:

a tourist destination from the information that is available on the Internet and through other sources is one of the most complex tasks for tourists when planning travel, both before and during travel. Previous Travel Recommendation Systems (TRSs) have attempted to solve this problem. However, some of the technical aspects such as system accuracy and the practical aspects such as usability and satisfaction have been neglected. To address this issue, it requires a full understanding of the tourists' decision-making and novel models for their information search process.

1.3.1 Proposed System:

The proposed DM framework consists of four phases including data acquisition, data pre-processing, data analysis, and result interpretation. (1) For data acquisition, the designed questionnaire, which has four parts, is distributed and collected from Chiang Mai, Thailand. (2) The collected data is pre-processed using several data pre-processing techniques involving data cleaning, data transformation, and feature selection methods. (3) The third phase involves the data analysis processes using a decision tree C4.5 as classifier. The aim of the third phase is to identify suitable features and find personalized systems have not been a focus of RS research.

- To overcome from above problem author is asking to use C4.5 decision tree algorithms which take experiences of previous users and then build a model and if new user enter his requirements then decision tree will predict best location based on his given input. Decision tree don't need new users past experience data.
- To implement decision tree model, we need to have dataset and this dataset sometime will have empty or garbage values and this values will put bad

effect on decision tree model so we can remove such empty or garbage values by applying pre-process techniques.

Sometime to predict or build model no need to use all columns (attributes) values from dataset and these unnecessary attributes can be remove by apply features selection algorithms and here we are using MRMR features selection algorithms to remove unnecessary attributes to reduce execution time of building model and to increase system accuracy.

2. INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with

the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

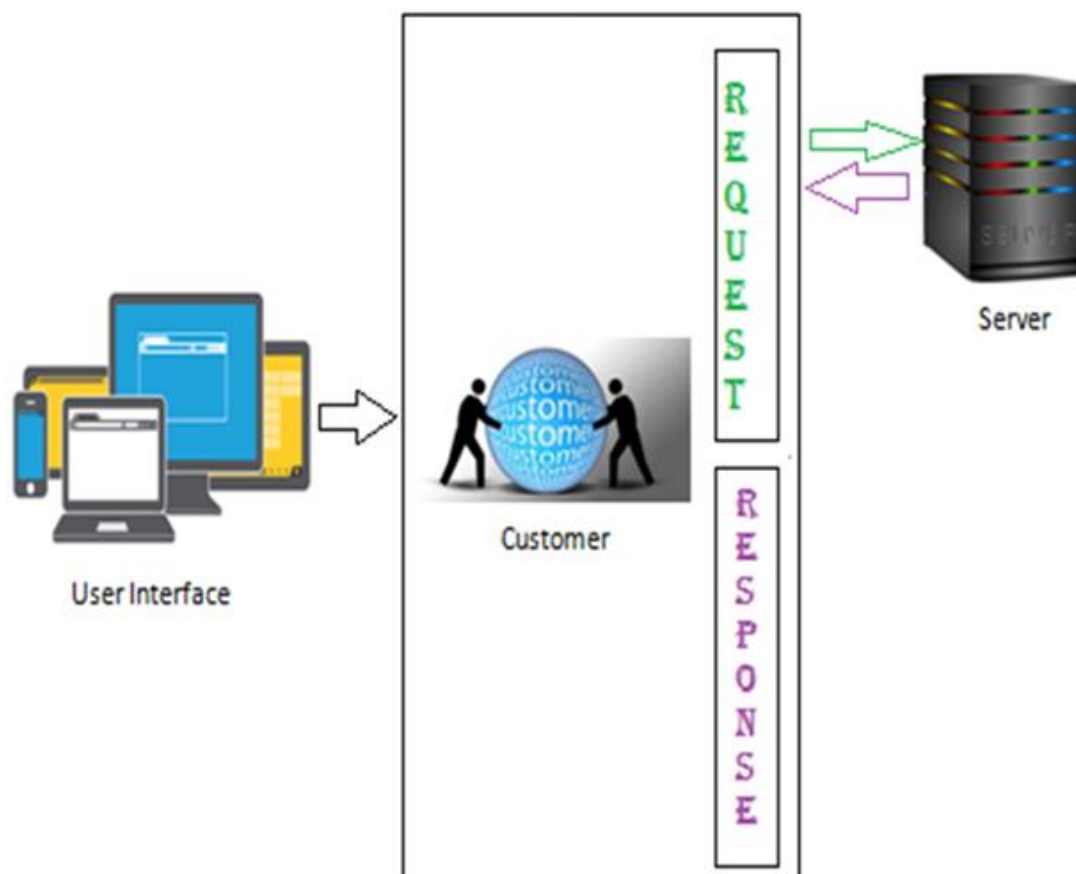
3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

3. SYSTEM DESIGN

3.1 System Architecture



3.2Moduledescription

Our application consists of three modules

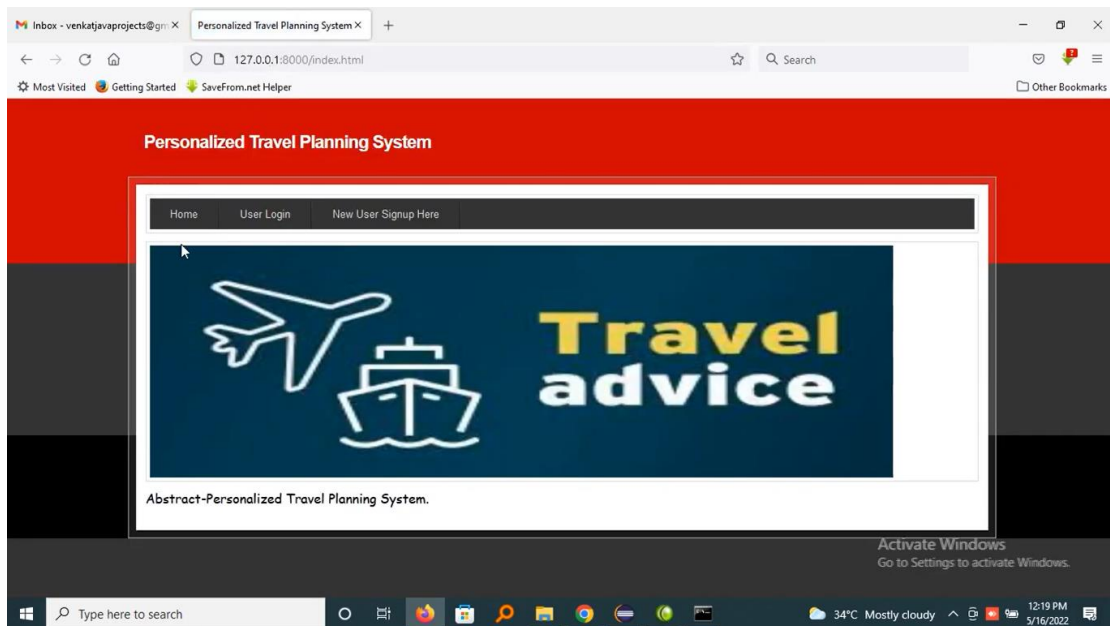
1. Customer module

Customer

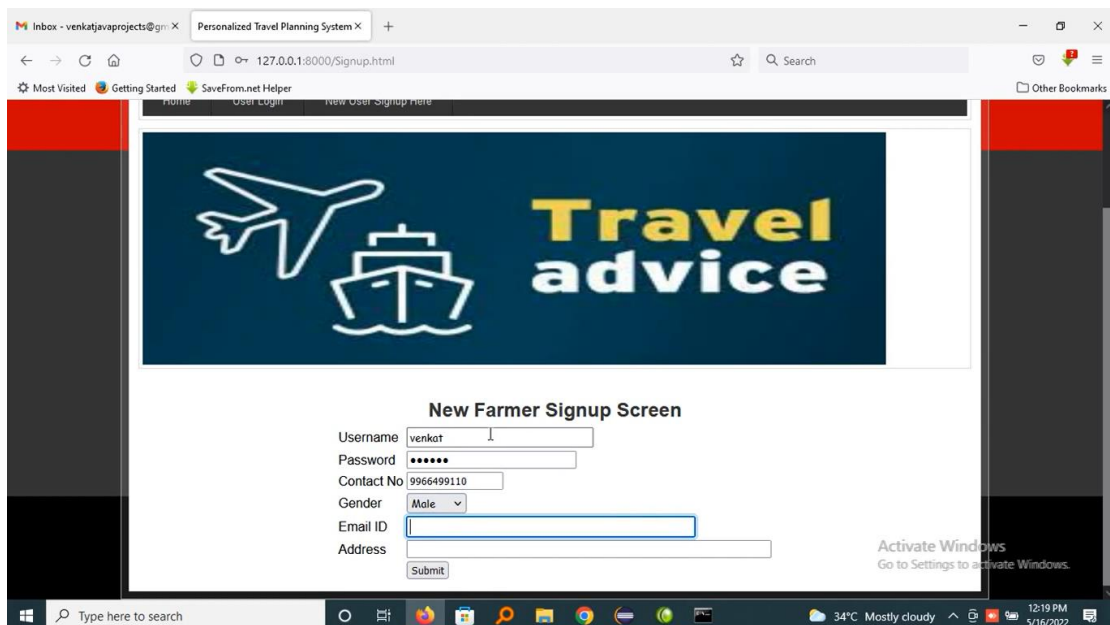
This module describes all about customers, by using this module any customer can perform operations like the upload dataset preprocess & MRMR Feature Selection Generate C4.5 Decision Tree Model Tourist Recommendation features Selection Graph.

4. RESULTS

Home Page



Rsgistration page



5. CONCLUSION & FUTURE WORK

a decision tree based tourist recommendation system has been presented in attempt of solving the current challenge of the destination TRS. The data set has been decomposed into two sub data sets using relevant tourism domain knowledge. This was done to increase classification accuracy rate and to reduce the complexity of the decision tree. The optimal decision trees

from NMIFS with the highest accuracy rate and simplicity (i.e. less number of leaf and tree size) have been constructed for destination choice. The decision rules from decision trees were extracted. It can be seen that NMIFS is the optimum method because it uses fewer number of feature than MRMR for both of the data sets. Finally, the experimental results confirm applicable of the proposed a TRS. The proposed TRS satisfies the tourists' requirements who plan to visit or during their visit the city of Chiang Mai.

6. REFERENCES

1. J.Chiverton, "Helmet Presence Classification with Motorcycle Detection And Tracking", IET Intelligent Transport Systems, Vol. 6, Issue 3, pp. 259–269, March 2012.
2. Rattapoom Waranusast, Nannaphat Bundon, Vasan Timtong and Chainarong Tangnoi, "Machine Vision techniques for Motorcycle Safety Helmet Detection", 28th International Conference on Image and Vision Computing New Zealand, pp 35-40, IVCNZ 2013.
3. Romuere Silva, Kelson Aires, Thiago Santos, Kalyf Abdala, Rodrigo Veras, André Soares, "Automatic Detection Of Motorcyclists without Helmet", 2013 XXXIX Latin America Computing Conference (CLEI).IEEE,2013.
4. Romuere Silva, "Helmet Detection on Motorcyclists Using Image Descriptors and Classifiers", 27th SIBGRAPI Conference on Graphics, Patterns and Images.IEEE, 2014.
5. Thepnimit Marayatr, Pinit Kumhom, "Motorcyclist"s Helmet Wearing Detection Using Image Processing", Advanced Materials Research Vol 931- 932,pp. 588-592,May-2014.
6. Amir Mukhtar, Tong Boon Tang, "Vision Based Motorcycle Detection using HOG features", IEEE International Conference on Signal and Image Processing Applications (ICSIPA).IEEE, 2015.



FLIGHT DELAY PREDICTION BASED ON AVIATION BIG DATA AND MACHINE LEARNING

Jakkamsetti Kiran Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

Accurate flight delay prediction is fundamental to establish the more efficient airline business. Recent studies have been focused on applying machine learning methods to predict the flight delay. Most of the previous prediction methods are conducted in a single route or airport. This paper explores a broader scope of factors which may potentially influence the flight delay, and compares several machine learning-based models in designed generalized flight delay prediction tasks. To build a dataset for the proposed scheme, automatic dependent surveillance broadcast (ADS-B) messages are received, pre-processed, and integrated with other information such as weather condition, flight schedule, and airport information. The designed prediction tasks contain different classification tasks and a regression task. Experimental results show that long short-term memory (LSTM) is capable of handling the obtained aviation sequence data, but overfitting problem occurs in our limited dataset. Compared with the previous schemes, the proposed random forest-based model can obtain higher prediction accuracy (90.2% for the binary classification) and can overcome the overfitting problem.

Index Terms:

Flight delay prediction, ADS-B, machine learning, LSTM neural network, random forest.

1. INTRODUCTION

A IR traffic load has experienced rapid growth in recent years, which brings increasing demands for air traffic surveillance system. Traditional surveillance technology such as primary surveillance radar (PSR) and secondary surveillance radar (SSR) cannot meet requirements of the future dense air traffic. Therefore, new technologies such as automatic dependent surveillance broadcast (ADS-B) have been proposed, where flights can periodically broadcast their current state information, such as international civil aviation organization (ICAO) identity number, longitude, latitude and speed [1]. Compared with the traditional radar-based schemes, the ADSB-based scheme is low cost, and the corresponding ADS-B receiver (at 1090 MHz or 978 MHz) can be easily connected to personal computers [2]. The received ADS-B message along with other collected data from the Internet can constitute a This work was supported by the Project Funded by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant TC190A3WZ-2, National Natural Science Foundation of China under Grant 61901228, Jiangsu Specially Appointed Professor Program under Grant RK002STP16001, Summit of the Six Top Talents Program of Jiangsu under Grant XYDXX-010, Program for High-Level Entrepreneurial and Innovative Talents Introduction under Grant CZ0010617002, and 1311 Talent Plan of Nanjing University of Posts and



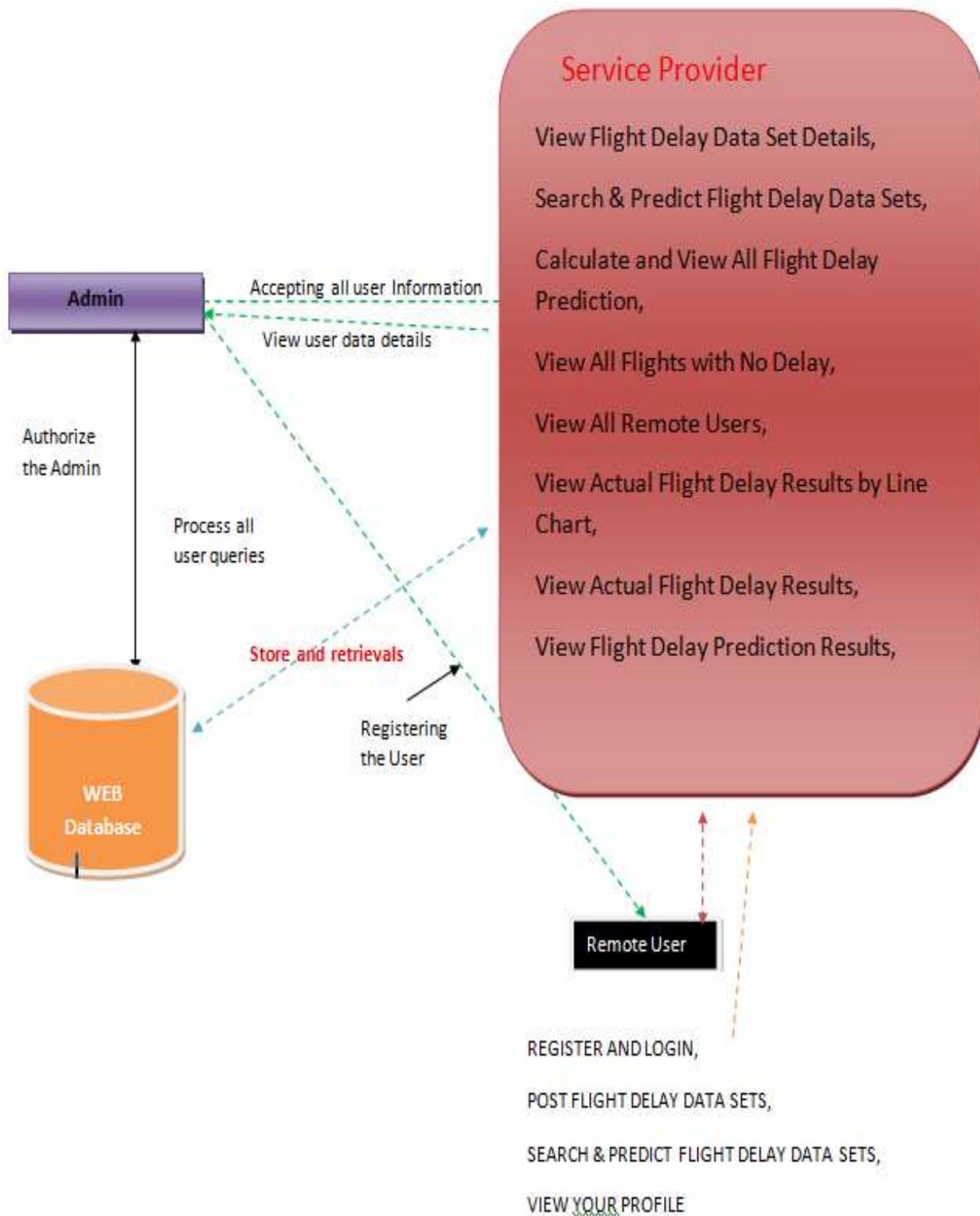
Telecommunications. (Corresponding authors: Jinlong Sun and Jie Yang) The authors are with College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China (E-mails: {guiguan, 1018010402, sunjinlong, jyang, 1218012005, 1218012004}@njupt.edu.cn). huge volumes of aviation data by which data mining can support military, agricultural, and commercial applications. In the field of civil aviation, the ADS-B can be used to increase precision of aircraft positioning and the reliability of air traffic management (ATM) system [3]. For example, malicious or fake messages can be detected with the use of multilateration (MLAT) [1], allowing open, free, and secure visibility to all the aircrafts within airspace [2]. Thus, the ADS-B provides opportunity to improve the accuracy of flight delay prediction which contains great commercial value. The flight delay is defined as a flight took off or arrived later than the scheduled time, which occurs in most airlines around the world, costing enormous economic losses for airline company, and bringing huge inconvenience for passenger. According to civil aviation administration of China (CAAC), 47.46% of the delays are caused by severe weather, and 21.14% of the delays are caused by air route problems. Due to the own problem of airline company or technical problems, air traffic control and other reasons account for 2.31% and 29.09%, respectively. Recent studies have been focused on finding a suitable way to predict probability of flight delay or delay time to better apply air traffic flow management (ATFM) [4] to reduce the delay level. Classification and regression methods are two main ways for modeling the prediction model. Among the classification models, many recent studies applied machine learning methods and obtained promising results [5]– [7]. For instance, L. Hao et al.

[8] used a regression model for the three major commercial airports in New York to predict flight delay. However, several reasons are restricting the existing methods from improving the accuracy of the flight delay prediction. The reasons are summarized as follows: the diversity of causes affecting the flight delay, the complexity of the causes, the relevancy between causes, and the insufficiency of available flight data. In [6], a public dataset named VRA [9] was used to compare the performance of several machine learning models including k-nearest neighbors (K-NN) [10], support vector machines (SVM) [11], naive Bayes classifier, and random forests for predicting flight delay, and achieved the best accuracy of 78.02% among the four schemes. However, the air route information (e.g., traffic flow and size of each route) was not considered in their model, which prevents them from obtaining higher accuracy. In [4], D. A. Pamplona et al. built an artificial neural network with 4 hidden layers, and achieved the highest accuracy of 87%; their proposed model suggested that the day of the week, block hour, and route has great influence on the flight delay. This model did not consider meteorological factors, so there is room for improvement. Y. J. Kim et al. [12] proposed a model with two stage. The first stage is to predict day-to-day delay status of specific airport by using deep RNN model, where the status was defined as an average delay of all flights arrived at each airport. The second stage is a layered neuron network model to predict the delay of each individual flight using the day-to-day delay status from the first stage and other information. The two stages of the model achieved accuracies of 85% and 87.42%, respectively. This study suggested that the deep learning model



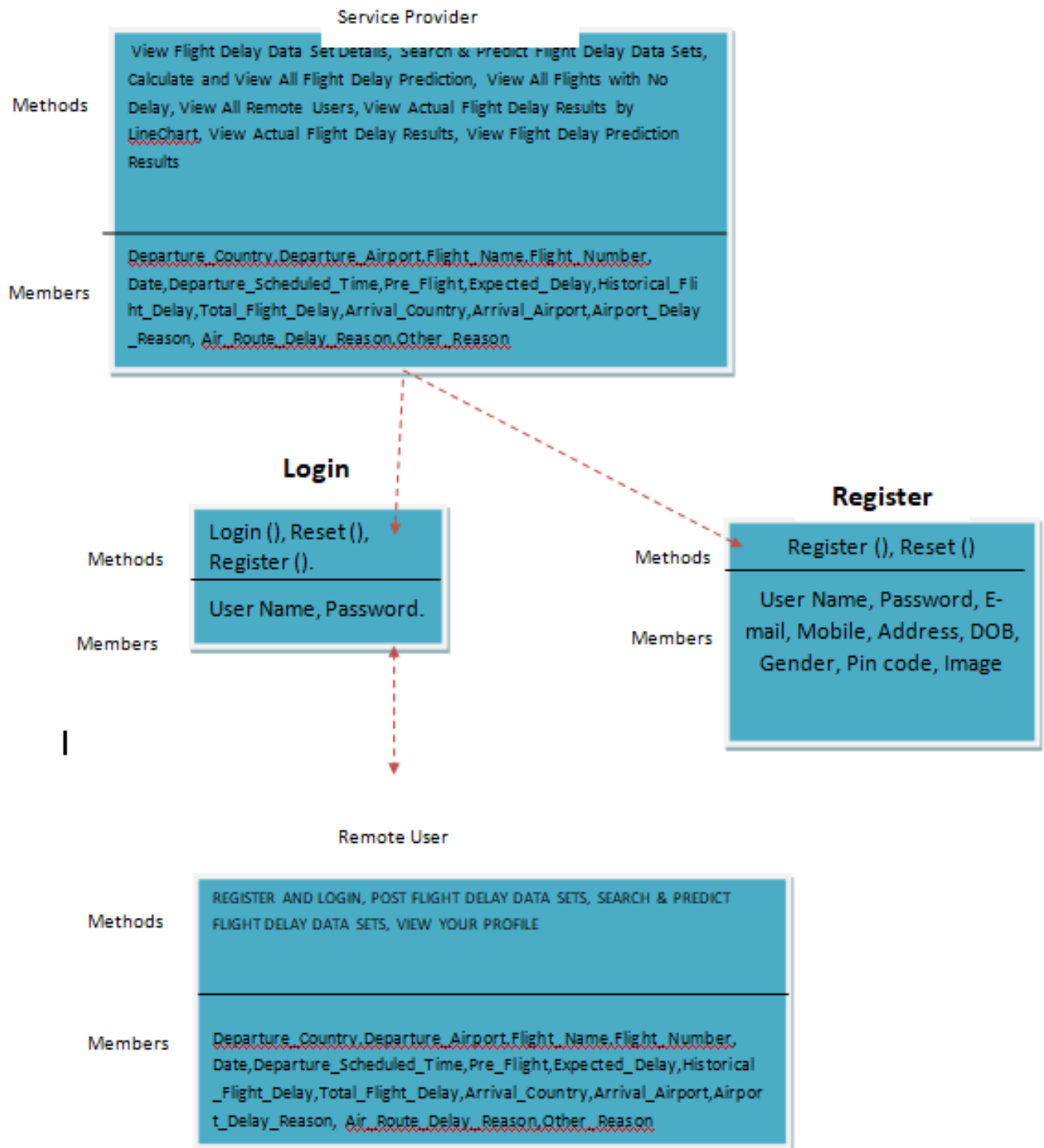
requires a great volumes of data. Otherwise, the model is likely to end up with poor performance or overfitting [13]. To address the problems in ATM, the received ADS-B messages can be coupled with weather information, traffic flow information, and other information to constitute an aviation data lake, which provides an opportunity to find a better approach to accurately predict the flight delay. Meanwhile, machine learning have made great progress and have obtain amazing performance in many domains, such as internet of things [14], heterogeneous network traffic control [15], autonomous driving [16], unmanned aerial vehicle [17]–[21], wireless communications [22]–[28], and cognitive radio [29]–[31]. The above successes motivate us to apply machine learning in the field of air traffic data analytic applications [12], [32]. Compared with the scenarios in wireless communications, the air traffic also faces dynamic environment and can be affected by many dynamic factors. Therefore, it is worthy to apply machine learning models for the flight delay prediction by making full use of the aviation data lake. By combining the advantages of all the available different data, we can feed the entire dataset into specific deep learning models, which allows us to find optimal solution in a larger and finer solution space and gain higher prediction accuracy of the flight delay. Our work benefits from considering as many factors as possible that may potentially influence the flight delay. For instance, airports information, weather of airports, traffic flow of airports, traffic flow of routes. The contributions of this paper can be summarized as follows: • We explore a broader scope of factors which may potentially influence the flight delay and quantize those selected factors. Thus we obtain an integrated aviation dataset. Our experimental results indicate that the multiple factors can be effectively used to predict whether a flight will delay. • Several machine learning based-network architectures are proposed and are matched with the established aviation dataset. Traditional flight prediction problem is a binary classification task. To comprehensively evaluate the performance of the architectures, several prediction tasks covering classification and regression are designed. • Conventional schemes mostly focused on a single route or a single airport [4], [6], [12]. However, our work covers all routes and airports which are within our ADS-B platform.

2. ARCHITECTURE DIAGRAM



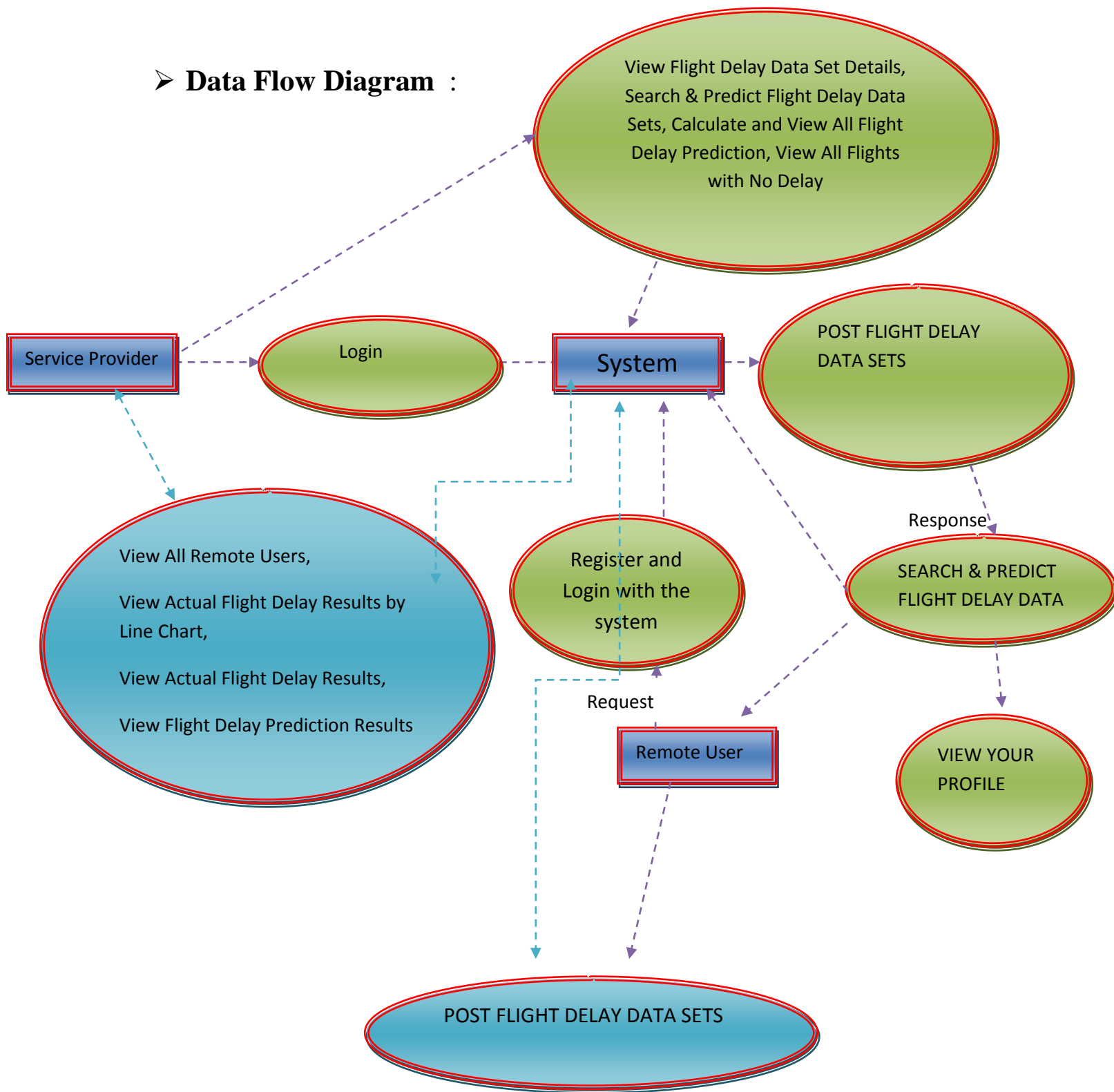


➤ Class Diagram :





➤ Data Flow Diagram :



3. CONCLUSIONS

In this paper, random forest-based and LSTM-based architectures have been implemented to predict individual flight delay. The experimental results show that the random forest based method can obtain good performance for the binary classification task and there are still room for improving the multi-categories classification tasks. The LSTM-based architecture can obtain relatively higher training accuracy, which suggests that the LSTM cell is an effective structure to handle time sequences. However, the over fitting problem occurred in the LSTM based architecture still needs to be solved. In summary, the random forest-based architecture presented better adaptation at a cost of the training accuracy when handling the limited dataset. In order to overcome the overfitting problem and to improve the testing accuracy for multi-categories classification tasks, our future work will focus on collecting or generating more training data, integrating more information like airport traffic flow, airport visibility into our dataset, and designing more delicate networks.

4. REFERENCES

- [1] M. Leonardi, "Ads-b anomalies and intrusions detection by sensor clocks tracking," IEEE Trans. Aerosp. Electron. Syst., to be published, doi: 10.1109/TAES.2018.2886616.
- [2] Y. A. Nijssure, G. Kaddoum, G. Gagnon, F. Gagnon, C. Yuen, and R. Mahapatra, "Adaptive air-to-ground secure communication system based on ads-b and wide-area multilateration," IEEE Trans. Veh. Technol., vol. 65, no. 5, pp. 3150–3165, 2015.
- [3] J. A. F. Zuluaga, J. F. V. Bonilla, J. D. O. Pabon, and C. M. S. Rios, "Radar error calculation and correction system based on ads-b and business intelligent tools," in Proc. Int. Carnahan Conf. Secur. Technol., pp. 1–5, IEEE, 2018.
- [4] D. A. Pamplona, L. Weigang, A. G. de Barros, E. H. Shiguemori, and C. J. P. Alves, "Supervised neural network with multilevel input layers for predicting of air traffic delays," in Proc. Int. Jt. Conf. Neural Networks, pp. 1–6, IEEE, 2018.
- [5] S. Manna, S. Biswas, R. Kundu, S. Rakshit, P. Gupta, and S. Barman, "A statistical approach to predict flight delay using gradient boosted decision tree," in Proc. Int. Conf. Comput. Intell. Data Sci., pp. 1–5, IEEE, 2017.
- [6] L. Moreira, C. Dantas, L. Oliveira, J. Soares, and E. Ogasawara, "On evaluating data preprocessing methods for machine learning models for flight delays," in Proc. Int. Jt. Conf. Neural Networks, pp. 1–8, IEEE, 2018.
- [7] J. J. Rebollo and H. Balakrishnan, "Characterization and prediction of air traffic delays," Transp. Res. Part C Emerg. Technol., vol. 44, pp. 231–241, 2014.
- [8] L. Hao, M. Hansen, Y. Zhang, and J. Post, "New york, new york: Two ways of estimating the delay impact of new york airports," Transp. Res. Part E Logist. Transp. Rev., vol. 70, pp. 245–260, 2014.



- [9] ANAC, “The Brazilian National Civil Aviation Agency.” anac.gov, 2017. [online] Available:<http://www.anac.gov.br/>.
- [10] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, “Efficient knn classification with different numbers of nearest neighbors,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 5, pp. 1774–1785, 2017.

AN EFFICIENT AND PRIVACY-PRESERVING BIOMETRIC

Jampala Lakshmi Thirupathamma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

Biometric identification has become increasingly popular in recent years. With the development of cloud computing, database owners are motivated to outsource the large size of biometric data and identification tasks to the cloud to get rid of the expensive storage and computation costs, which however brings potential threats to users' privacy. In this paper, we propose an efficient and privacy-preserving biometric identification outsourcing scheme. Specifically, the biometric data is encrypted and outsourced to the cloud server. To execute a biometric identification, the database owner encrypts the query data and submits it to the cloud. The cloud performs identification operations over the encrypted database and returns the result to the database owner. A thorough security analysis indicates the proposed scheme is secure even if attackers can forge identification requests and collude with the cloud. Compared with previous protocols, experimental results show the proposed scheme achieves a better performance in both preparation and identification procedures.

1.INTRODUCTION

BIOMETRIC identification has raised increasingly attention since it provides a promising way to identify users. Compared with traditional authentication methods based on passwords and identification cards, biometric identification is considered to be more reliable and convenient [1]. Additionally, biometric identification has been widely applied in many fields by using biometric traits such as fingerprint [2], iris [3], and facial patterns [4], which can be collected from various sensors [5]–[9]. In a biometric identification system, the database owner such as the FBI who is responsible to manage the national fingerprints database, may desire to outsource the enormous biometric data to the cloud server (e.g., Amazon) to get rid of the expensive storage and computation costs. However, to preserve the privacy of biometric data, the biometric data has to be encrypted before outsourcing. Whenever a FBI's partner (e.g., the police station) wants to

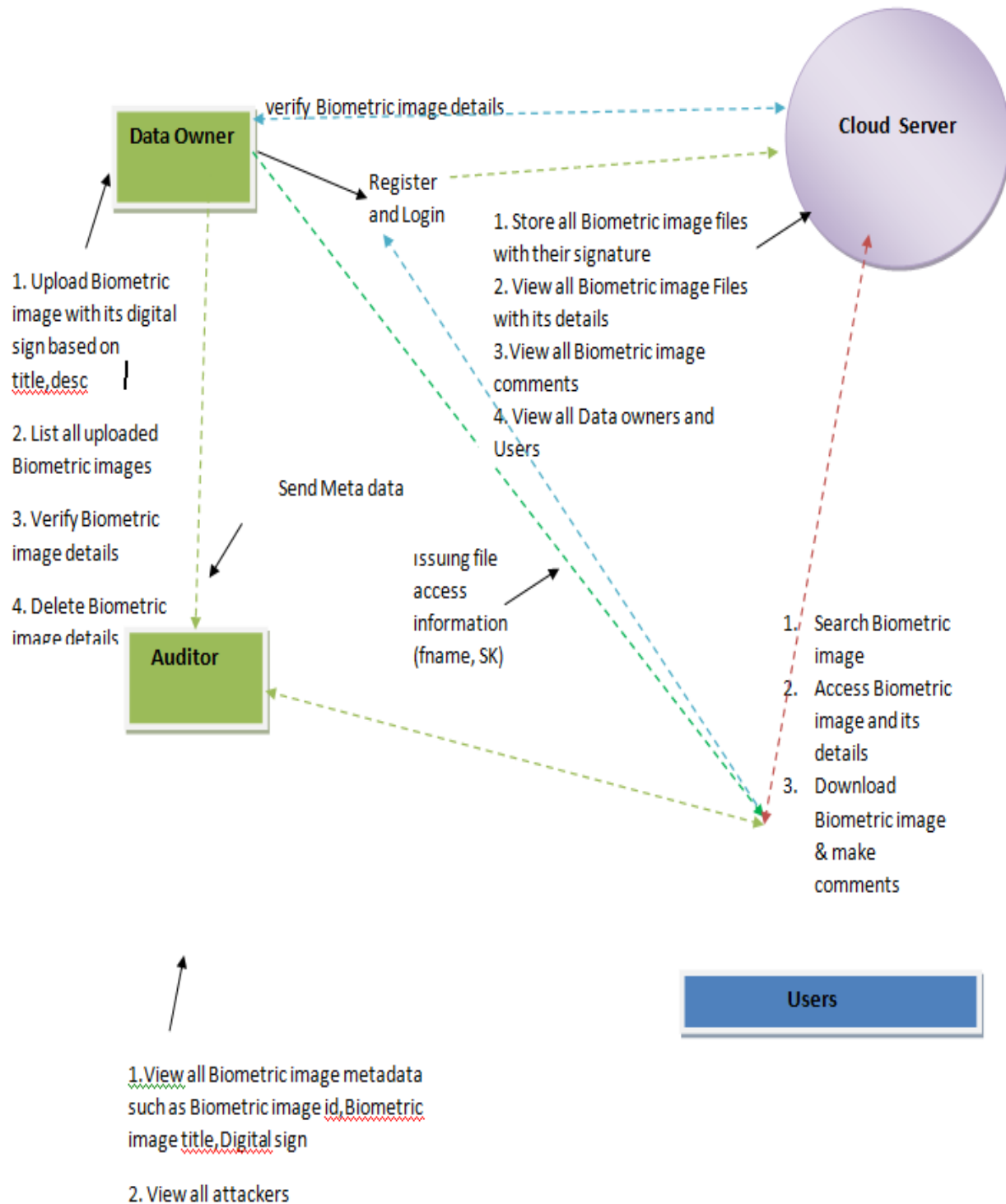
authenticate an individual's identity, he turns to the FBI and generates an identification query by using the individual's biometric traits (e.g., fingerprints, irises, voice patterns, facial patterns etc.). Then, the FBI encrypts the query and submits it to the cloud to find the close match. Thus, the challenging problem is how to design a protocol which enables efficient and privacy-preserving biometric identification in the cloud computing. A number of privacy-preserving biometric identification solutions [10]–[17] have been proposed. However, most of them mainly concentrate on privacy preservation but ignore the efficiency, such as the schemes based on homomorphic encryption and oblivious transfer in [10], [11] for fingerprint and face image identification respectively. Suffering from performance problems of local devices, these schemes are not efficient once the size of the database is larger than 10 MB. Later, Evans et al. [12] presented a biometric identification scheme by utilizing circuit design and ciphertext packing techniques to achieve efficient identification for a larger database of up to 1GB. Additionally, Yuan and Yu [13].

Propose an efficient privacy preserving biometric identification scheme. Specifically, they constructed three modules and designed a concrete protocol to achieve the security of fingerprint trait. To improve the efficiency, in their scheme, the database owner outsources identification matching tasks to the cloud. However, Zhu et al. [18] pointed out that Yuan and Yu's protocol can be broken by a collusion attack launched by a malicious user and cloud. Wang et al. [14] proposed the scheme CloudBI-II which used random diagonal matrices to realize biometric identification. However, their work was proven insecure in [15], [16]. In this paper, we propose an efficient and privacy preserving biometric identification scheme which can resist the collusion attack launched by the users and the cloud. Specifically, our main contributions can be summarized as follows:

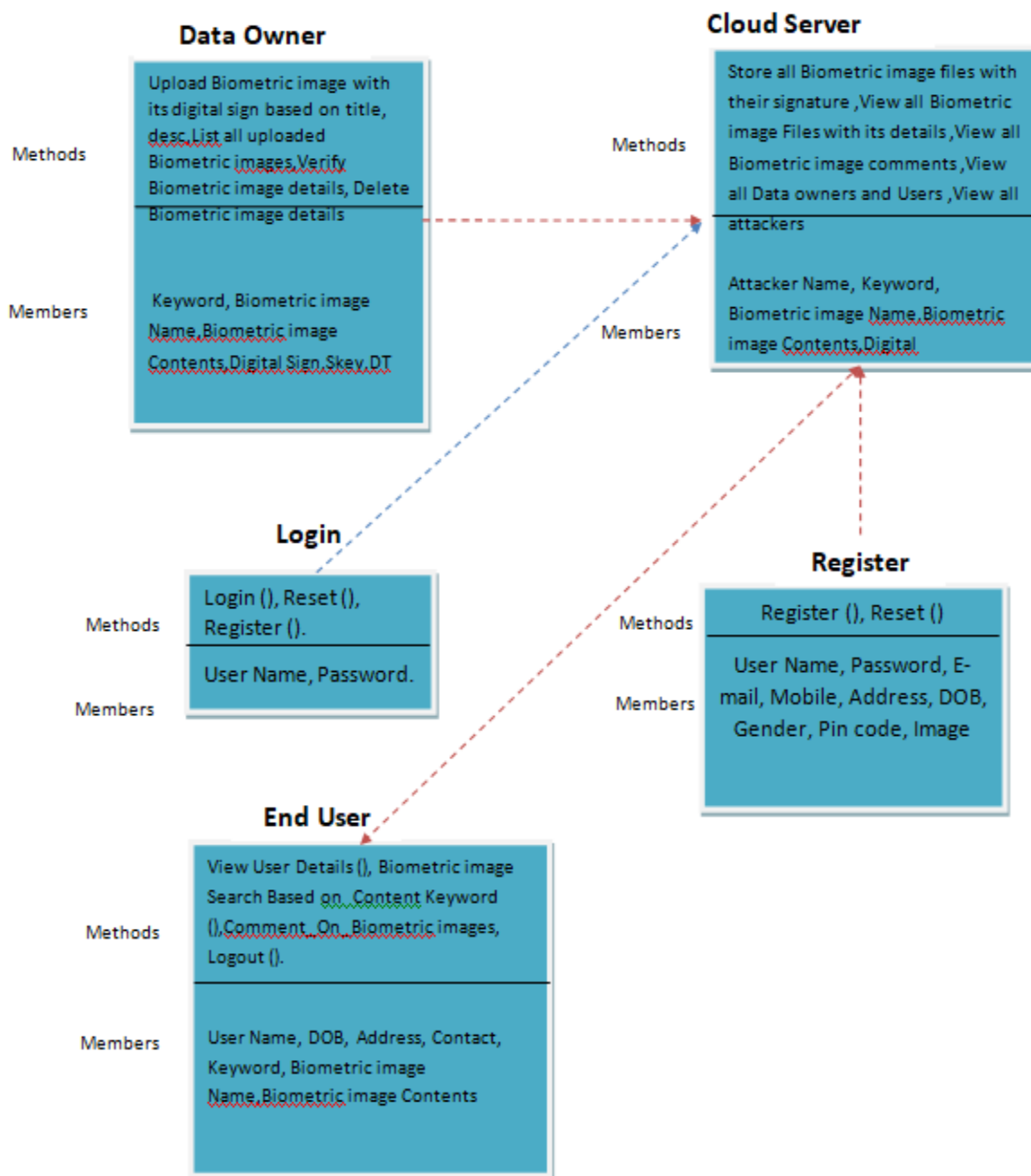
- We examine the biometric identification scheme [13] and show its insufficiencies and security weakness under the proposed level 3 attack. Specifically, we demonstrate that the attacker can recover their secret keys by colluding with the cloud, and then decrypt the biometric traits of all users.
- We present a novel efficient and privacy-preserving biometric identification scheme. The detailed security analysis shows that the proposed scheme can achieve a required level of privacy protection. Specifically, our scheme is secure under the biometric identification outsourcing model and can also resist the attack proposed by [18].
- Compared with the existing biometric identification schemes, the performance analysis shows

that the proposed scheme provides a lower computational cost in both preparation and identification procedures. The remainder of this paper is organized as follows:

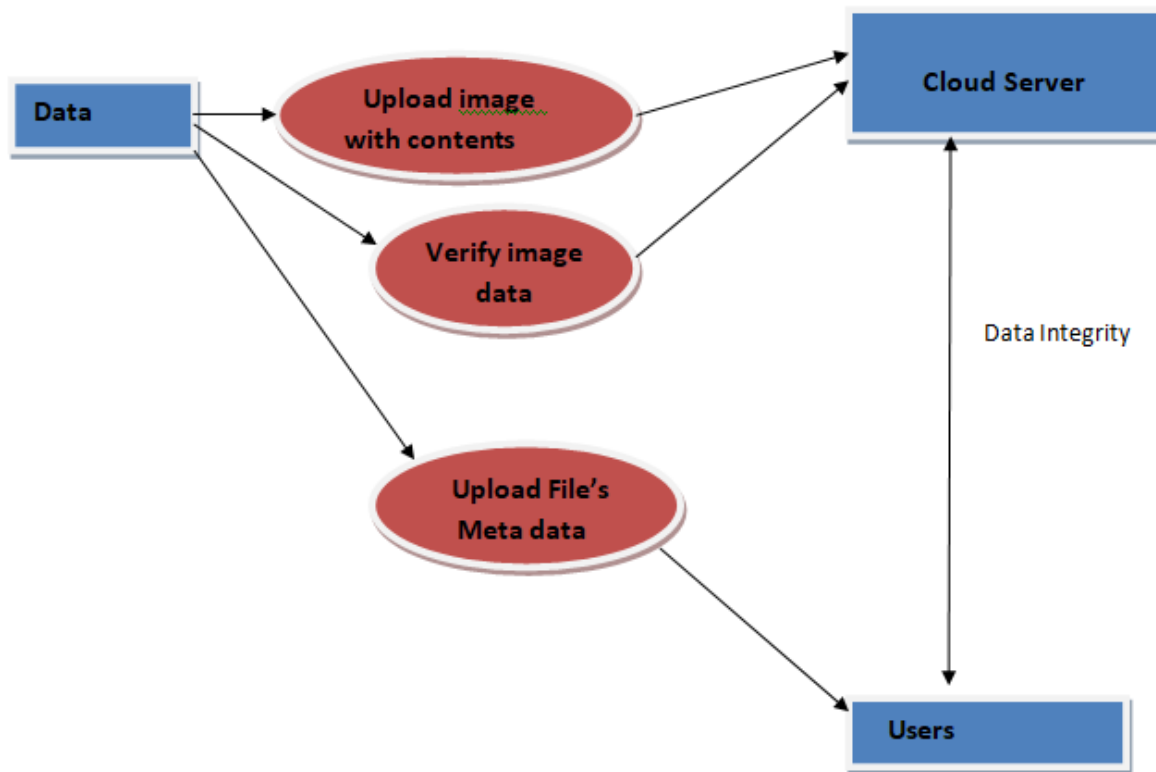
Architecture Diagram



Class Diagram



Data Flow Diagram



2.SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3.SYSTEM TESTING

Software once validated must be combined with other system elements (e.g. Hardware, people, database). System testing verifies that all the elements are proper and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

UNIT TESTING

In unit testing different are modules are tested against the specifications produced during the design for the modules. Unit testing is essential for verification of the code produced during the coding phase, and hence the goals to test the internal logic of the modules. Using the detailed design description as a guide, important Conrail paths are tested to uncover errors within the boundary of the modules. This testing is carried out during the programming stage itself. In this type of testing step, each module was found to be working satisfactorily as regards to the expected output from the module.

In Due Course, latest technology advancements will be taken into consideration. As part of technical build-up many components of the networking system will be generic in nature so that future projects can either use or interact with this. The future holds a lot to offer to the development and refinement of this project.

4.CONCLUSION

In this paper, we proposed a novel privacy-preserving biometric identification scheme in the cloud computing. To realize the efficiency and secure requirements, we have designed a new encryption algorithm and cloud authentication certification. The detailed analysis shows it can resist the potential attacks. Besides, through performance evaluations, we further demonstrated the proposed scheme meets the efficiency need well.

5 .REFERENCES

- [1] A. Jain, L. Hong and S. Pankanti, "Biometric identification," Communications of the ACM, vol. 43, no. 2, pp. 90-98, 2000.
- [2] R. Allen, P. Sankar and S. Prabhakar, "Fingerprint identification technology," Biometric Systems, pp. 22-61, 2005.

- [3] J. de Mira, H. Neto, E. Neves, et al., "Biometric-oriented Iris Identification Based on Mathematical Morphology," *Journal of Signal Processing Systems*, vol. 80, no. 2, pp. 181-195, 2015.
- [4] S. Romdhani, V. Blanz and T. Vetter, "Face identification by fitting a 3d morphable model using linear shape and texture error functions," in *European Conference on Computer Vision*, pp. 3-19, 2002.
- [5] Y. Xiao, V. Rayi, B. Sun, X. Du, F. Hu, and M. Galloway, "A survey of key management schemes in wireless sensor networks," *Journal of Computer Communications*, vol. 30, no. 11-12, pp. 2314-2341, 2007.
- [6] X. Du, Y. Xiao, M. Guizani, and H. H. Chen, "An effective key management scheme for heterogeneous sensor networks," *Ad Hoc Networks*, vol. 5, no. 1, pp. 24-34, 2007.
- [7] X. Du and H. H. Chen, "Security in wireless sensor networks," *IEEE Wireless Communications Magazine*, vol. 15, no. 4, pp. 60-66, 2008.
- [8] X. Hei, and X. Du, "Biometric-based two-level secure access control for implantable medical devices during emergency," in *Proc. of IEEE INFOCOM 2011*, pp. 346-350, 2011.
- [9] X. Hei, X. Du, J. Wu, and F. Hu, "Defending resource depletion attacks on implantable medical devices," in *Proc. of IEEE GLOBECOM 2010*, pp. 1-5, 2010.
- [10] M. Barni, T. Bianchi, D. Catalano, et al., "Privacy-preserving fingercode authentication," in *Proceedings of the 12th ACM workshop on Multimedia and security*, pp. 231-240, 2010.

DETECTION OF STROKE DISEASE USING ML

Jonnada Suresh (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract: A stroke is a medical condition in which poor blood flow to the brain results in cell death. It is now a day a leading cause of death all over the world. Several risk factors believe to be related to the cause of stroke has been found by inspecting the affected individuals. Using these risk factors, a number of works have been carried out for predicting and classifying stroke diseases. Most of the models are based on data mining and machine learning algorithms. In this work, we have used four machine learning algorithms to detect the type of stroke that can possibly occur or occurred form a person's physical state and medical report data. We have collected a good number of entries from the hospitals and use them to solve our problem. The classification result shows that the result is satisfactory and can be used in real time medical report. We believe that machine learning algorithms can help better understanding of diseases and can be a good healthcare companion.

Index Terms—Stroke, machine learning, WEKA, Naive Bayes, J48, k-NN, Random Forest.

1. INTRODUCTION

A stroke occurs due to poor blood flow to the brain which results in cell death. Two main types of stroke are ischemic stroke and haemorrhagic stroke. Ischemic stroke occurs due to lack of blood flow and haemorrhagic stroke occurs due to bleeding [1]. Another type of stroke is transient ischemic attack. Ischemic stroke has two categories- embolic stroke and thrombotic stroke. An embolic stroke occurs by forming a clot in any part of the body and moves toward the brain and blocks blood flow. A thrombotic stroke caused by a clot that weakens blood flow in an artery. Haemorrhagic stroke is classified into two types- subarachnoid haemorrhage and intracerebral haemorrhage. Transient ischemic attack is also known as "ministroke".

A large number of people lose their life due to stroke and it is increasing in developing countries [3]. There are several stroke risk factors that regulate different types of stroke. Predictive algorithms help to understand the relation between these risk factors to types of strokes. The machine learning algorithm can improve patients' health through early detection

and treatment. We have used several machine learning algorithms to detect the type of stroke that can occur in a patient or already occurred from their clinical report and statistical data. We have built a stroke dataset by collecting data from various sources validated by medical experts. Then the dataset was processed to be used with the machine learning algorithms. We have built several models of classification. The result of the models is satisfactory and can be used in a realtime patient's stroke classification.

2. LITERATURE SURVEY

1. Thrombophilia testing in young patients with ischemic stroke :

The possible significance of thrombophilia in ischemic stroke remains controversial. We aimed to study inherited and acquired thrombophilias as risk factors for ischemic stroke, transient ischemic attack (TIA) and amaurosis fugax in young patients.

We included patients aged 18 to 50 years with ischemic stroke, TIA or amaurosis fugax referred to thrombophilia investigation at Aarhus University Hospital, Denmark from 1 January 2004 to 31 December 2012 (N = 685). Clinical information was obtained from the Danish Stroke Registry and medical records. Thrombophilia investigation results were obtained from the laboratory information system. Absolute thrombophilia prevalences and associated odds ratios (OR) with 95% confidence intervals (95% CI) were reported for ischemic stroke (N = 377) and TIA or amaurosis fugax (N = 308). Thrombophilia prevalences for the general population were obtained from published data.

2. Classification of stroke disease using machine learning algorithms :

This paper presents a prototype to classify stroke that combines text mining tools and machine learning algorithms. Machine learning can be portrayed as a significant tracker in areas like surveillance, medicine, data management with the aid of suitably trained machine learning algorithms. Data mining techniques applied in this work give an overall review about the tracking of information with respect to semantic as well as syntactic perspectives. The proposed idea is to mine patients' symptoms from the case sheets and train the system with the acquired data. In the data collection phase, the case sheets of 507 patients were collected from Sugam Multispecialty Hospital, Kumbakonam, Tamil Nadu, India. Next, the case sheets were mined using tagging and maximum entropy methodologies, and the proposed stemmer extracts the common and unique set of attributes to classify the strokes. Then, the processed data were fed into various machine learning algorithms such as artificial neural networks, support vector machine, boosting and bagging and random forests. Among

these algorithms, artificial neural networks trained with a stochastic gradient descent algorithm outperformed the other algorithms with a higher classification accuracy of 95% and a smaller standard deviation of 14.69.

3. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4. SYSTEM DESIGN

4.1 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

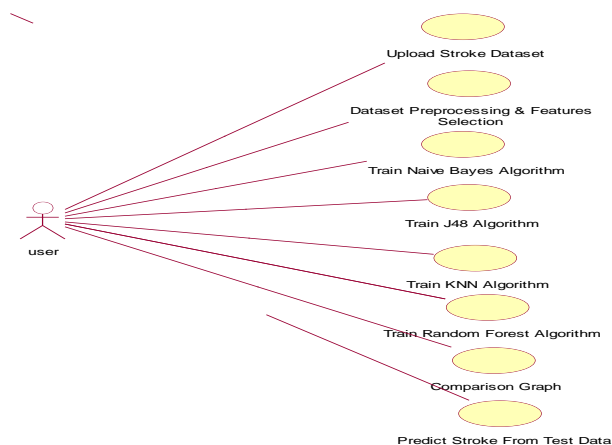
GOALS:

The Primary goals in the design of the UML are as follows:

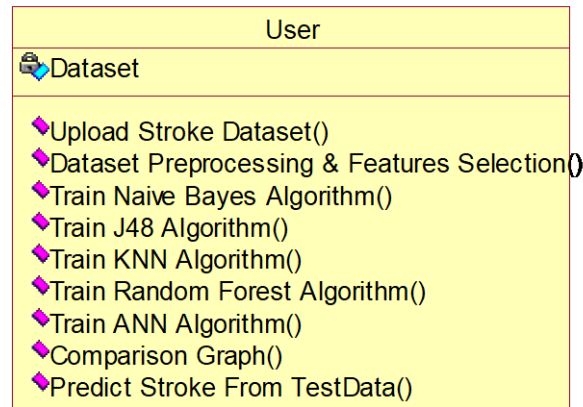
1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

**CLASS DIAGRAM:**

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



5. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the

combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

5.1 IMPLEMENTATION: MODULES

To implement this project we have designed following modules

1) Upload Stroke Dataset:

using this module we will upload dataset to application

2) Dataset Preprocessing & Features Selection:

using this module we will clean dataset by replacing missing values with 0 and then apply label encoding algorithm to convert non-numeric values to numeric values and then select features from dataset and then split dataset into train and test where application used 80% data for training and 20% for testing

3) Train Naive Bayes Algorithm:

above training data will be input to Naïve Bayes algorithm to train a model and this model will be applied on test data to calculate accuracy

4) Train J48 Algorithm:

above training data will be input to J48 algorithm to train a model and this model will be applied on test data to calculate accuracy

5) Train KNN Algorithm:

above training data will be input to KNN algorithm to train a model and this model will be applied on test data to calculate accuracy

6) Train Random Forest Algorithm:

above training data will be input to Random Forest algorithm to train a model and this model will be applied on test data to calculate accuracy

7) Train ANN Algorithm:

above training data will be input to ANN algorithm to train a model and this model will be

6. CONCLUSION

In this paper, a sufficiently large dataset of stroke attacked patients has been classified accurately. Four classifiers such as TABLE XI: Confusion matrix for Random Forest algorithm.

	a	b	c	d
a	436	0	0	1
b	0	301	1	0
c	0	0	142	0
d	0	0	0	177

Naive Bayes, J48, k-NN, and Random Forest were used for detection of stroke disease. From the performance analysis we see that Naive Bayes performs better than other methods. The novelty and the main contribution of our work are collecting this dataset and preparing them to use with WEKA. The model can help people with a cautionary indication of being affected by stroke. Healthcare industries generate huge amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. Which is very difficult to relate to one another even by a field expert. It will help the clinician to better understand the type of disease. The limitations of our method are that the dataset is not perfectly symmetrical. However, it did not affect the predicted accuracy for the other algorithms. Naive Bayes algorithm didn't work as we expected.

In future work, it is possible to extend the research by using different classification techniques. Moreover, the prediction of stroke can be done by adding some non-stroke data with the existing dataset.

7. REFERENCES

- [1] S. H. Pahus, A. T. Hansen, and A.-M. Hvas, "Thrombophilia testing in young patients with ischemic stroke," *Thrombosis research*, vol. 137, pp. 108–112, 2016.

- [2] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Computing and Applications*, pp. 1–12.
- [3] L. T. Kohn, J. Corrigan, M. S. Donaldson, et al., *To err is human: building a safer health system*, vol. 6. National academy press Washington, DC, 2000.
- [4] R. Jeena and S. Kumar, "Stroke prediction using svm," in *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 600–602, IEEE, 2016.
- [5] P. A. Sandercock, M. Niewada, and A. Członkowska, "The international stroke trial database," *Trials*, vol. 13, no. 1, pp. 1–1, 2012.
- [6] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 158–161, IEEE, 2017.
- [7] S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," *Int J Comput Appl*, vol. 149, no. 10, pp. 26–31, 2016.
- [8] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," *International Journal of Computer Applications*, vol. 43, no. 14, pp. 26–31, 2012.
- [9] G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," *International Journal of Computer Applications*, vol. 98, no. 22, 2014.
- [10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

PROTOTYPING MOBILE APP

Kadali Kalyan (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y.Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract: This paper presents the prototyping and evaluation of a mobile application that is being developed to promote territorial-based innovation. It also describes the methodological procedures adopted for the application design and evaluation phases involved in its construction. To carry out a proof of concept of the prototype with a focus on usability, remote accessibility and microcopy evaluations were carried. From these evaluations, it was possible to gather several inputs enabling the improvement of the prototype interface, resulting in a more robust, reliable and feasible version of the application, thus contributing to meet the goals of the CeNTER program.

Keywords: Hypermediation, Digital Technologies, Territorial Innovation, Community-led initiatives, UserCentered Design, Mobile Application, Accessibility, Microcopy, Remote Testing.

1. INTRODUCTION

The general contribution of this paper is to present the concept of a mobile application, designed and prototyped under the scope of the CeNTER research program, to promote territorial innovation. CeNTER – Community-led Networks for Territorial Innovation – is an interdisciplinary research program running on the University of Aveiro, Portugal, since 2017, and one of its goals is to identify and develop digital tools that can best leverage the potential of the Centro region of Portugal.

The article reports the development and validation processes of the prototype of a mobile application to foster territorial innovation. The prototype simulates the functionalities of a digital hypermediation platform designed to support community-led initiatives focusing on Tourism and Health and Well-being sectors.

The article is organized as follows: section 2 introduces the theoretical background concerning user-centered design approach to the development of mobile applications adopted and also summarizes some related work; section 3 formalizes the methodological procedures adopted on the CeNTER app design and evaluation; finally, section 4 presents the main conclusions and discusses some future research paths.

Prototype The prototyping process began with the choice of the smartphone as the target user's device. This choice led to the analysis of the challenges inherent to the development of an application for mobile devices, considering concepts such as interaction design, usability and interface design. Thereafter, the methodological procedure to be employed was conceived having a focus on a UCD approach, as aforementioned. Based on the previously defined app CeNTER requirements, previously defined directly involving different

stakeholders (Renó et al, 2019), the conceptual phase of the prototype started by drawing in paper a first set of mockups. These mockups were then used to consolidate the basic requirements and specifications of the application design. With the Sketch software, wireframes that outline the skeleton of the different screens were developed, such as text boxes, images and other graphic elements, which compose the interface of each screen. With these wireframes, along with the support of Principle software, the interaction and styles of the prototype were developed. Furthermore, also using Principle, it was possible to conceive a prototype with greater fidelity, deploying the style and interaction like what was intended to be obtained in the final product (Carvalho et al., 2020).

2. METHODOLOGY FOR DESIGN AND EVALUATION

The first phase of the project included the following procedures: i) systematic literature review (Silva et al. 2020); ii) mapping of innovative initiatives in the Centro region of Portugal (Tymoshchuk et al., 2019c); iii) benchmarking of social networks, applications and websites (Martínez-Rolán et al., 2019; Renó et al., 2021 ; Tymoshchuk et al., 2019a); iv) interviews with leaders of local communities and entities (Renó et al., 2018; Silva et al., 2018; Tymoshchuk et al., 2019b); v) organization of two focus groups with representatives of small and large community initiatives (Silva et al. 2019). The focus of this article is on the second phase of the project. As discussed before, a user-centered approach was adopted in the design process. Following this approach, a digital platform prototype was developed to support different usage scenarios (use cases). This phase had four main steps. I) A medium fidelity prototype was specified and developed for iPhone using the Principle software. For this prototype, a set of (78) screens were developed to simulate the specified functionalities offering navigation through elements and allowing an effective interaction. II) After achieving a stable version of the prototype, a heuristic testing was conducted with experts. This prototype intended to offer a simulated look and feel of a real mobile application, allowing the user to engage himself in a pleasant experience (Seifi, 2015). Continuous feedback from the users, in the initial stages of development of technological products, is crucial for detecting possible problems of a system. The prototype was previously evaluated by an internal team to obtain validation before moving on to the next steps. III) An accessibility assessment with an expert was remotely conducted. This process was moderated using ApowerMirror software, which allowed the expert to access the prototype, as well as the recording of the interactions. The results were transcribed and analysed using a grid, defining the levels of priority and complexity of the problems to be fixed. One of the main problems highlighted by the expert was the low contrast between graphic and text elements. Those issues were solved using contrast checker tools aligned with success criteria of WCAG 2.0. IV) Finally, microcopy methodology was used to evaluate prototype's content by an expert. This evaluation, that covered the 215 textual elements presented on the screens, was intended to identify whether the textual content used was adequate and correct, considering the context of the project, and both lexical and grammatical aspects. Furthermore, the expert was invited to provide suggestions for alternatives to the textual elements which, in his opinion, should be modified. The results obtained lead to several improvements of the final version of the prototype.

3. INPUT AND OUTPUT DESIGN INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is

designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the □ Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

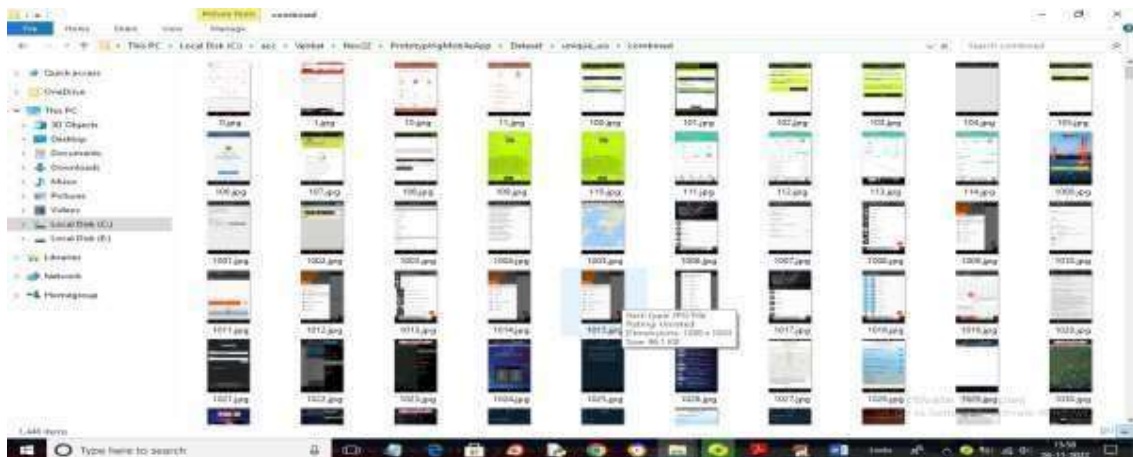
4. RESULT

Machine Learning-Based Prototyping of Graphical User Interfaces for Mobile Apps

In propose paper author is applying CNN Neural Network algorithm to predict code for given Android GUI screen. It's common for developer to generate code for given graphical user interface but this process will take lot of experience and manual work. To overcome from above issue author is training CNN with RICO dataset which consists of CODE in JSON format and GUI images. After training we can apply this CNN model on any android screen to generate or predict code.

Predicted code will be in the form of JSON and we can use below ANDROID APP to convert that JSON code to ANDROID layout <https://github.com/flipkart-incubator/proteus>

In above link we can see by giving JSON code we can get android code. In below screen I am showing images used to train CNN



Above images are from RICO dataset which can I downloaded from KAGGLE by typing RICO dataset on Google.

To implement this project we have designed following modules

- 1) Upload RICO Dataset: using this module we will upload dataset to application
- 2) Preprocess Dataset: using this module we will read each image and then resize and normalize all pixel values from the image
- 3) Shuffling, Splitting & Dataset Normalization: using this module we will shuffle dataset and then split dataset into train and test where application will used 80% dataset for training and 20% for testing
- 4) Run CNN Algorithm: now 80% train data will be input to train CNN and then apply 20% test data to calculate CNN prediction accuracy confusion matrix
- 5) CNN Training Graph: using this module we will plot CNN training accuracy and loss graph
- 6) Predict Code from Image: using this module we will upload test GUI screen and then CNN will predict android code in JSON format.

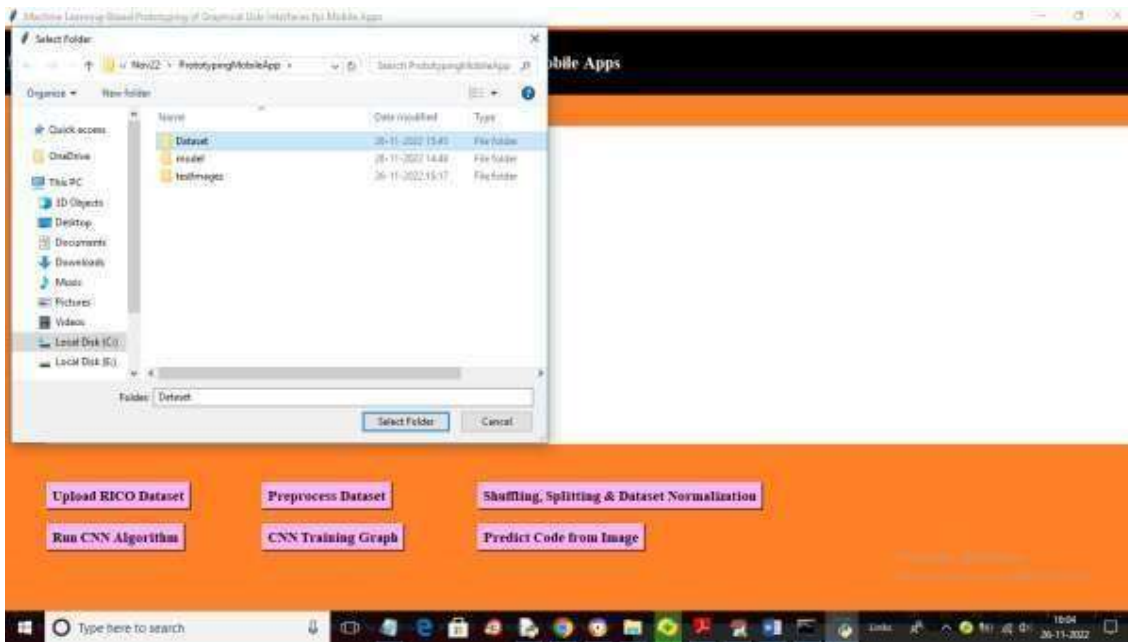
NOTE: in paper also author used RICO dataset and this dataset given training code in JSON format so we can get generated code also in JSON format

SCREEN SHOTS

To run project double click on 'run.bat' file to get below screen



In above screen click on 'Upload RICO Dataset' button to upload dataset and get below output



In above screen selecting and uploading 'Dataset' folder and then click on 'Select Folder' button to load dataset and get below output



In above screen click on 'Preprocess Dataset' button to read and process each image and get below output

4. CONCLUSIONS

This article reported the main aspects of the prototyping and evaluation processes of a mobile application for territorial innovation, developed under the scope of the CeENTER research program. The theoretical framework employed for designing and developing the mobile app

was an UCD approach. The prototype was developed using the Principle software for Mac OS and targeting the use of mobile phones. The accessibility evaluation was carried out through expert evaluation and the use of accessibility evaluation tools. Several relevant issues, regarding screen contrast, involving foreground and background colours and thickness, were identified and corrected. Regarding microcopy evaluation, it was possible to achieve results that enabled the improvement of the textual elements used in the prototype. Accessibility and microcopy evaluations with experts were adapted and conducted remotely because of the pandemic situation.

Currently, another set of evaluations is being prepared with different use cases related to the areas of Tourism and Health and Well-being. These evaluations will be conducted with a group of evaluators that represent public or private enterprises and organizations, and also individuals, with direct connection with the areas under scrutiny.

Future research plans include the design and development of the full CeNTER digital platform, and the design of the database that will support the server side of the CeNTER platform is already underway. As a limitation of this study, the prototype was only developed for iPhone interfaces, ignoring its impacts on other platforms like Android, tablets or PC. Nevertheless, this does not devalue its potentiality to be implemented in the future as a crossplatform application.

Furthermore, the number of testers that participated in the evaluation limited the generalization of the results, so further refinement of the evaluations is strongly recommended.

5. REFERENCES

- Ballantyne, M., Jha, A., Jacobsen, A., Hawker, J. S. & El-Glaly, Y. (2018). Study of accessibility guidelines of mobile applications. In Proceedings of the 17th international conference on mobile and ubiquitous multimedia, (pp. 305-315).
- Billi, M., Burzagli, L., Catarci, T., Santucci, G., Bertini, E., Gabbanini, F. & Palchetti, E. (2010). A unified methodology for the evaluation of accessibility and usability of mobile applications. *Universal Access in the Information Society*, 9(4), 337-356.
- Carvalho, D., Oliveira, E., Tymoshchuk, O., Antunes, M. J., Pedro, L., Almeida, M., Carvalho, D. & Ramos, F. (2020). Prototipagem de uma Plataforma Digital para a Promoção da Inovação Territorial de Base Comunitária. *Journal of Digital Media & Interaction*, 3(6), 53-71. <https://doi.org/10.34624/jdmi.v3i6.15517>
- Fonseca, M. J., Campos, P., & Gonçalves, D. (2012). *Introdução ao Design de Interfaces* (2a Edição). FCA - Editora de Informática. <https://www.fca.pt/pt/catalogo/informatica/designmultimedia/introducao-ao-design-deinterfaces/>
- ISO. (2018). *Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts* (ISO 9241-11:2018).
- Knight, W. (2019). *UX for Developers*. Northampton: Apress.



- Martínez-Rolán, X., Tymoshchuk, O., Piñero-Otero, T., & Renó, D. (2019). Instagram como red de promoción e hipermediación del turismo rural: el caso de Aldeias Históricas. *Revista Latina de Comunicación Social*, 74, (1610-1632).
- Moeller, S., Joseph, A., Lau, J., & Carbo, T. (2011). Towards Media and Information Literacy Indicators. In *Background Document of the Expert Meeting Towards* (p. 53). Paris: UNESCO. Retrieved from <https://www.ifla.org/publications/towards-media-and-information-literacy-indicators?iframe=true&width=95%25&height=95%25>
- Nielsen, J. (1994). 10 Usability Heuristics for User Interface Design. Retrieved from <https://www.nngroup.com/articles/ten-usability-heuristics/>
- Nielsen, J. (2012). Usability 101: Introduction to usability (2012). Retrieved from <http://www.nngroup.com/articles/usability-101-introduction-to-usability/>
- Norman, D. A. (2006). *O Design do dia a dia*. Rio de Janeiro: Editora Rocco.
- Oliveira, E., Castello Branco, A., Carvalho, D., Tymoshchuk, O., Almeida, M., Antunes, M. J., Pedro, L., & Ramos, F. (2021, in press). Accessibility and microcopy remote testing of mobile applications: The case of the CeNTER platform. In *Proceedings of CISTI'2021 - 16th Iberian Conference on Information Systems and Technologies*.
- Preece, J., Sharp, H., & Rogers, Y. (2019). *Interaction Design - beyond human-computer interaction* (5th ed.). Indianapolis: John Wiley & Sons.

SCA-SYBIL BASED COLLISION ATTACKS

Kadali Vineesha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract—With the massive amounts of data generated by Industrial Internet of Things (IIoT) devices at all moments, federated learning (FL) enables these distributed distrusted devices to collaborate to build machine learning model while maintaining data privacy. However, malicious participants still launch malicious attacks against the security vulnerabilities during model aggregation. This paper is the first to propose sybil-based collusion attacks (SCA) in the IIoT-FL system for the vulnerabilities mentioned above. The malicious participants use label flipping attacks to complete local poisoning training. Meanwhile, they can virtualize multiple sybil nodes to make the local poisoning models aggregated with the greatest possibility during aggregation. They focus on making the joint model misclassify the selected attack class samples during the testing phase, while other non-attack classes kept the main task accuracy similar to the nonpoisoned state. Exhaustive experimental analysis demonstrates that our SCA has superior performance on multiple aspects than the state-of-the-art.

1.INTRODUCTION

WITH the fast development of industry 4.0 and the widespread popularity of industrial Internet of Things (IIOT) applications makes applications such as smart transportation and smart healthcare thrive and also makes the data generated by the industrial devices exponentially grow. Such as autonomous driving technology [1], it needs to train all data generated by sensor and camera devices to build a stable joint model to identify road conditions. And the distributed IIOT devices can generate a large amount of data in a short time [2]. In order to take into account the efficiency of processing big data and protect the privacy of clients. A novel machine learning paradigm named federated learning (FL) [3] was proposed, which is a new solution based on distributed training to alleviate the performance bottleneck and privacy risk caused by centralized processing. Traditional machine learning methods [4] usually store and run these data centrally, which will generate considerable computational and communication overhead in involving millions of mobile devices or

massive data. This makes it unacceptable for sensitive IIOT applications (e.g., autonomous driving, intelligent robots, smart medical) that require real-time data transmission [5]. In addition, relying on centralized storage will cause a huge risk of private leakage [6]. Generally, when FL performs the collaborative training process of multiple distributed participants (e.g., IIOT devices), the sensitive information and private data of each client are kept locally [7]. FL has demonstrated excellent performance in the distributed execution process, while ensuring the privacy of participants by performing independent local training and model updates, so as to implement collaborative calculating in a joint environment that includes malicious participants. This also makes FL attract much attention in many fields including smart healthcare [8] [9], smart feature prediction [10], and Internet of Things in smart homes

2.EXISTING SYSTEM

Xie et al. [22] manipulated a subset of training data by injecting adversarial triggers to perform the wrong prediction on images embedded with triggers in a distributed heterogeneous dataset. Sun et al. [23] injected backdoor tasks into a part of the images to damage the global model's performance on the target task. Although it has a high attack success rate, it can cause much overhead to inject backdoor triggers into large-scale training samples. In addition, the goal of our attack is to misclassify the selected attack class samples. So in this work, we use the label flipping poisoning attacks. Malicious adversaries can perform label flipping attacks without conducting parameter interaction, changing the FL architecture, and pre-training. They use the dirty data with the wrong label for training locally. This attack method is both concealed and direct.

Jiang et al. [25] proposed a sybilbased attacks method. Sybil clients compromised the infected device to update the poisoning model directly. They proved their effectiveness on several advanced defense methods, while also slowing down the convergence of the global model. Fung et al. [26] also designed a novel sybil-based attacks technology, it has shown the effectiveness on multiple recent distributed machine learning fault tolerance protocols. The sybil attacks also showed an excellent attack effect in IoT applications [27]. Although they have shown reliability in the attack effect, the drift gradient of their local poisoning model is very easy to detect and remove. In this paper, we integrate the sybil-based collusion attacks

technology to make the local poisoning model have a higher possibility of aggregation and help malicious participants better obscure the attack behavior.

Taheri et al. [28] proposed two dynamic poisoning attack strategies that integrate Generative Adversarial Network (GAN) and Federated Generative Adversarial Network (FedGAN) on the side of the participants, and evaluated them on IIoT applications. Lim et al. [29] studied the collusion attacks between dishonest participants and the server. The malicious participant uploads the poisoning model during the aggregation stage, and the server also leaks the parameters of other participants to the malicious participant. They aim to achieve the purpose of reducing the global model's performance while analyzing the local model of other participants to avoid anomaly detection [30] during the poisoning process

Disadvantages

The system is not implemented the use the cloning properties of sybil that all sybil nodes virtualized by malicious participants will perform the same malicious operations during the training process and have equal attack influence.

The system is not implemented SCA on IIoT-FL model.

3.PROPOSED SYSTEM

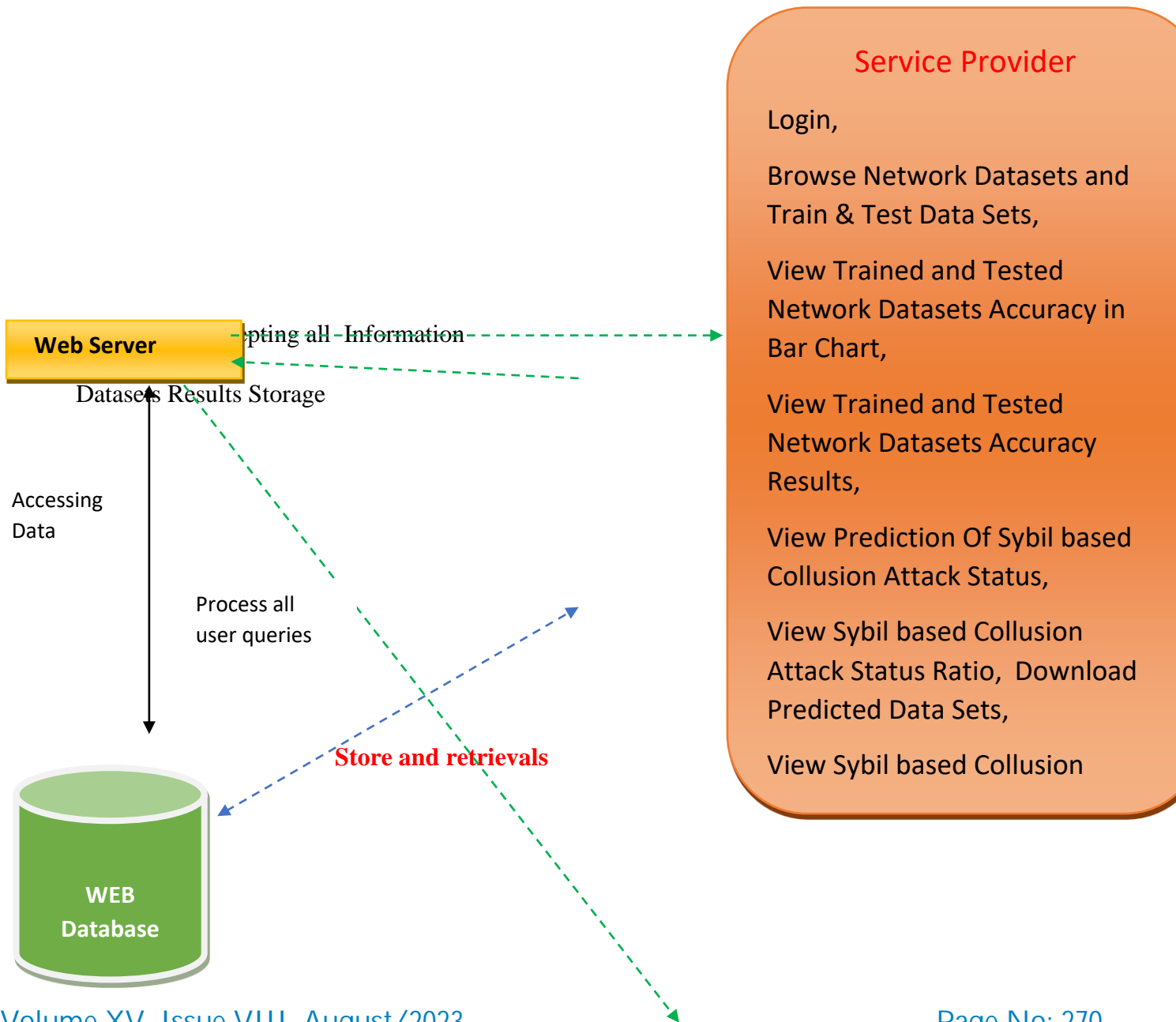
- The proposed system explores sybil-based collusion attacks of IIoT data poisoning for the IIoT-FL application, and implement poisoning training and model collusion attacks in this IIoT-FL system.
- The proposed system makes minimal malicious assumptions for malicious adversaries and integrate the label flipping poisoning attacks to make the global model misclassify the selected attack class samples while maintaining the main task accuracy of other non-attack classes.
- The proposed system further propose an efficient sybil-based collusion attacks (SCA) method, which aims to make the poisoning collusion models to be aggregated with greater probability during aggregation, and successfully obscure their attack behavior.
- The proposed system utilizes F-MNIST and CIFAR-10 datasets to represent the data generated by IIoT devices. Exhaustive experimental analysis demonstrates that our SCA has superior performance than the state-of-the-art.

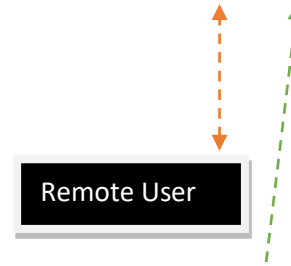
Advantages

The system is implemented SCA Based on Label Flipping Poisoning Attacks which is more secured and safe.

In the proposed system, the system is implemented scenario that the malicious adversary uses the label flipping strategy to train the poisoning data locally and collude with other poisoning models.

Architecture Diagram





REGISTER AND LOGIN,

PREDICT SYBYL BASED COLLUSION ATTACK STATUS,

VIEW YOUR PROFILE.

4.SYSTEM TESTING

Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing. During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

The following are the types of Integration Testing:

Top Down Integration

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner.

In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

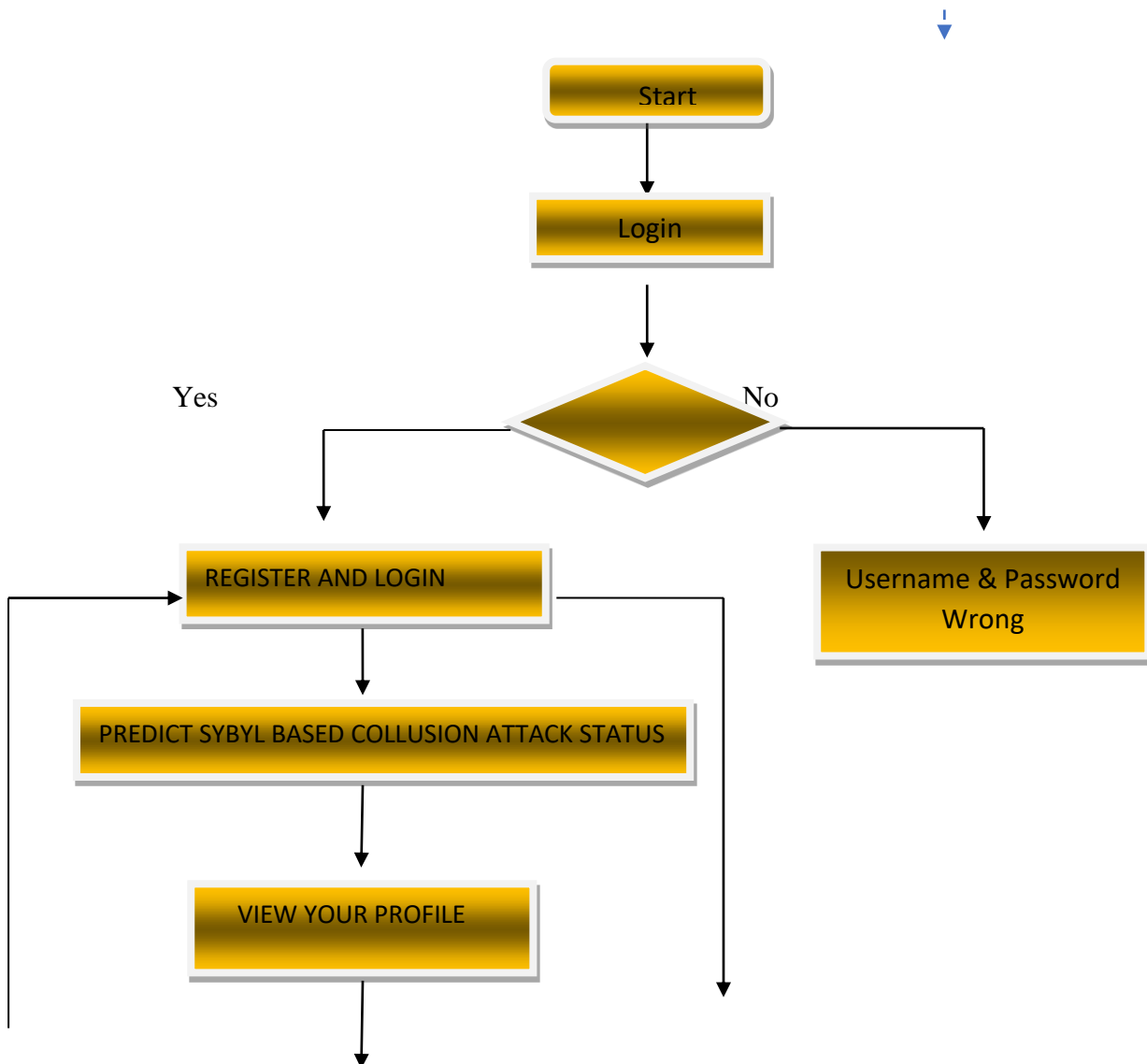
2. Bottom-up Integration

This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the need for stubs is eliminated. The bottom up integration strategy may be implemented with the following steps:

The low-level modules are combined into clusters into clusters that perform a specific Software sub-function. A driver (i.e.) the control program for testing is written to coordinate test case input and output. The cluster is tested.

Drivers are removed and clusters are combined moving upward in the program structure

Flow Chart : Remote User



[Logout](#)

5. CONCLUSION

This paper analyzed the security vulnerabilities of joint training in the IIOT-FL system, then proposed a sybil-based collusion attacks (SCA) approach for the vulnerabilities. Meanwhile, we also gave further details on the execution of related algorithms, model architecture, and analysis of the effectiveness of the experiment. In this work, malicious participants in our federated system can virtualize multiple Sybil nodes and perform malicious collusion attacks. The purpose is to make the local poisoning model be aggregated with a greater possibility. They aim to make the samples of the selected attack class be misclassified, while other non-attack classes maintain similar accuracy as before. Compared with the state-of-the-art, our SCA can achieve a more substantial attack effect under the condition of fewer malicious participants performing collusion, and can successfully obscure their attack behavior. Extensive experimental results show that our SCA has a more robust attack performance on several evaluation metrics.

6. REFERENCES

- [1] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato and H. V. Poor, "Federated learning for industrial internet of things in future industries," *IEEE Wireless communications magazine*, 2021.
- [2] P. Zhang, C. Wang, C. Jiang, and Z. Han. "Deep reinforcement learning assisted federated learning algorithm for data management of IIoT," *IEEE Transactions on Industrial Informatics (TII)*, 2021.
- [3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273-1282.
- [4] Y. Roh, G. Heo, and S. E. Whang, "A survey on data collection for machine learning: A big data - AI integration perspective," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 33, no. 4, pp. 1328-1347, 2021.

- [5] B. Jia, X. Zhang, J. Liu, Y. Zhang, K. Huang, and Y. Liang, "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT," *IEEE Transactions on Industrial Informatics (TII)*, 2021.
- [6] V. Mothukuri, R. M. Parizi, S. Pouriye, Y. Huang, A. Dehghantaha, and G. Srivastava, "A survey on security and privacy of federated learning," *ELSEVIER Future Generation Computer Systems (FGCS)*, vol. 115, pp. 619-640, 2021.
- [7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50-60, 2020.
- [8] W. S. Zhang, T. Zhou, Q. H. Lu, X. Wang, C. S. Zhu, H. Y. Sun, Z. P. Wang, S. K. Lo, and F. Y. Wang, "Dynamic fusion-based federated learning for COVID-19 detection," *IEEE Internet of Things Journal (IoTJ)*, 2021.
- [9] M. Parimala, M. S. Swarna, P. V. Quoc, D. Kapal, M. Praveen, T. Gadekallu, and T. H. Thien, "Fusion of federated learning and industrial internet of things: A survey," *arXiv preprint arXiv:2101.00798*, 2021.
- [10] M. X. Duan, K. L. Li, A. J. Ouyang, K. N. Win, K. Q. Li and Q. Tian,^b "EGroupNet: A feature-enhanced network for age estimation with novel age group schemes," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 16, no. 2, 2020.

FAKE IMAGE DETECTION USING MACHINE LEARNING

Kadimi Nani Vijatesh (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract: Online social networks (OSNs) have become an integral mode of communication among people and even nonhuman scenarios can also be integrated into OSNs. The ever growing rise in the popularity of OSNs can be attributed to the rapid growth of Internet technology. OSN becomes the easiest way to broadcast media (news/content) over the Internet. In the wake of emerging technologies, there is dire need to develop methodologies, which can minimize the spread of fake messages or rumors that can harm society in any manner. In this article, a model is proposed to investigate the propagation of such messages currently coined as fake news. The proposed model describes how misinformation gets disseminated among groups with the influence of different misinformation refuting measures. With the onset of the novel coronavirus-19 pandemic, dubbed COVID-19, the propagation of fake news related to the pandemic is higher than ever. In this article, we aim to develop a model that will be able to detect and eliminate fake news from OSNs and help ease some OSN users stress regarding the pandemic. A system of differential equations is used to formulate the model. Its stability and equilibrium are also thoroughly analyzed. The basic reproduction number (R_0) is obtained which is a significant parameter for the analysis of message spreading in the OSNs. If the value of R_0 is less than one ($R_0 < 1$), then fake message spreading in the online network will not be prominent, otherwise if $R_0 > 1$ the rumor will persist in the OSN. Real world trends of misinformation spreading in OSNs are discussed. In addition, the model discusses the controlling mechanism for untrusted message propagation. The proposed model has also been validated through extensive simulation and experimentation.

1. INTRODUCTION

In The 20th century, the Internet has become the most powerful tool for communication. It facilitates efficient and effective transfer of media from one location to another. With the development of Internet technology, social networks such as Facebook, WhatsApp, Twitter, Instagram, and Google plus have become a vital platform for information exchange [1]. Nowadays, people are connected through online social networks (OSNs) and exchange

information in a cost efficient manner through data transfer. However, information exchanged on OSN platforms may comprise rumors that may affect the social lives of people [2]. Take COVID-19 as an example, where the proliferation of fake news related to the virus has left many people skeptical of any information they read information related to the virus on social media [3]. Some recent fake news related to a cure for COVID 19 has spread through Facebook [4].

Due to this type of misinformation, people from different corners of the world died. The impact of fake news on people related to a well-known Zika virus case study was presented by Sommariva et al. [5]. The authors found that the speed of fake news spread on OSNs is tremendous and tends to cover large audiences. One major challenge that is associated with OSNs is verification of messages exchanged as well as the authenticity of users.

Some messages that are spread through these social networks may create horrible situations regarding peace and harmony in society. Such messages, currently coined as fake news, can also be life-threatening. These kinds of messages are in essence just rumors/misinformation which are propagated through different means [6], [7] either just for entertainment or maliciously as well. Due to such messages, unnecessary anxiety uprise among the public and countries may also face economic loss [8]–[11] as is seen currently with COVID-19 [12]. This can be attributed to the fact that the rate of information dissemination on OSNs is very quick and information can spread globally within seconds [13], [14]. Several instances exist where the spread of fake news on OSNs created undesirable and detrimental situations for society. For instance, two bombs exploded in the White House injuring the U.S. president (23 April 2013) and incurring a loss of 10 billion USD [15]. Another example from India can be of a rumour on OSN that claimed, “Sonam Gupta is unfaithful.”

Due to this message in social networks, the personal life of a random girl whose name is Sonam Gupta was affected. Such types of comments should not be accepted in a civilized society. This is a type of public shaming on OSNs and can lead to malicious consequences even if unintended. To overcome these types of issues, Basak et al. [16] suggested a mechanism of blocking/muting of shamer’s attacks on victims on Twitter. Liang et al. [17] investigated the rumor identification problem in microblogs. The authors proposed a method for identification of rumor rumormonger in the microblogs. Their scheme is based on the hidden behavior of users. More recently, the drug vaccine trial in the U.K. for COVID-19

was harmed when it was falsely reported that the first patient injected with the vaccine has died [18].

2. EXISTING SYSTEM

- ❖ The improved SIR model has been discussed by Zhang *et al.* [29] who considered the variable rate of infection and the resultant function for infected individuals and nonlinear Ordinary Differential Equation (ODE) is developed. This model also discusses the crowding effect on OSN and also derives an expression for the basic reproduction number. This model has been used for the analysis of rumor spreading dynamics in social network and predicts the spreading behavior of rumor. They discussed the control strategies of rumor spread in social networks.
- ❖ Zhu *et al.* [41] proposed an epidemic SIRS model, in which they described joining and leaving of users in OSNs. This article considers the dynamics of demography and the model is validated by simulation. More epidemic models are discussed related to rumors. Some of the researchers examined the temporal dynamics using the ODE [47]. Singh and Singh [48] discussed the spatial and temporal dynamics of rumor propagation and developed a strategy for countermeasures using. They used partial differential equation for the study of rumor propagation dynamics in the social network. Huang and Su [44] proposed an epidemic model for the study of news propagation on OSN and also suggested a method for controlling the rumor. They explained the effects of rumor spreading on OSN. For the study of rumor spreading in OSN, they evaluated the value of basic reproduction number and observed that if its value is less than one then the OSN will be free from unauthenticated news, otherwise unauthenticated news will be present in the OSN forever. The result of the proposed model has been verified by numerical calculation as well as simulation results.
- ❖ Dong *et al.* [49] analyzed the rumor spreading dynamics on OSN by SEIR epidemic model. They considered the varying user's number on OSN with time. The joining and deactivation rate of user in this model is discussed. They also found the basic reproduction number and exact equilibrium points of the model. The effect of user variation on rumor spreading in OSN is explained. They found that the new incoming users influence the rumor spreading rate in OSN. The proposed model is verified by simulation results.

- ❖ Furthermore, Zhu *et al.* [50] using the same model as in [49] obtained a local and global equilibrium as well as calculated the basic reproduction number using the next generation matrix concept. The authors explained the effect of time delay on rumor propagation and developed an effective control mechanism. A hesitating mechanism-based SEIR model is proposed by Liu *et al.* [51] for the study of rumor spreading in OSN. They used mean field theory for analysis of rumor spreading in OSN. They discussed the rumor-free equilibrium condition and global stability of the OSN and also obtained the value of basic reproduction number. They also analyzed the effects of feedback method on rumor spreading. They established the analysis feedback mechanism to reduce the rate of rumor spreading but were not able to reduce the value of basic reproduction number.

Disadvantages

- In the existing work, Identify when the user after the spreading rumor in the network.
- This system is less performance due to the standard susceptible-infected-recovered (SIR) model which is not used primarily to its generalization and efficacy.

3. PROPOSED SYSTEM

The key objectives of the proposed model are to monitor the presence of fake news/misinformation as well as spreaders in OSNs and apply a suitable corrective method for blocking and/or removal of these types misinformation and spreaders. Our contributions can be summarized as follows:

- 1) Formulate a mathematical model for monitoring fake news/misinformation as well as spreaders in OSNs and develop a method to prevent spreading of fake news;
- 2) Suggest the concept of verification through verified state for verification of users in OSNs;
- 3) analyze the effect of a verified state on a given OSN's responsiveness and investigate its role in the prevention of fake news spreading in OSNs;
- 4) analyze the effectiveness of a recovered state (blocking/ removing/leaving of a spreader group) on fake news as well as a spreader in OSNs;
- 5) Investigate social network stability under various conditions and verify theoretical findings through extensive simulation results.

Advantages

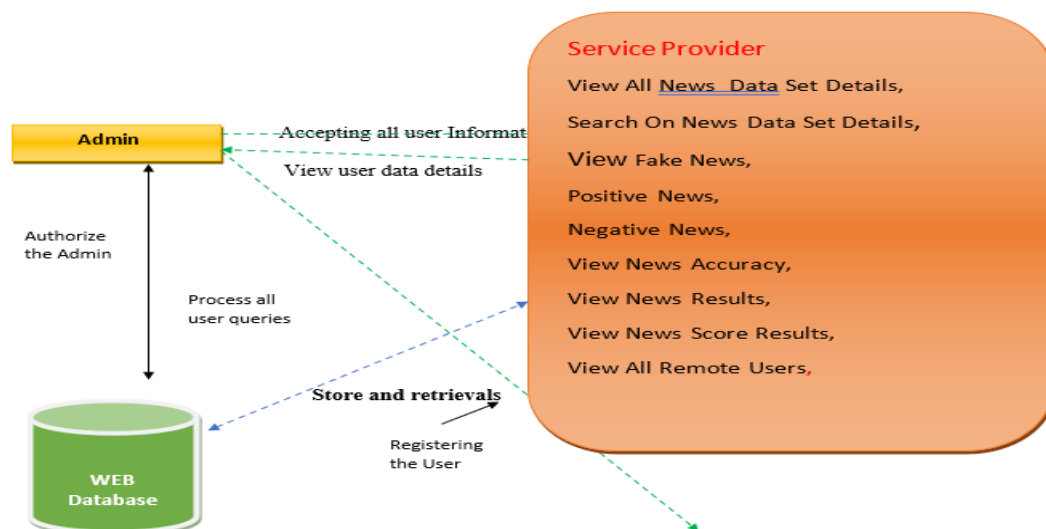
- ❖ For detection and controlling of misinformation (rumor) in OSN, a susceptible-verified-infected-recovered (SVIR) model is proposed which is more effective.
- ❖ The system is more effective due to presence of the mechanisms for the removal of rumors (an “infection of the mind”) has been used.

4. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

Architecture Diagram



5. CONCLUSIONS

The research work presented in this article proposes a mathematical model to study the dynamic spreading and controlling activities of message transmission in OSNs. The proposed model employs differential equations for investigating the effect of verification and blocking of users and the spread of messages on OSNs. The expression for basic reproduction R_0 is

obtained, which is used to analyze the status of rumor in the social network. Results obtained indicates that if R_0 is less than 1, then rumors and fake news will be eliminated and OSNs gets stabilized locally. The local stability of rumor free equilibrium is established by the Jacobian matrix. It is found that if the eigen values of the matrix are less than zero then the network will be asymptotically stabilize locally in nature and free from the rumors. The Lyapunov function used to establish the global asymptotic stable status of the social network. Mathematical analysis has been performed to depict the accuracy of the rumor-free equilibrium. The activities of different classes of users have also been examined in the social network. In future, the method of latent and isolation can be used for the prevention of social network from rumor spread and fake news propagation. The issues examined in this article are of direct current concern, and the pandemic COVID-19 is creating a global crisis in rumors and fake news propagating freely on OSNs which may continue until it is cured/handled. Real world data clearly show that fake news propagation can be harmful for people, businesses, and many other facets of society. The results in this article therefore, may help solve some of the current global issues related to fake news spread.

6. REFERENCES

- [1] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modeling propagation dynamics of social network worms," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 8, pp. 1633–1643, Aug. 2013.
- [2] E. Lebensztayn, F. P. Machado, and P. M. Rodríguez, "On the behaviour of a rumour process with random stifling," *Environ. Model. Softw.*, vol. 26, no. 4, pp. 517–522, Apr. 2011.
- [3] L. Li et al., "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on weibo," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020.
- [4] A. Legon and A. Alsalman. How Facebook Can Flatten the Curve of the Coronavirus Infodemic. Accessed: Apr. 20, 2020. [Online]. Available: https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/
- [5] S. Sommariva, C. Vamos, A. Mantzarlis, L. U.-L. Dào, and D. Martinez Tyson, "Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study," *Amer. J. Health Edu.*, vol. 49, no. 4, pp. 246–255, Jul. 2018, doi: 10.1080/19325037.2018.1473178.

- [6] G. Whitehouse, "Pete/Repeat Tweet/Retweet Blog/reblog: A hoax reveals media mimicking," *J. Mass Media Ethics*, vol. 27, no. 1, pp. 57–59, Jan. 2012.
- [7] M. Kosfeld, "Rumours and markets," *J. Math. Econ.*, vol. 41, no. 6, pp. 646–664, Sep. 2005.
- [8] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, "Rumor propagation dynamic model based on evolutionary game and antirumor," *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, Jan. 2019.
- [9] A. V. Banerjee, "The economics of rumours," *Rev. Econ. Stud.*, vol. 60, no. 2, pp. 309–327, Apr. 1993.
- [10] K. Dietz, "Epidemics and rumours: A survey," *J. Roy. Stat. Soc., A (Gen.)*, vol. 130, no. 4, pp. 505–528, 1967.



E-PILOT A SYSTEM TO PREDICT HARD LANDING DURING THE APPROACH PHASE OF COMMERCIAL FLIGHTS

Kalaga Pavan Venkata Sai Ravi Teja (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract:

More than half of all commercial aircraft operation accidents could have been prevented by executing a go-around. Making timely decision to execute a go-around manoeuvre can potentially reduce overall aviation industry accident rate. In this paper, we describe a cockpit-deployable machine learning system to support flight crew go-around decision-making based on the prediction of a hard landing event. This work presents a hybrid approach for hard landing prediction that uses features modelling temporal dependencies of aircraft variables as inputs to a neural network. Based on a large dataset of 58177 commercial flights, the results show that our approach has 85% of average sensitivity with 74% of average specificity at the go-around point. It follows that our approach is a cockpit-deployable recommendation system that outperforms existing approaches.

1. INTRODUCTION

Between 2008-2017, 49% of fatal accidents involving commercial jet worldwide occurred during final approach and landing, and this statistic has not changed in several decades [1]. A considerable proportion of approach and landing accidents/incidents involved runway excursions, which has been identified as one of the top safety concerns shared by European Union Aviation Safety Agency (EASA) member states [2], as well as US National Transportation Safety Board and US Federal Aviation Administration [3].

According to EASA [2], there are several known precursors to runway excursions during landing. These include unstable approach, hard landing, abnormal attitude or bounce at landing, aircraft lateral deviations at high speed on the ground, and short rolling distance at landing. Some precursors can occur in isolation, but they

can also cause the other precursors, with unstable approach being the predominant one. Boeing reported that whilst only 3% of approaches in commercial aircraft operation met the criteria of an unstable approach, 97% of them continued to landing rather than executing a go-around [4]. A study conducted by Blajev and Curtis [5] found that 83% of runway excursion accidents in their 16-year analysis period could have been avoided by a go-around decision. Therefore, making timely decision to execute a go-around manoeuvre could therefore potentially reduce the overall aviation industry accident rate [4].

A go-around occurs when the flight crew makes the decision not to continue an approach or a landing, and follows procedures to conduct another approach or to divert to another airport. Go-around decision can be made by either flight crew members, and can be executed at any point



from the final approach fix point to wheels touching down on the runway (but prior to activation of brakes, spoilers, or thrust reversers). In addition to unstable approaches, traffic, blocked runway, or adverse weather conditions are other reasons for a go-around. Despite a clear policy and training on go-around policies in most airlines, operational data show that flight crew decision-making process in deciding for a go-around could be influenced by many other factors. These include fatigue, flight schedule pressure, time pressure, excessive a head-down work, incorrect anticipation of aircraft deceleration, visual illusions, organizational policy/culture, inadequate training or practice, excessive confidence in the ability to stabilize approach, and Crew Resource Management issues [5]. It is for these reasons that on-board real-time performance monitoring and alerting systems that could assist the flight crew with the landing/go-around decision are needed [5], [6].

Such on-board systems could utilize the huge and ever-increasing amount of data collected from aircraft systems and the exponential advances in machine learning methods and artificial intelligence. EASA is anticipating a huge impact of machine learning on aviation, including helping the crew to take decisions in particular in high workload circumstances (e.g. go-around, or diversion [7]). Artificial Intelligence in aviation is considered one of the strategic priorities in the European Plan for Aviation Safety 2020–2024 [8].

Under the hypothesis that a hard-landing (HL) occurrence has precursors and, thus, it can be predicted, this paper presents a

cockpit deployable machine learning system to predict hard landings considering the aircraft dynamics and configuration. In particular, this paper evaluates three main hypothesis. A primary hypothesis is to assess to what extent HL can be predicted at DH for go-around recommendation from the analysis of the variables recorded from FMS. A second hypothesis is to analyze if precursors are particular to aircraft types. A third hypothesis is to validate if the variability on the aircraft state variables can provide enough information to predict a HL regardless of the operational context (like environmental conditions and automation factors).

2. INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?



- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be

displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

3. SYSTEM DESIGN

UML DIAGRAMS:

UML represents Unified Modeling Language. UML is an institutionalized universally useful showing dialect in the subject of article situated programming

designing. The fashionable is overseen, and become made by way of, the Object Management Group.

The goal is for UML to become a regular dialect for making fashions of item arranged PC programming. In its gift frame UML is contained two noteworthy components: a Meta-show and documentation. Later on, a few type of method or system can also likewise be brought to; or related with, UML.

The Unified Modeling Language is a popular dialect for indicating, Visualization, Constructing and archiving the curios of programming framework, and for business demonstrating and different non-programming frameworks.

The UML speaks to an accumulation of first-rate building practices which have verified fruitful in the showing of full-size and complicated frameworks.

The UML is a essential piece of creating gadgets located programming and the product development method. The UML makes use of commonly graphical documentations to specific the plan of programming ventures.

GOALS:

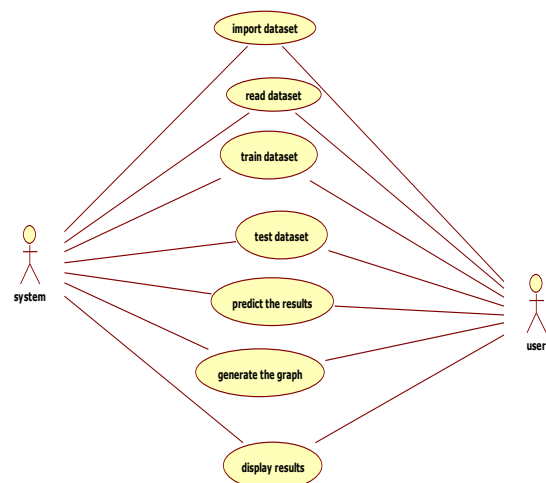
The Primary goals inside the plan of the UML are as in step with the subsequent:

1. Provide clients a prepared to-utilize, expressive visual showing Language on the way to create and change massive models.
2. Provide extendibility and specialization units to make bigger the middle ideas.
3. Be free of specific programming dialects and advancement manner.

4. Provide a proper cause for understanding the displaying dialect.
5. Encourage the improvement of OO gadgets exhibit.
6. Support large amount advancement thoughts, for example, joint efforts, systems, examples and components.
7. Integrate widespread procedures.

USE CASE DIAGRAM:

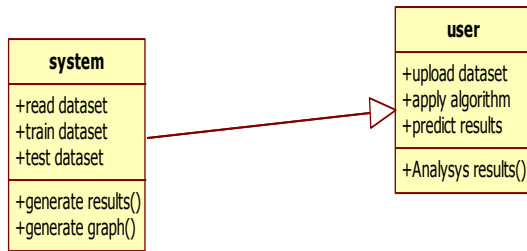
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



CLASS DIAGRAM:

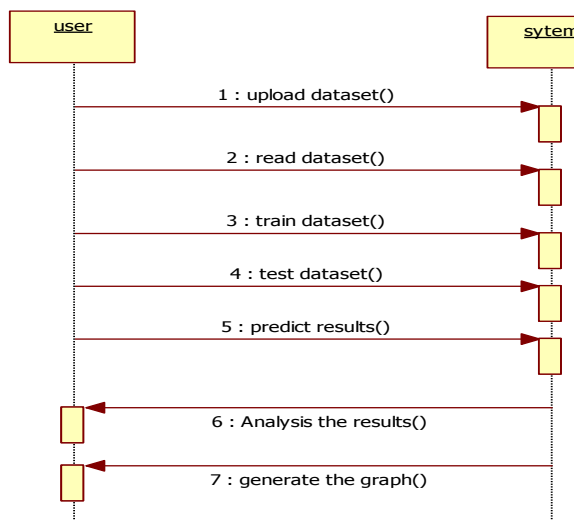
In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the

structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



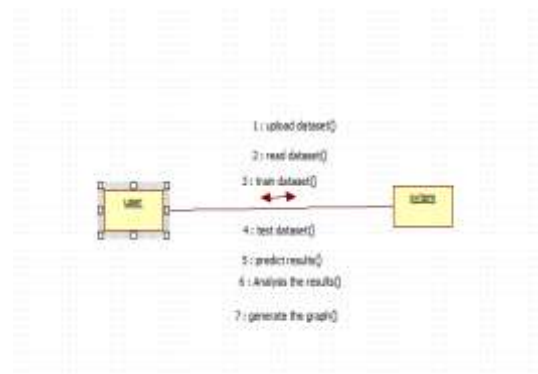
SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.



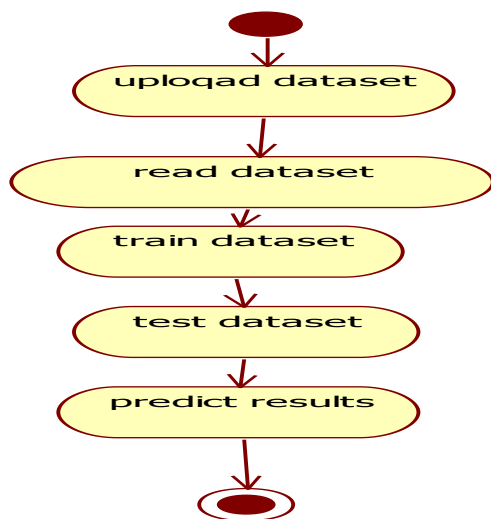
COLLABORATION DIAGRAM:

In collaboration diagram the method call sequence is indicated by some numbering technique as shown below. The number indicates how the methods are called one after another. We have taken the same order management system to describe the collaboration diagram. The method calls are similar to that of a sequence diagram. But the difference is that the sequence diagram does not describe the object organization where as the collaboration diagram shows the object organization.



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



DEPLOYMENT DIAGRAM:

Deployment diagram represents the deployment view of a system. It is related to the component diagram. Because the components are deployed using the deployment diagrams. A deployment diagram consists of nodes. Nodes are nothing but physical hardware's used to deploy the application.



4. CONCLUSION

The following conclusions can be extracted from the analysis carried out in this paper.

The analysis of automation factors (autopilot, flight director and auto-thrust) suggests that these factors do not have any influence on the probability of a HL event and, thus, it might not be necessary to incorporate them into models.

Experiments for the optimization of architectures show that the configurations

that achieve higher sensitivity are the ones with the lowest number of neurons. As reported in the literature [23] increasing the number of layers and neurons does not improve the performance of neither classifiers nor regressors.

Models using only Physical variables achieve an average recall of 94% with a specificity of 86% and outperform state-of-the-art LSTM methods. This brings confidence into the model for early prediction of HL in a cockpit deployable system. Regarding capability for go-around recommendation before DH, even if we perform better than existing methods, there is a significant drop in recall and specificity due to the dynamic nature of a landing approach and factors influencing HL close to TD.

Comparing classifiers and regression approaches, experiments show that a low MSE error in estimation of maxG does not guarantee accurate HL predictions. Experiments for assessing the capability of models for early detection of HL show that classifiers are able to accurately predict HL before DH. This is not the case of regressors, which predict maxG more accurately if data close to TD is considered into the model. The study suggests that classifiers are a better approach for early prediction of hard landing.

Neural networks performance could be increased if they were used to extract deep learning features from continuous signals by using one dimensional convolutional networks and different architectures for a better combination of the three categories of variables. Also, models should incorporate additional parameters such as aircraft mass and centre of gravity position



which are known to impact vehicle dynamics.

Finally, there are some issues that have not been covered in this work, that remain as future work, and should be further developed. Among such cases, stand out the robustness of the classifier (regressor) to unseen cases and its behavior under a drifting data environment. In a safety demanding environment as aviation, it surely be needed to investigate such issues and we expect to do in further works. In the future, such a system could be expanded to also include Air Traffic Management in which the information is shared with the Air Traffic Controller in order to anticipate the likely scenario and optimize runway use.

5. REFERENCES

1. Statistical Summary of Commercial Jet Airplane Accidents–Worldwide Operations|1959–2017, Seattle, WA, USA, 2018.
2. Developing standardised FDM-based indicators, Cologne, Germany, 2016.
3. Advisory circular ac no: 91-79a mitigating the risks of a runway overrun upon landing, Washington, DC, USA, 2016.
- 4.M. Coker and L. S. Pilot, "Why and when to perform a go-around maneuver", *Boeing Edge*, vol. 2014, pp. 5-11, 2014.
- 5.T. Blajev and W. Curtis, *Go-around decision making and execution project: Final report to flight safety foundation*, Mar. 2017.
- 6.European action plan for the prevention of runway excursions, Brussels, Belgium, 2013.
- 7.Artificial intelligence roadmap—A human-centric approach to ai in aviation, Cologne, Germany, 2020.
- 8.The European plan for aviation safety (EPAS 2020–2024), Cologne, Germany, 2019.
- 9.L. Wang, C. Wu and R. Sun, "Pilot operating characteristics analysis of long landing based on flight QAR data", *Proc. Int. Conf. Eng. Psychol. Cognit. Ergonom.*, pp. 157-166, 2013.
10. L. Li, J. Hansman, R. Palacios and R. Welsch, "Anomaly detection via a Gaussian mixture model for flight operation and safety monitoring", *Transp. Res. C Emerg. Technol.*, vol. 64, pp. 45-57, Mar. 2016.

HOTEL REVIEW ANALYSIS FOR THE PREDICTION OF BUSINESS USING DEEP LEARNING APPROACH

Kalidindi Satya Srikanth Vamsi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:- Sentiment analysis is a widely used topic in Natural Language Processing that allows identifying the opinions or sentiments from a given text. Social media is the scope for the customers to share their opinion over the products or services as part of customer reviews. Dissect this review has become an important factor for business analysis since online business is exponentially growing in today's techno-friendly competitive market. A large number of algorithms have been found in recent articles. Among those deep learning is an important approach. In the proposed methodology, long short-term memory (LSTM) and Gated recurrent units (GRUs) have been used to train the hotel review data where the accuracy rate of identifying customer opinion is 86%, and 84% respectively. The dataset is also tested by using Naïve Bayes, Decision Tree, Random Forest, and SVM. For Naïve Bayes obtains an accuracy of 75%, for Decision Tree obtains an accuracy of 71%, for Random Forest the accuracy is 82% and for SVM our accuracy result is 71%. Deep learning is used to obtain better business performance and also get the review from customers and also to predict the sentiment about customer review. Our algorithm works properly and gives better accuracy.

Keywords— Natural Language Processing, Machine Learning, Deep Learning, Artificial Intelligent, LSTM, GRU

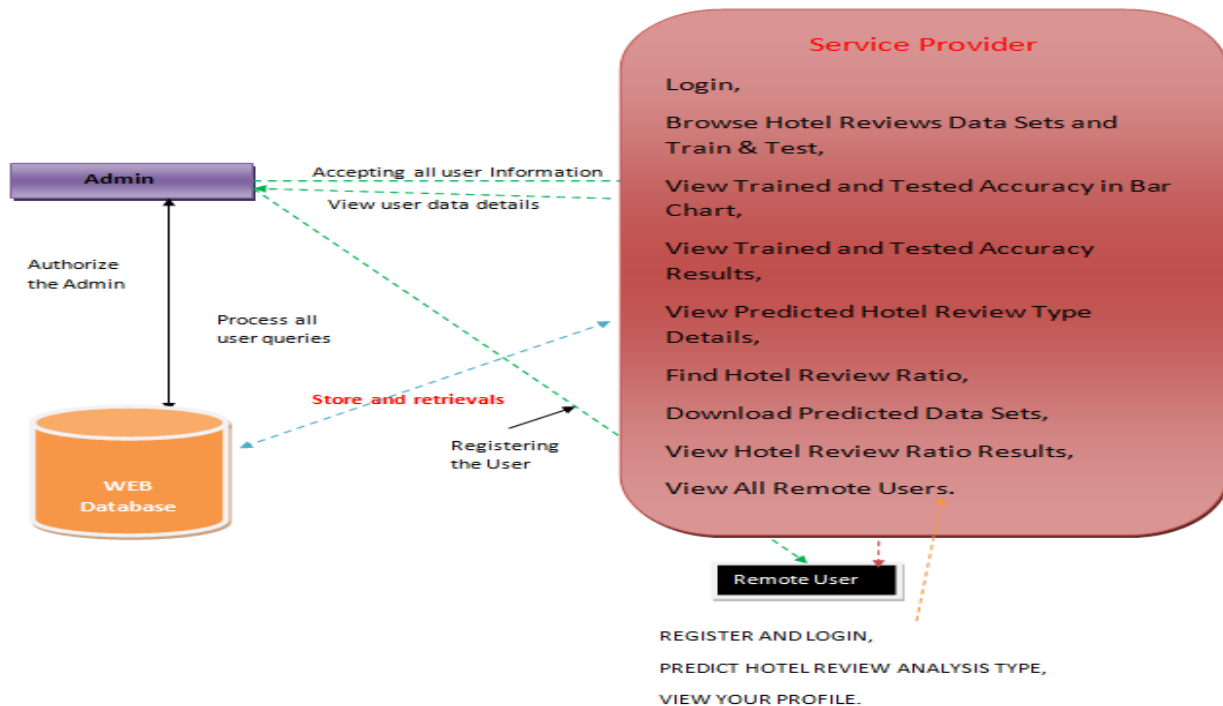
1.INTRODUCTION

In the age of modern science, everything is based on online and on the internet. Internet-based shopping has become easier and more popular because of better quality, and fast logistic systems. Internet-based shopping and booking are very comfortable. People can easily make a booking without going outside. The most effective side part of online-based work is that people can give a review. Recognizing reviews allows others to easily understand the emotions of others and obtain the rationality result of different products [10].

In the hotel review, the prediction of business using Deep Learning was analyzed [24]. Many start-up businesses became failure due to lack of analysis and the sentiment of the customer. Sentiment Analysis is the most significant to improve a business site. Here, different type of data from social media as well as from the Hotel Management Website was collected using Unamo tools. And also some supervised and unsupervised data is used to predict the best result. This article will help to improve the business.

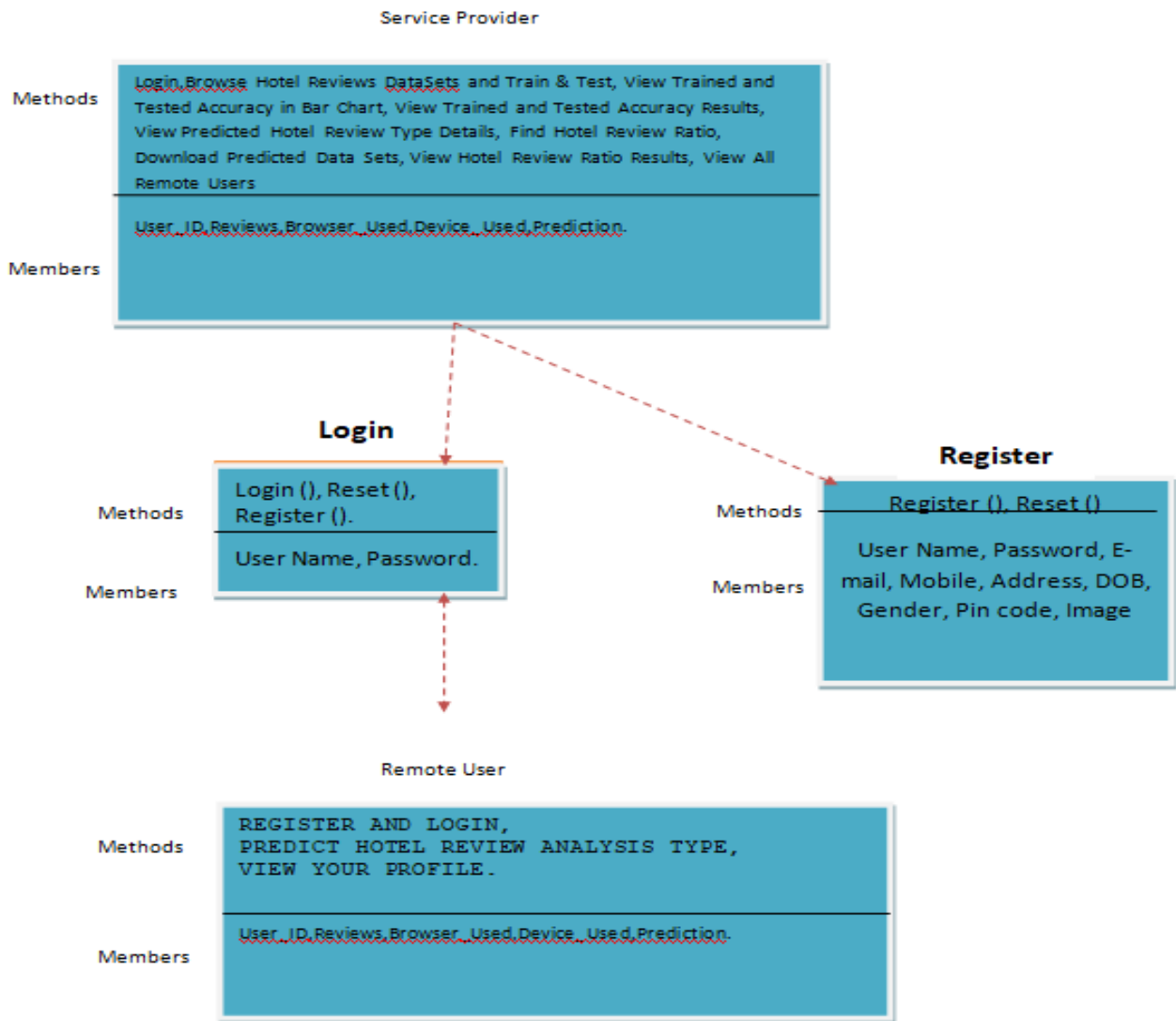
At present, online-based opinions can easily analysis with the help of Sentiment Analysis (SA). It is the management of sentiments, different opinions, subjective text, and different emoji used for giving reviews. People can easily get the comprehension information related to people reviews. Mainly Sentiment analysis is one kind of tool that helps to get the public sentiment. By capturing reviews of product or location or person might be found from a different internet-based site like Face book, Amazon. Sentiment Analysis is used to increase the requirement of analyzing and structuring hidden information which comes from social media in the form of unstructured data. A huge amount of data is used due to the capability of automation and can handle a huge amount of data. A different type of font [23] of review are further classified.

Architecture Diagram

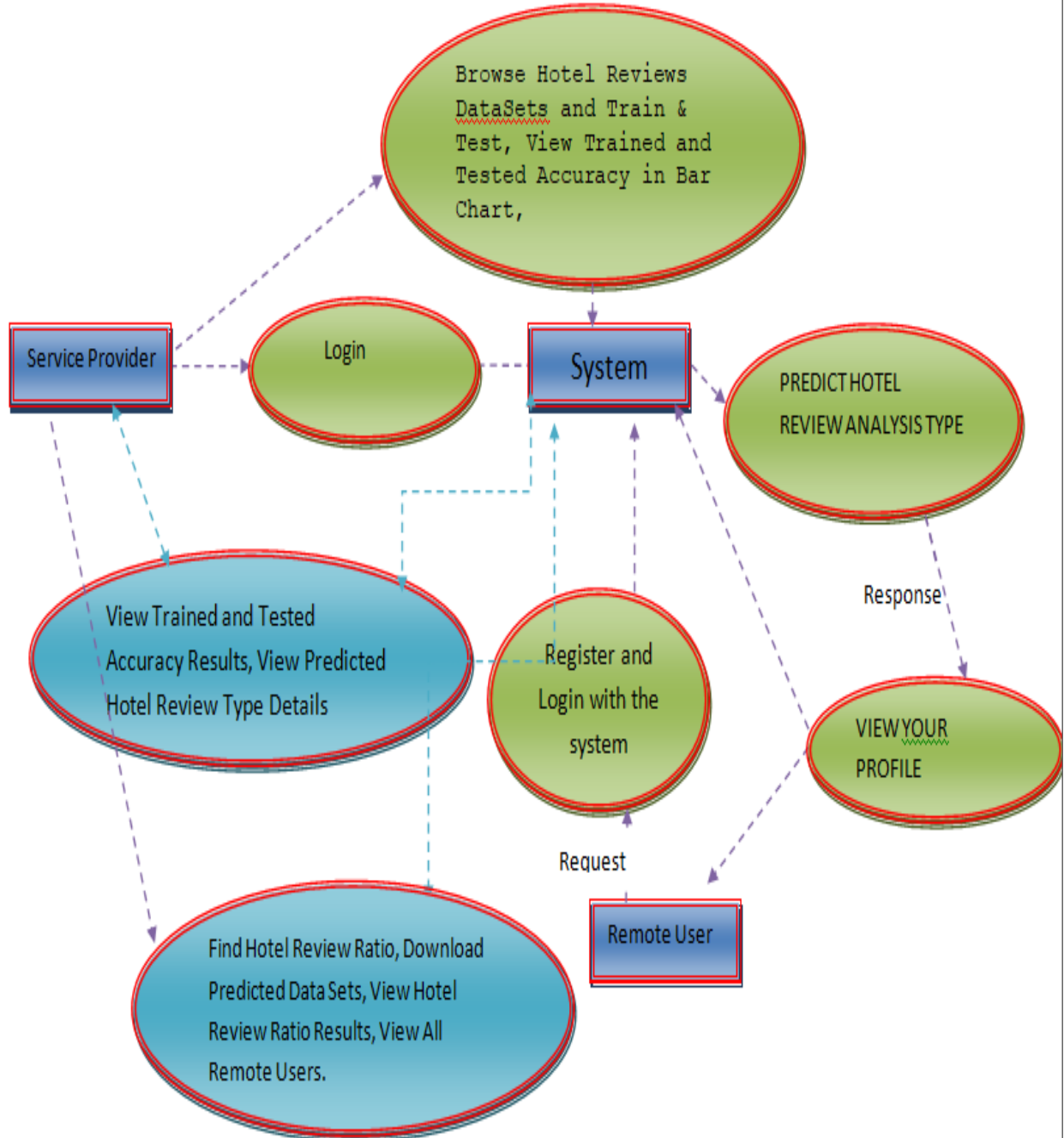


Algorithms Used(Trained and Tested With)
Naive Bayes, SVM, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, SGD Classifier, KNeighborsClassifier

Class Diagram



Data Flow Diagram



2.SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

3.SYSTEM TESTING

Software once validated must be combined with other system elements (e.g. Hardware, people, database). System testing verifies that all the elements are proper and that overall system function performance is achieved. It also tests to find discrepancies between the system and its original objective, current specifications and system documentation.

UNIT TESTING

In unit testing different are modules are tested against the specifications produced during the design for the modules. Unit testing is essential for verification of the code produced during the coding phase, and hence the goals to test the internal logic of the modules. Using the detailed design description as a guide, important Conrail paths are tested to uncover errors within the boundary of the modules. This testing is carried out during the programming stage itself. In this type of testing step, each module was found to be working satisfactorily as regards to the expected output from the module.

In Due Course, latest technology advancements will be taken into consideration. As part of technical build-up many components of the networking system will be generic in nature so that future projects can either use or interact with this. The future holds a lot to offer to the development and refinement of this project.

4.CONCLUSIONS

The present age is the modern age. Everything in the age is now technology dependent and every person in the country is able to familiarize themselves with this technology. With the help of that technology, online marketing has become popular in today's world, which has easily become

popular among people. People are now getting a lot of things through their hands very easily. One part of online marketing is the online hotel booking system. With this people can easily prebook the hotel of their choice and they can easily go to their hotel without having to bother to search for the place. It has become the most popular among people and this led to an increase in the number of people traveling around. And at the same time, they can able to view different beautiful places of the world by taking advantage of this hotel booking. In the future, many more features can be added to the project and ensure more popular things.

5. REFERENCES

- [1] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 452-455.
- [2] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in *IEEE Access*, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
- [3] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 700-704, doi: 10.1109/ICCS45141.2019.9065812.
- [4] Z. Singla, S. Randhawa and S. Jain, "Statistical and sentiment analysis of consumer product reviews," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi:10.1109/ICCCNT.2017.8203960.
- [5] C. Nanda, M. Dua and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 1069-1072, doi:10.1109/ICCSP.2018.8524223.
- [6] B. Seetharamulu, B. N. K. Reddy and K. B. Naidu, "Deep Learning for Sentiment Analysis Based on Customer Reviews," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi:10.1109/ICCCNT49239.2020.9225665.
- [7] Rahul, V. Raj and Monika, "Sentiment Analysis on Product Reviews," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2019, pp. 5-9, doi: 10.1109/ICCCIS48478.2019.8974527.

- [8] Y. Saito and V. Klyuev, "Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019, pp. 681-685, doi: 10.23919/ICACT.2019.8702039.
- [9] A. Salinca, "Business Reviews Classification Using Sentiment Analysis," 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, 2015, pp. 247-250, doi: 10.1109/SYNASC.2015.46.
- [10] Chhaya Chauhan, Smriti Sehgal "SENTIMENT ANALYSIS ON PRODUCT REVIEWS", International Conference on Computing, Communication and Automation (ICCCA2017) ISBN:978-1-5090-6471-7/17/\$31.00 ©2017 IEEE

QOS RECOMMENDATION IN CLOUD SERVICES

Kanakala Chandra Vamsi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

As cloud computing becomes increasingly popular, cloud providers compete to offer the same or similar services over the Internet. Quality of service (QoS), which describes how well a service is performed, is an important differentiator among functionally equivalent services. It can help a firm to satisfy and win its customers. As a result, how to assist cloud providers to promote their services and cloud consumers to identify services that meet their QoS requirements becomes an important problem. In this paper, we argue for QoS-based cloud service recommendation, and propose a collaborative filtering approach using the Spearman coefficient to recommend cloud services. The approach is used to predict both QoS ratings and rankings for cloud services. To evaluate the effectiveness of the approach, we conduct extensive simulations. Results show that the approach can achieve more reliable rankings, yet less accurate ratings, than a collaborative filtering approach using the Pearson coefficient.

1. INTRODUCTION

Cloud computing refers to a large pool of virtualized resources that can be dynamically reconfigured to provide elastic services over the Internet [14]. It has the potential to increase business agility, improve efficiencies, and reduce costs. As cloud computing becomes increasingly popular, cloud providers, including leading IT companies like Amazon, Google, and Microsoft, compete to offer the same or similar services over the Internet.

As an example, Amazon Simple Storage Service (Amazon S3) offers durable and massively scalable object storage. Google Cloud Storage provides durable and highly available object storage. Microsoft Azure Storage provides reliable and economical storage for small and big data. Indeed, there are more than a dozen cloud providers offering online storage services, and the number is still growing.

As the cloud market becomes more open and competitive, quality will be more important. According to the American Society for Quality, quality is "the totality of features and

characteristics of a product or service that bears on its ability to satisfy stated or implied needs." [5]. It can help companies to obtain a competitive advantage by improving business operations, building good reputation, reducing product liability, and competing effectively in the global economy.

2. EXISTING SYSTEM

In the Infrastructure as a Service (IaaS) paradigm of cloud computing, computational resources are available for rent. Although it offers a cost efficient solution to virtual network requirements, low trust on the rented computational resources prevents users from using it. To reduce the cost, computational resources are shared, i.e., there exists multi-tenancy. As the communication channels and other computational resources are shared, it creates security and privacy issues. A user may not identify a trustworthy co-tenant as the users are anonymous. The user depends on the Cloud Provider (CP) to assign trustworthy co-tenants. But, it is in the CP's interest that it gets maximum utilization of its resources. Hence, it allows maximum co-tenancy irrespective of the behaviours of users. In this paper, we propose a robust reputation management mechanism that encourages the CPs in a federated cloud to differentiate between good and malicious users and assign resources in such a way that they do not share resources. We show the correctness and the efficiency of the proposed reputation management system using analytical and experimental analysis.

3. PROPOSED SYSTEM

As the cloud market becomes more open and competitive, quality will be more important. According to the American Society for Quality, quality is "the totality of features and characteristics of a product or service that bears on its ability to satisfy stated or implied needs. It can help companies to obtain a competitive advantage by improving business operations, building good reputation, reducing product liability, and competing effectively in the global economy. In cloud computing, Quality of Service (QoS) is non-functional properties of cloud services, which describe how well a service is performed, such as availability, reliability, responsiveness, and security. Indeed, QoS is an important differentiator among functionally equivalent services. It can help arm to satisfy and win its customers. As a result, how to assist cloud providers to promote their services and cloud consumers to identify services that meet their QoS requirements becomes an important problem. Recommender systems, which have been developed to alleviate the information overload problem, can help users to and useful information and products. They can generate

suggestions that match users' interests and preferences. Recommender systems are personalized information altering techniques, which are employed to either predict whether a user will like an item (prediction problem) or find a set of items that will be of interest to a user (top- N recommendation problem).

spearman approach coefficient:

we argue for qos based cloud service recommendation and propose collaborative filtering approach using spearman approach coefficient

pearson coefficient:

- can achieve more reliable ranking
- less accurate rating
- collaborative filtering

Collaborative filtering approach:

we studied electronic e-commerce many years

it recommendation to user based on the opinion of set of user sharing the same and similar interest

achievement of using collaborative filtering approach:

- quality of item
- recommend the best item to user

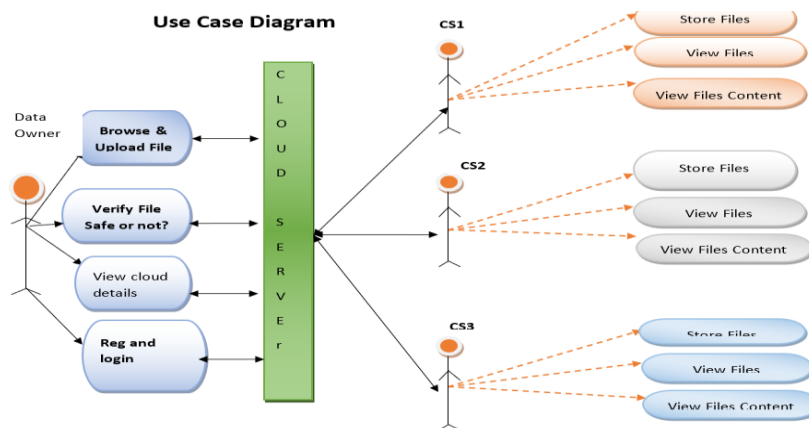
4. LITERATURE SURVEY

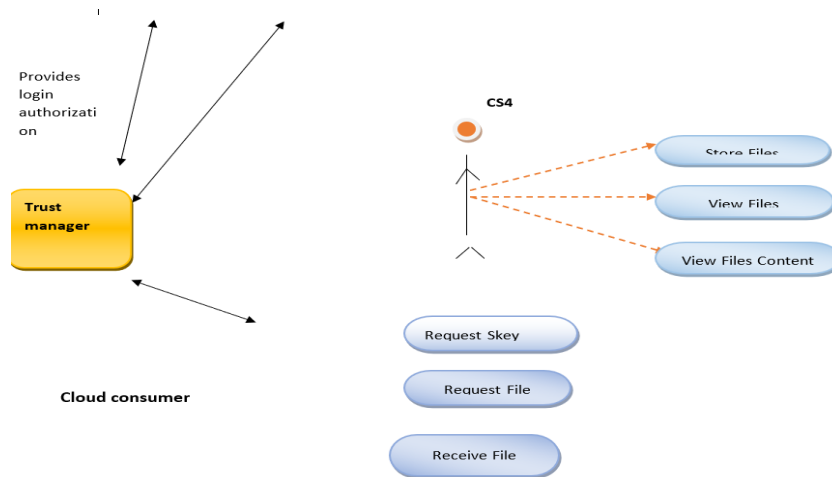
Privacy-Preserving Fine-Grained Access Control in Public Clouds

With many economical benefits of cloud computing, many organizations have been considering moving their information systems to the cloud. However, an important problem in public clouds is how to selectively share data based on fine-grained attribute based access control policies while at the same time assuring confidentiality of the data and preserving the privacy of users from the cloud. In this article, we briefly discuss the drawbacks of approaches based on well known cryptographic techniques in addressing such problem and then present two approaches that address these drawbacks with different trade-offs.

Universally Composable Multiparty Computation with Partially Isolated Parties

It is well known that universally composable multiparty computation cannot, in general, be achieved in the standard model without setup assumptions when the adversary can corrupt an arbitrary number of players. One way to get around this problem is by having a trusted third party generate some global setup such as a common reference string (CRS) or a public key infrastructure (PKI). The recent work of Katz shows that we may instead rely on physical assumptions, and in particular tamper-proof hardware tokens. In this paper, we consider a similar but strictly weaker physical assumption. We assume that a player (Alice) can partially isolate another player (Bob) for a brief portion of the computation and prevent Bob from communicating more than some limited number of bits with the environment. For example, isolation might be achieved by asking Bob to put his functionality on a tamper-proof hardware token and assuming that Alice can prevent this token from communicating to the outside world. Alternatively, Alice may interact with Bob directly but in a special o_{ce} which she administers and where there are no high-bandwidth communication channels to the outside world. We show that, under standard cryptographic assumptions, such physical setup can be used to UC-realize any two party and multiparty computation in the presence of an active and adaptive adversary corrupting any number of players. We also consider an alternative scenario, in which there are some trusted third parties but no single such party is trusted by all of the players. This compromise allows us to significantly limit the use of the physical set-up and hence might be preferred in practice





5. SYSTEM TESTING

TESTING METHODOLOGIES

Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

Functional test

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes

6. CONCLUSION

As cloud computing becomes popular, the same or similar services are delivered over the Internet. QoS is an important differentiator among functionally equivalent services. In this paper, recommender systems are employed to assist cloud providers to promote their services and cloud consumers to identify services that meet their QoS requirements.

Collaborative filtering is the most successful and widely used technique to build recommender systems. In the paper, we argue for QoS-based cloud service recommendation, and propose a ranking-based CF approach using the Spearman coefficient. The approach can predict both ratings and rankings for cloud services. To demonstrate the effectiveness of the approach, we conduct extensive simulations, and compare the approach with a rating-based CF approach using the Pearson coefficient. Results show that the CF approach using the Spearman coefficient can achieve more reliable rankings, yet less accurate ratings, than the CF approach using the Pearson coefficient.

To achieve better performance, we plan to use a mixed approach in our next step. In other words, we first use the CF approach using the Spearman coefficient to predict rankings, and then use the CF approach using the Pearson coefficient to predict ratings. In this way, the mixed approach could achieve more accurate ratings, while still obtaining reliable rankings. In addition, we plan to compare the CF approach using the Spearman coefficient with other ranking-based approaches in our future work.

7. REFERENCES

- [1] F. CACHEDA, V. CARNEIRO, D. FERNANDEZ, and V. FORMOSO, "Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems," *ACM Trans. Web*, vol. 5, no. 1, p. 2, Feb. 2011.
- [2] C. A. GOMEZ-URIBE and N. HUNT, "The netix recommender system: Algorithms, business value, and innovation," *ACM Trans. Manage. Inf. Syst.*, vol. 6, no. 4, p. 13, Jan. 2015.
- [3] M. DESHPANDE and G. KARYPIS, "Item-based top-n recommendation algorithms," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 143177, Jan. 2004.
- [4] D. GONZALES, J. KAPLAN, E. SALTZMAN, Z. WINKELMAN, and D. WOODS, "Cloud-trustA security assessment model for infrastructure as a service (IaaS) clouds," *IEEE Trans. Cloud Comput.*, to be published, doi: 10.1109/TCC.2015.2415794.
- [5] J. HEIZER and B. RENDER, *Operations Management*, 7th ed. Upper Saddle River, NJ, USA: Pearson, 2004.

- [6] K.Hwang, G. C. Fox, and J. J. Dongarra, Distributed and Cloud Computing From Parallel Computing to the Internet of Things, 1st ed. Waltham, MA, USA: Morgan Kaufmann, 2012.
- [7] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender Systems: An Introduction, 1st ed. New York, NY, USA: Cambridge Univ. Press, 2010.
- [8] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Comput. Surv.*, vol. 47, no. 1, p. 3, Jul. 2014.



PREDICTING USED CAR PRICE TYPE

Kandi Naga Venkata Sriram (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The number of cars on Mauritian roads has been rising consistently by 5% during the last decade. In 2014, 173 954 cars were registered at the National Transport Authority. Thus, one Mauritian in every six owns a car, most of which are second hand reconditioned cars and used cars. The aim of this study is to assess whether it is possible to predict the price of second-hand cars using artificial neural networks. Thus, data for 200 cars from different sources was gathered and fed to four different machine learning algorithms. We found that support vector machine regression produced slightly better results than using a neural network or linear regression. However, some of the predicted values are quite far away from the

actual prices, especially for higher priced cars. Thus, more investigations with a larger data set are required and more experimentation with different network type and structures is still required in order to obtain better predictions.

1. INTRODUCTION

According to the data obtained from the National Transport Authority (2014), there has been an increase of 254% in the number of cars from 2003 (68, 524) to 2014 (173, 954), as shown in Figure 1. We can thus infer that the sale of second-hand imported (reconditioned) cars and second-hand used cars has eventually increase given that new cars represent only a very small percentage of the total number of cars sold each year. Most individuals in Mauritius who buy new cars also want to know about the resale value of their cars after some years so that they can sell it in the used car market.

Price prediction of second-hand cars depends on numerous factors. The most important ones are manufacturing year, make, model, mileage, horsepower and country of origin. Some other factors are type and amount of fuel per usage, the type of braking system, its acceleration, the interior style, its physical state, volume of

cylinders (measured in cubic centimeters), size of the car, number of doors, weight of the car, consumer reviews, paint colour and type, transmission type, whether it is a sports car, sound system, cosmic wheels, power steering, air conditioner, GPS navigator, safety index etc. In the Mauritian context, there are some special factors that are also usually considered such as who were the previous owners and whether the car has had any serious accidents.

Thus, predicting the price of second-hand cars is a very laudable enterprise. In this paper, we will assess whether neural networks can be used to accurately predict the price of secondhand cars. The results will also be compared with other methods like linear regression and support vector regression.

This paper proceeds as follows. In this system, various works on neural networks and price prediction have been summarized. The methodology and data



collection are described in this system. The system presents the results for price prediction of second-hand cars. Finally, we end the paper with a conclusion and some ideas towards future works.

2. EXISTING SYSTEM

Predicting the price of second-hand cars has not received much attention from academia despite its huge importance for the society. Bharambe and Dharmadhikari (2015) used artificial neural networks (ANN) to analyse the stock market and predict market behaviour. They claimed that their proposed approach is more accurate than existing ones by 25%.

Pudaruth (2014) used four different supervised machine learning techniques namely kNN (k-Nearest Neighbour), Naïve Bayes, linear regression and decision trees to

predict the price of second-hand cars. The best result was obtained using kNN which had a mean error of 27000 rupees.

Jassbi et al. (2011) used two different neural networks and regression methods to predict the thickness of paint coatings on cars. The error for the final thickness of the paint was found to be 2/99 microns for neural networks and 17/86 for regression. Ahangar et al. (2010) also compared the use of neural networks with linear regression in order to predict the stock prices of companies in Iran. They also found that neural networks had superior performance both in terms of accuracy and speed compared to linear regression. Listiani (2009) used support vector machines (SVM) to predict the price of leased cars.

They showed that SVM performed better than simple linear regression and

multivariate regression. Iseri and Karlik (2009) used neural networks to predict the price of automobiles and achieved a mean square error of 8% compared with 14.4% for regression. Yeo (2009) used neural networks to predict the retention rate for policy holders of automobile insurance. The neural network was able to predict which customers were likely to renew their policy and which ones would terminate soon. Doganis et al. (2006) used artificial neural networks and genetic algorithm in order to predict the sales of fresh milk with an accuracy of 95.4%. Rose (2003) used neural networks to predict the production of cars for different manufacturers.

Disadvantages

An existing methodology doesn't implement DATA PRE-PROCESSING & LABELLING method.

The system not implemented an effective ML Classifiers for predictions in the datasets.

3. PROPOSED SYSTEM

In order to carry out this study, data have been obtained from different car websites and from the small adverts sections found in daily newspapers like L'Express and Le Defi. The data was collected in less than one month interval (i.e. in the month of August in 2014) because like other goods, the price of cars also changes with time. Two hundred records were collected. The data comprises of different features for second-hand cars such as the year (YEAR) in which it was manufactured, the make (MAKE), engine capacity (ENGINE) measured in cubic centimetres, paint (PAINT) type (normal or metallic), transmission (T/N) type (manual or automatic), mileage (MILEAGE) (number

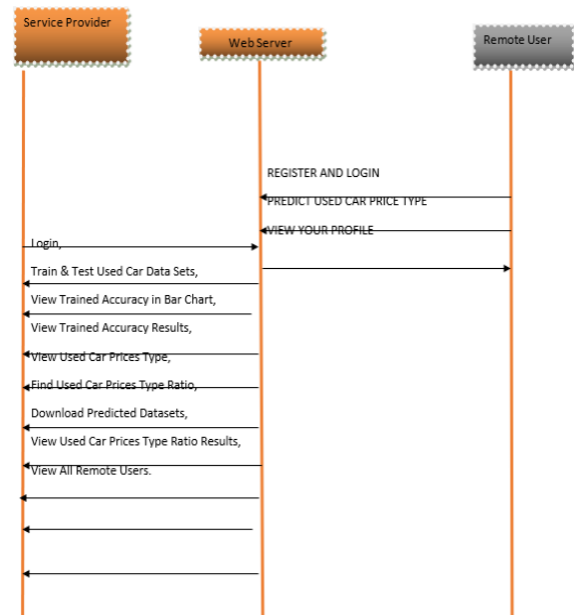
of kilometres the car has been driven) and its price (PRICE) in Mauritian rupees. A large number of experiments have been conducted in order to find the best network structure and the best parameters for the neural network. We found that a neural network with 1 hidden layer and 2 nodes produced the smallest mean absolute error among various neural network structures that were experimented with. However, we found that Support Vector Regression and a multilayer perception with back-propagation produced slightly better predictions than linear regression while the k-Nearest Neighbour algorithm had the worst accuracy among these four approaches. All experiments were performed with a cross validation value of 10 folds.

Advantages

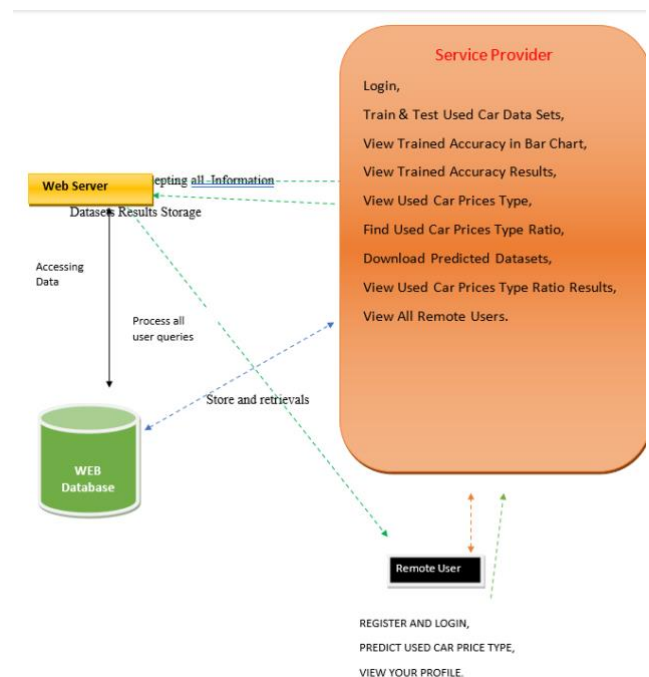
The purpose of linear regression, support vector regression which are more effective for testing and training accuracy.

In this work, the system will assess whether neural networks can be used to accurately predict the price of secondhand cars.

Sequence Diagram



4. ARCHITECTURE DIAGRAM



5. CONCLUSION

The aim of this paper was to predict the price of second-hand reconditioned and second-hand used cars in Mauritius. The car market has been increasing steadily by



around 5% for the last ten years, showing the high demand for cars by the Mauritian population. There are hundreds of car websites in Mauritius but none of them provide such a facility to predict the price of used cars based on their attributes. Our dataset of 200 records was used with the cross-validation technique with ten folds. The car make, year manufactured, paint type, transmission type, engine capacity and mileage have been used to predict the price of second-hand cars using four different machine learning algorithms. The average residual value was reasonably low for all four approaches. Thus, we conclude that predicting the price of second-hand cars is a very risky enterprise but which is feasible. This system will be very useful to car dealers and car owners who need to assess the value of their cars. In the future, we intend to collect more data and more features and to use a larger variety of machine learning algorithms to do the prediction.

6. REFERENCES

- [1] NATIONAL TRANSPORT AUTHORITY. 2015. Available at: <http://nta.govmu.org/English/Statistics/Pages/Archives.aspx>. [Accessed 24 April 2015].
- [2] Bharambe, M. M. P., and Dharmadhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". *Fourth Post Graduate Conference, 24-25th March 2015, Pune, India*.
- [3] Pudaruth, S. (2014) "Predicting the Price of Used Cars using Machine Learning Techniques". *International Journal of Information & Computation Technology*, Vol. 4, No. 7, pp.753- 764.
- [4] Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". *Journal of Industrial Engineering International*, Vol. 7, No. 14, pp. 1-6, November 2010
- [5] Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". *International Journal of Computer Science and Information Security*, Vol.7, No. 2, pp. 38-46.
- [6] Listiani, M. (2009) "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application". Thesis (MSc). Hamburg University of Technology.
- [7] Iseri, A. and Karlik, B. (2009) "An Artificial Neural Network Approach on Automobile Pricing". *Expert Systems with Application: ScienceDirect Journal of Informatics*, Vol. 36, pp. 155-2160, March 2009.
- [8] Yeo, C. A. (2009) "Neural Networks for Automobile Insurance Pricing". *Encyclopedia of Information Science and Technology*, 2nd Edition,



PREDICTING USED CAR PRICE TYPE

Kandi Naga Venkata Sriram (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The number of cars on Mauritian roads has been rising consistently by 5% during the last decade. In 2014, 173 954 cars were registered at the National Transport Authority. Thus, one Mauritian in every six owns a car, most of which are second hand reconditioned cars and used cars. The aim of this study is to assess whether it is possible to predict the price of second-hand cars using artificial neural networks. Thus, data for 200 cars from different sources was gathered and fed to four different machine learning algorithms. We found that support vector machine regression produced slightly better results than using a neural network or linear regression. However, some of the predicted values are quite far away from the

actual prices, especially for higher priced cars. Thus, more investigations with a larger data set are required and more experimentation with different network type and structures is still required in order to obtain better predictions.

1. INTRODUCTION

According to the data obtained from the National Transport Authority (2014), there has been an increase of 254% in the number of cars from 2003 (68, 524) to 2014 (173, 954), as shown in Figure 1. We can thus infer that the sale of second-hand imported (reconditioned) cars and second-hand used cars has eventually increase given that new cars represent only a very small percentage of the total number of cars sold each year. Most individuals in Mauritius who buy new cars also want to know about the resale value of their cars after some years so that they can sell it in the used car market.

Price prediction of second-hand cars depends on numerous factors. The most important ones are manufacturing year, make, model, mileage, horsepower and country of origin. Some other factors are type and amount of fuel per usage, the type of braking system, its acceleration, the interior style, its physical state, volume of

cylinders (measured in cubic centimeters), size of the car, number of doors, weight of the car, consumer reviews, paint colour and type, transmission type, whether it is a sports car, sound system, cosmic wheels, power steering, air conditioner, GPS navigator, safety index etc. In the Mauritian context, there are some special factors that are also usually considered such as who were the previous owners and whether the car has had any serious accidents.

Thus, predicting the price of second-hand cars is a very laudable enterprise. In this paper, we will assess whether neural networks can be used to accurately predict the price of secondhand cars. The results will also be compared with other methods like linear regression and support vector regression.

This paper proceeds as follows. In this system, various works on neural networks and price prediction have been summarized. The methodology and data



collection are described in this system. The system presents the results for price prediction of second-hand cars. Finally, we end the paper with a conclusion and some ideas towards future works.

2. EXISTING SYSTEM

Predicting the price of second-hand cars has not received much attention from academia despite its huge importance for the society. Bharambe and Dharmadhikari (2015) used artificial neural networks (ANN) to analyse the stock market and predict market behaviour. They claimed that their proposed approach is more accurate than existing ones by 25%.

Pudaruth (2014) used four different supervised machine learning techniques namely kNN (k-Nearest Neighbour), Naïve Bayes, linear regression and decision trees to

predict the price of second-hand cars. The best result was obtained using kNN which had a mean error of 27000 rupees.

Jassbi et al. (2011) used two different neural networks and regression methods to predict the thickness of paint coatings on cars. The error for the final thickness of the paint was found to be 2/99 microns for neural networks and 17/86 for regression. Ahangar et al. (2010) also compared the use of neural networks with linear regression in order to predict the stock prices of companies in Iran. They also found that neural networks had superior performance both in terms of accuracy and speed compared to linear regression. Listiani (2009) used support vector machines (SVM) to predict the price of leased cars.

They showed that SVM performed better than simple linear regression and

multivariate regression. Iseri and Karlik (2009) used neural networks to predict the price of automobiles and achieved a mean square error of 8% compared with 14.4% for regression. Yeo (2009) used neural networks to predict the retention rate for policy holders of automobile insurance. The neural network was able to predict which customers were likely to renew their policy and which ones would terminate soon. Doganis et al. (2006) used artificial neural networks and genetic algorithm in order to predict the sales of fresh milk with an accuracy of 95.4%. Rose (2003) used neural networks to predict the production of cars for different manufacturers.

Disadvantages

An existing methodology doesn't implement DATA PRE-PROCESSING & LABELLING method.

The system not implemented an effective ML Classifiers for predictions in the datasets.

3. PROPOSED SYSTEM

In order to carry out this study, data have been obtained from different car websites and from the small adverts sections found in daily newspapers like L'Express and Le Defi. The data was collected in less than one month interval (i.e. in the month of August in 2014) because like other goods, the price of cars also changes with time. Two hundred records were collected. The data comprises of different features for second-hand cars such as the year (YEAR) in which it was manufactured, the make (MAKE), engine capacity (ENGINE) measured in cubic centimetres, paint (PAINT) type (normal or metallic), transmission (T/N) type (manual or automatic), mileage (MILEAGE) (number

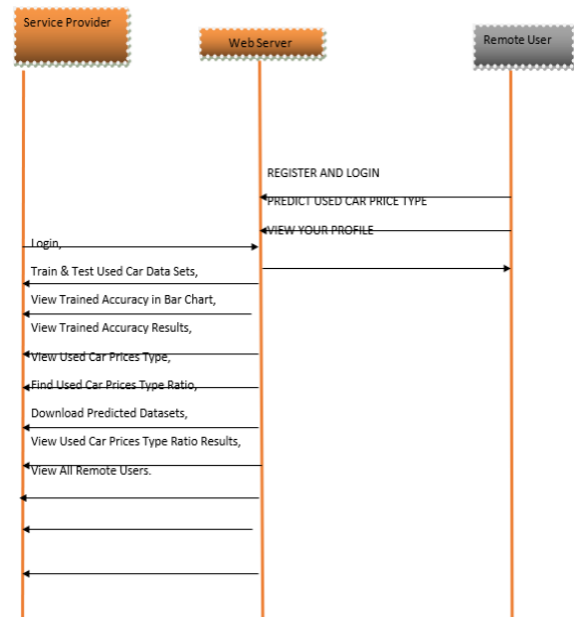
of kilometres the car has been driven) and its price (PRICE) in Mauritian rupees. A large number of experiments have been conducted in order to find the best network structure and the best parameters for the neural network. We found that a neural network with 1 hidden layer and 2 nodes produced the smallest mean absolute error among various neural network structures that were experimented with. However, we found that Support Vector Regression and a multilayer perception with back-propagation produced slightly better predictions than linear regression while the k-Nearest Neighbour algorithm had the worst accuracy among these four approaches. All experiments were performed with a cross validation value of 10 folds.

Advantages

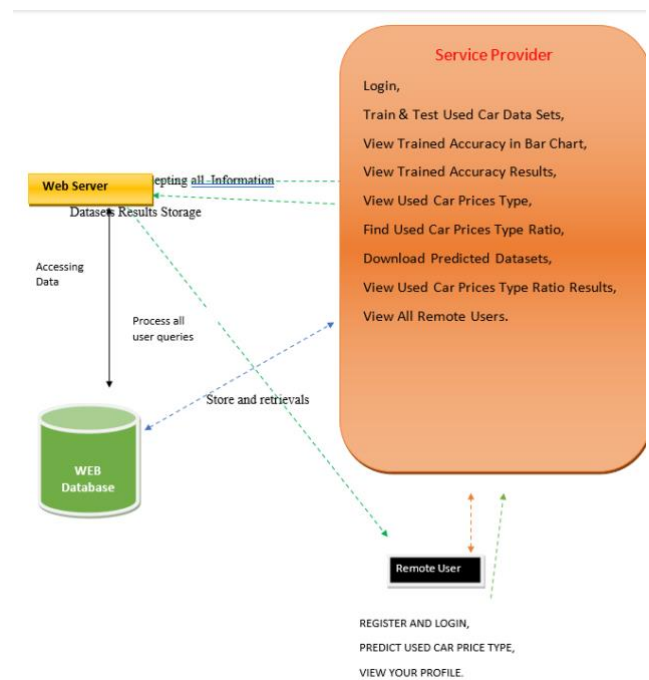
The purpose of linear regression, support vector regression which are more effective for testing and training accuracy.

In this work, the system will assess whether neural networks can be used to accurately predict the price of secondhand cars.

Sequence Diagram



4. ARCHITECTURE DIAGRAM



5. CONCLUSION

The aim of this paper was to predict the price of second-hand reconditioned and second-hand used cars in Mauritius. The car market has been increasing steadily by



around 5% for the last ten years, showing the high demand for cars by the Mauritian population. There are hundreds of car websites in Mauritius but none of them provide such a facility to predict the price of used cars based on their attributes. Our dataset of 200 records was used with the cross-validation technique with ten folds. The car make, year manufactured, paint type, transmission type, engine capacity and mileage have been used to predict the price of second-hand cars using four different machine learning algorithms. The average residual value was reasonably low for all four approaches. Thus, we conclude that predicting the price of second-hand cars is a very risky enterprise but which is feasible. This system will be very useful to car dealers and car owners who need to assess the value of their cars. In the future, we intend to collect more data and more features and to use a larger variety of machine learning algorithms to do the prediction.

6. REFERENCES

- [1] NATIONAL TRANSPORT AUTHORITY. 2015. Available at: <http://nta.govmu.org/English/Statistics/Pages/Archives.aspx>. [Accessed 24 April 2015].
- [2] Bharambe, M. M. P., and Dharmadhikari, S. C. (2015) "Stock Market Analysis Based on Artificial Neural Network with Big data". *Fourth Post Graduate Conference, 24-25th March 2015, Pune, India*.
- [3] Pudaruth, S. (2014) "Predicting the Price of Used Cars using Machine Learning Techniques". *International Journal of Information & Computation Technology*, Vol. 4, No. 7, pp.753- 764.
- [4] Jassibi, J., Alborzi, M. and Ghoreshi, F. (2011) "Car Paint Thickness Control using Artificial Neural Network and Regression Method". *Journal of Industrial Engineering International*, Vol. 7, No. 14, pp. 1-6, November 2010
- [5] Ahangar, R. G., Mahmood and Y., Hassen P.M. (2010) "The Comparison of Methods, Artificial Network with Linear Regression using Specific Variables for Prediction Stock Prices in Tehran Stock Exchange". *International Journal of Computer Science and Information Security*, Vol.7, No. 2, pp. 38-46.
- [6] Listiani, M. (2009) "Support Vector Regression Analysis for Price Prediction in a Car Leasing Application". Thesis (MSc). Hamburg University of Technology.
- [7] Iseri, A. and Karlik, B. (2009) "An Artificial Neural Network Approach on Automobile Pricing". *Expert Systems with Application: ScienceDirect Journal of Informatics*, Vol. 36, pp. 155-2160, March 2009.
- [8] Yeo, C. A. (2009) "Neural Networks for Automobile Insurance Pricing". *Encyclopedia of Information Science and Technology*, 2nd Edition,

CREDIT CARD FRAUD DETECTION USING STATE-OF-THE-ART MACHINE LEARNING

Kanuboyina Prem Sai (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

People can use credit cards for online transactions as it provides an efficient and easy-to-use facility. With the increase in usage of credit cards, the capacity of credit card misuse has also enhanced. Credit card frauds cause significant financial losses for both credit card holders and financial companies. In this research study, the main aim is to detect such frauds, including the accessibility of public data, high-class imbalance data, the changes in fraud nature, and high rates of false alarm. The relevant literature presents many machine learning based approaches for credit card detection, such as Extreme Learning Method, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression and XG Boost. However, due to low accuracy, there is still a need to apply state of the art deep learning algorithms to reduce fraud losses. The main focus has been to apply the recent development of deep learning algorithms for this purpose. Comparative analysis of both machine learning and deep learning algorithms was performed to find efficient outcomes. The detailed empirical analysis is carried out using the European card benchmark dataset for fraud detection. A machine learning algorithm was first applied to the dataset, which improved the accuracy of detection of the frauds to some extent. Later, three architectures based on a convolutional neural network are applied to improve fraud detection performance. Further addition of layers further increased the accuracy of detection. A comprehensive empirical analysis has been carried out by applying variations in the number of hidden layers, epochs and applying the latest models. The evaluation of research work shows the improved results achieved, such as accuracy, f1-score, precision and AUC Curves having optimized values of 99.9%,85.71%,93%, and 98%, respectively. The proposed model outperforms the state-of-the-art machine learning and deep learning algorithms for credit card detection problems. In addition, we have performed experiments by balancing the data and

applying deep learning algorithms to minimize the false negative rate. The proposed approaches can be implemented effectively for the real-world detection of credit card fraud.

INDEX TERMS- Fraud detection, deep learning, machine learning, online fraud, credit card frauds, transaction data analysis.

1. INTRODUCTION

Credit card fraud (CCF) is a type of identity theft in which someone other than the owner makes an unlawful transaction using a credit card or account details. A credit card that has been stolen, lost, or counterfeited might result in fraud. Card-not-present fraud, or the use of your credit card number in e-commerce transactions has also become increasingly common as a result of the increase in online shopping. Increased fraud, such as CCF, has resulted from the expansion of e-banking and several online payment environments, resulting in annual losses of billions of dollars. In this era of digital payments, CCF detection has become one of the most important goals. As a business owner, it cannot be disputed that the future is heading towards a cashless culture. As a result, typical payment methods will no longer be used in the future, and therefore they will not be helpful for expanding a business. Customers will not always visit the business with cash in their pockets. They are now placing a premium on debit and credit card payments. As a result, companies will need to update their environment to ensure that they can take all types of payments. In the next years, this situation is expected to become much more severe [1].

In 2020, there were 393,207 cases of CCF out of approximately 1.4 million total reports of identity theft [4]. CCF is now the second most prevalent sort of identity theft recorded as of this year, only following government documents and benefits fraud [5]. In 2020, there were 365,597 incidences of fraud perpetrated using new credit card accounts [10]. The number of identity theft complaints has climbed by 113% from 2019 to 2020, with credit card identity theft reports increasing by 44.6% [14]. Payment card theft cost the global economy \$24.26 billion last year. With 38.6% of reported card fraud losses in 2018, the United States is the most vulnerable country to credit theft.

As a result, financial institutions should prioritize equipping themselves with an automated fraud detection system. The goal of supervised CCF detection is to create a machine learning (ML) model based on existing transactional credit card payment data. The model should distinguish between fraudulent and non fraudulent transactions, and use this information to decide whether an incoming transaction is fraudulent or not. The issue involves a variety of

fundamental problems, including the system's quick reaction time, cost sensitivity, and feature pre-processing. ML is a field of artificial intelligence that uses a computer to make predictions based on prior data trends [1]

ML models have been used in many studies to solve numerous challenges. Deep learning (DL) algorithms applied applications in computer network, intrusion detection, banking, insurance, mobile cellular networks, health care fraud detection, medical and malware detection, detection for video surveillance, location tracking, Android malware detection, home automation, and heart disease prediction. We explore the practical application of ML, particularly DL algorithms, to identify credit card thefts in the banking industry in this paper. For data categorisation challenges, the support vector machine (SVM) is a supervised ML technique. It is employed in a variety of domains, including image recognition [25], credit rating [5], and public safety [16]. SVM can tackle linear and nonlinear binary classification problems, and it finds a hyper plane that separates the input data in the support vector, which is superior to other classifiers. Neural networks were the first method used to identify credit card theft in the past [4]. As a result, (DL), a branch of ML, is currently focused on DL approaches.

In recent years, deep learning approaches have received significant attention due to substantial and promising outcomes in various applications, such as computer vision, natural language processing, and voice. However, only a few studies have examined the application of deep neural networks in identifying CCF. [3]. It uses a number of deep learning algorithms for detecting CCF. However, in this study, we choose the CNN model and its layers to determine if the original fraud is the normal transaction of qualified datasets. Some transactions are common in datasets that have been labelled fraudulent and demonstrate questionable transaction behavior . As a result, we focus on supervised and unsupervised learning in this research paper.

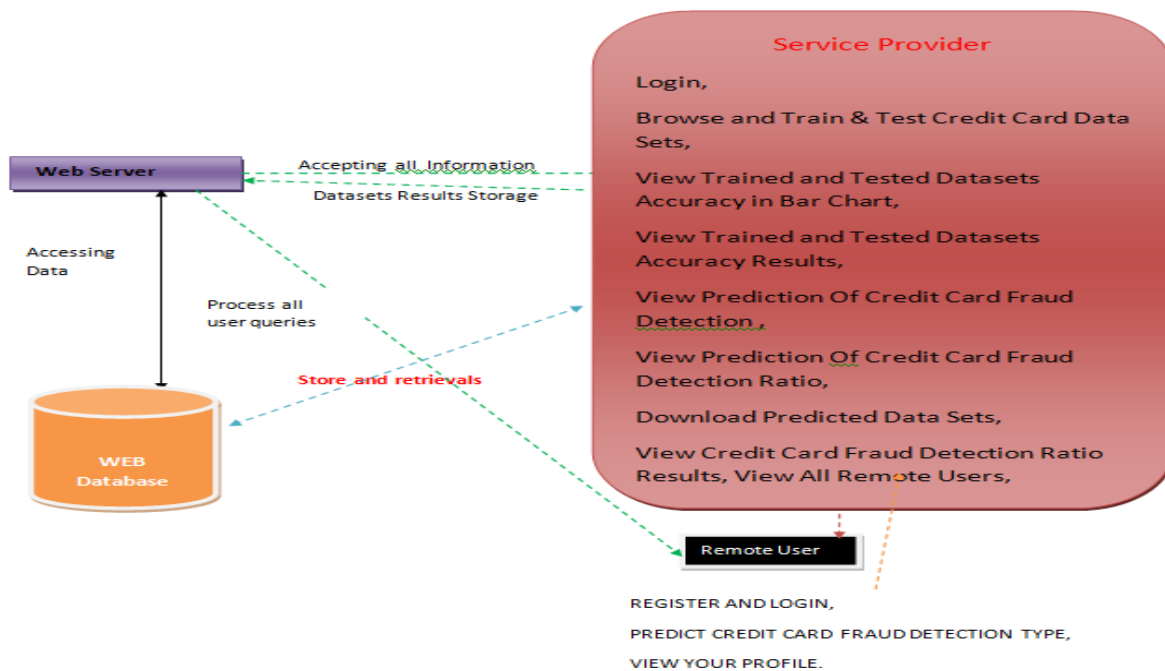
The class imbalance is the problem in ML where the total number of a class of data (positive) is far less than the total number of another class of data (negative). The classification challenge of the unbalanced dataset has been the subject of several studies. An extensive collection of studies can provide several answers. Therefore, to the best of our knowledge, the problem of class imbalance has not yet been solved. We propose to alter the DL algorithm of the CNN model by adding the additional layers for features extraction and the classification of credit card transactions as fraudulent or otherwise. The top attributes from the prepared dataset are ranked using feature selection techniques. After that, CCF is classified using several supervised machine-driven and deep learning models.

In this study, the main aim is to detect fraudulent transactions using credit cards with the help of ML algorithms and deep learning algorithms. This study makes the following contributions:

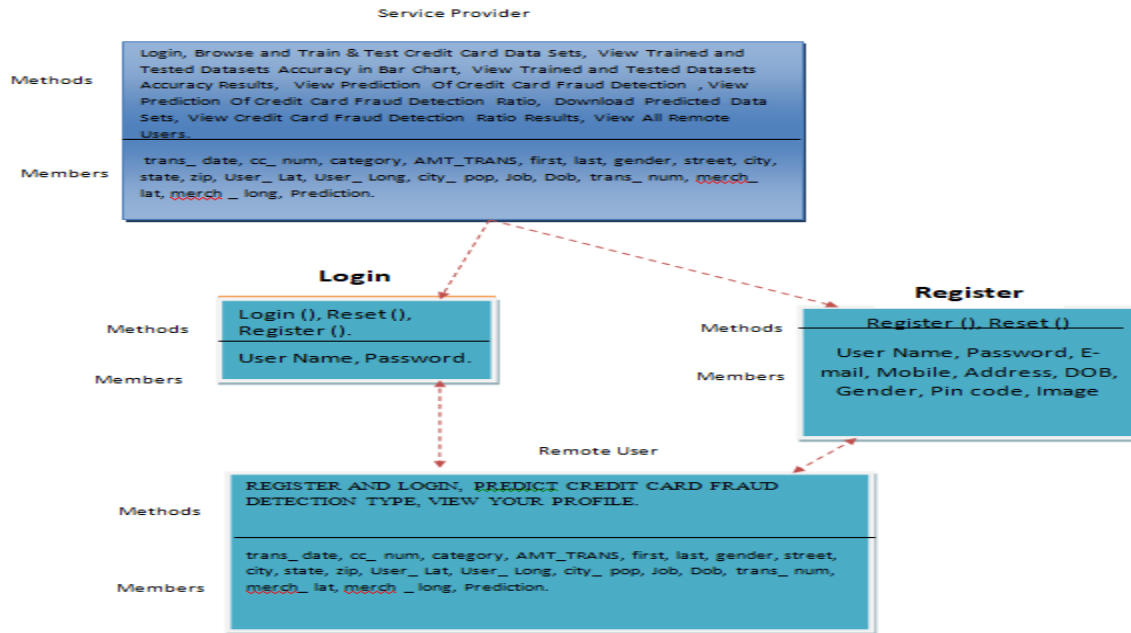
- _ Feature selection algorithms are used to rank the top features from the CCF transaction dataset, which help in class label predictions.
- _ The deep learning model is proposed by adding a number of additional layers that are then used to extract the features and classification from the credit card fraud detection dataset.
- _ To analyse the performance CNN model, apply different architecture of CNN layers.
- _ To perform a comparative analysis between ML with DL algorithms and proposed CNN with baseline model, the results prove that the proposed approach outperforms existing approaches.
- _ To assess the accuracy of the classifiers, performance evaluation measures, accuracy, precision, and recall are used. Experiments are performed on the latest credit cards dataset.

The rest of the paper is structured as follows: The second section examines the related works. The proposed model and its methodology are described in depth in Section 3. The dataset and evaluation measures are described in Section 4. It also shows the outcomes of our tests on a real dataset, as well as the analysis.

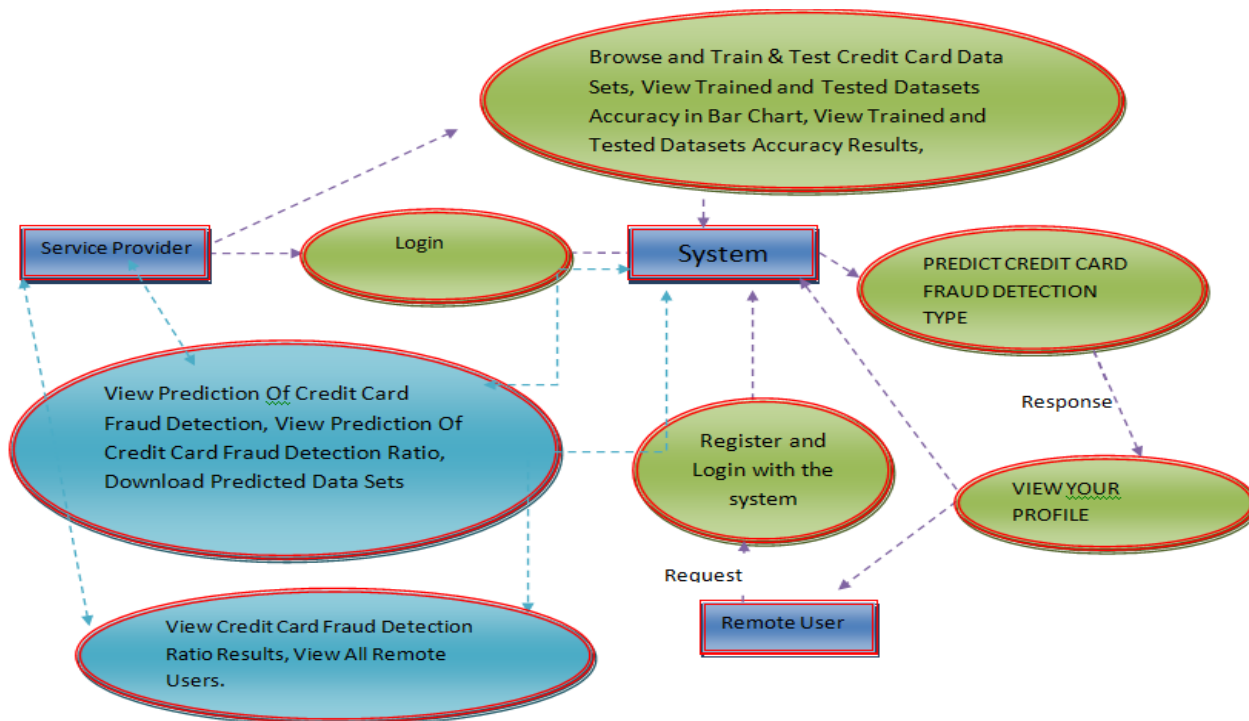
2. ARCHITECTURE DIAGRAM



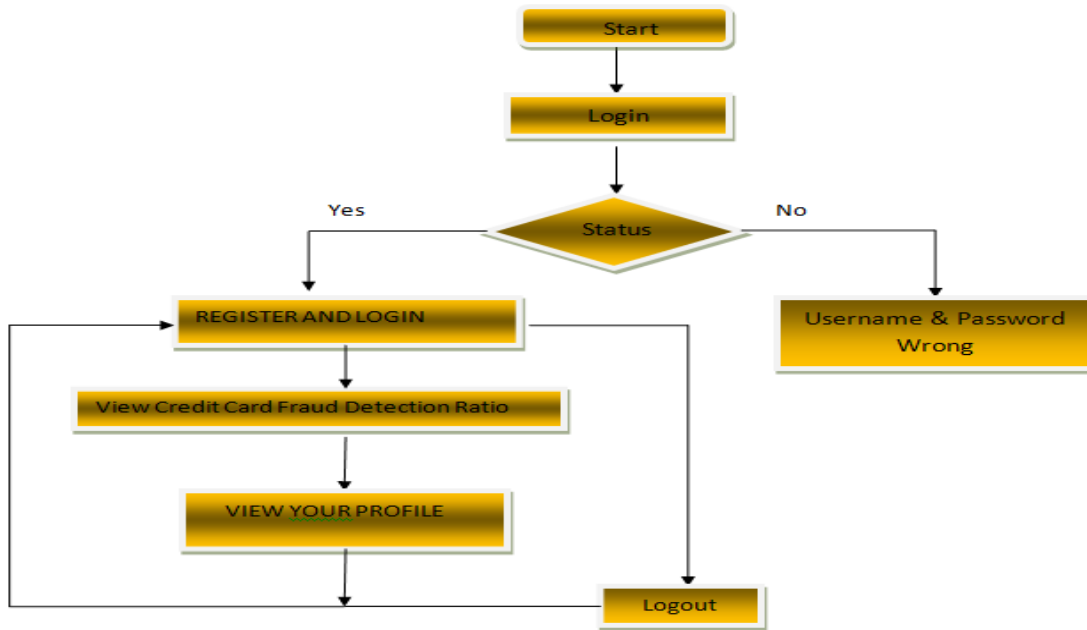
3. CLASS DIAGRAM



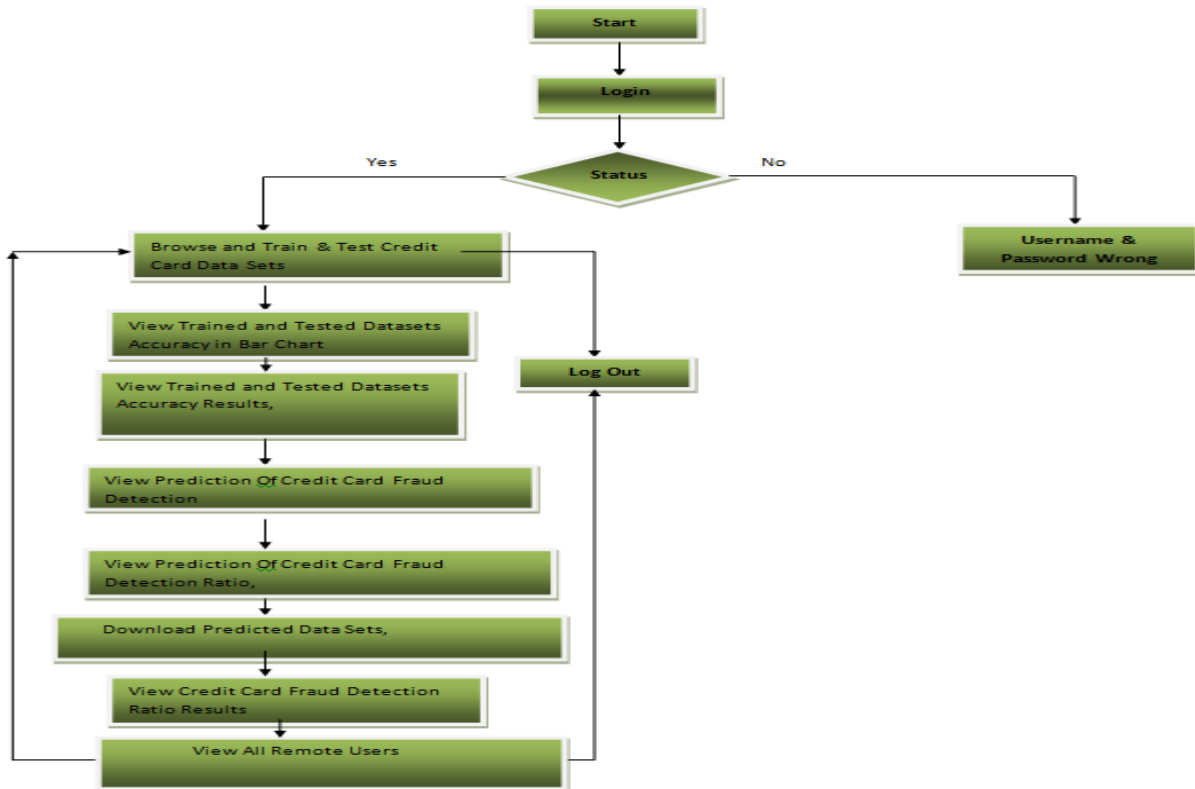
4. DATA FLOW DIAGRAM



➤ **Flow Chart : Remote User**



➤ **Flow Chart : Service Provider**



5. CONCLUSION

CCF is an increasing threat to financial institutions. Fraudsters tend to constantly come up with new fraud methods. A robust classifier can handle the changing nature of fraud. Accurately predicting fraud cases and reducing false-positive cases is the foremost priority of a fraud detection system. The performance of ML methods varies for each individual business case. The type of input data is a dominant factor that drives different ML methods. For detecting CCF, the number of features, number of transactions, and correlation between the features are essential factors in determining the model's performance. DL methods, such as CNNs and their layers, are associated with the processing of text and the baseline model. Using these methods for the detection of credit cards yields better performance than traditional algorithms. Comparing all the algorithm performances side to side, the CNN with 20 layers and the baseline model is the top method with an accuracy of 99.72%. Numerous sampling techniques are used to increase the performance of existing examples, but they significantly decrease on the unseen data. The performance on unseen data increased as the class imbalance increased. Future work associated may explore the use of more state of art deep learning methods to improve the performance of the model proposed in this study.

6. REFERENCES

- [1] Y. Abakarim, M. Lahby, and A. Attioui, "An efficient real time model for credit card fraud detection based on deep learning," in *Proc. 12th Int. Conf. Intell. Systems: Theories Appl.*, Oct. 2018, pp. 1_7, doi: 10.1145/3289402.3289530.
- [2] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Inter- discipl. Rev., Comput. Statist.*, vol. 2, no. 4, pp. 433_459, Jul. 2010, doi: 10.1002/wics.101.
- [3] V. Arora, R. S. Leekha, K. Lee, and A. Kataria, "Facilitating user authorization from imbalanced data logs of credit cards using artificial intelligence," *Mobile Inf. Syst.*, vol. 2020, pp. 1_13, Oct. 2020, doi: 10.1155/2020/8885269.
- [4] A. O. Balogun, S. Basri, S. J. Abdulkadir, and A. S. Hashim, "Performance analysis of feature selection methods in software defect prediction: A search method approach," *Appl. Sci.*, vol. 9, no. 13, p. 2764, Jul. 2019, doi: 10.3390/app9132764.
- [5] B. Bandaranayake, "Fraud and corruption control at education system level: A case study of the Victorian department of education and early childhood development in Australia," *J. Cases Educ. Leadership*, vol. 17, no. 4, pp. 34_53, Dec. 2014, doi: 10.1177/1555458914549669.

- [6] J. B^aaszczy[«]ski, A. T. de Almeida Filho, A. Matuszyk, M. Szelg_„, and R. S^aowi[«]ski, "Auto loan fraud detection using dominance-based rough set approach versus machine learning methods," *Expert Syst. Appl.*, vol. 163, Jan. 2021, Art. no. 113740, doi: 10.1016/j.eswa.2020.113740.
- [7] B. Branco, P. Abreu, A. S. Gomes, M. S. C. Almeida, J. T. Ascens^o, and P. Bizarro, "Interleaved sequence RNNs for fraud detection," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2020, pp. 3101_3109, doi: 10.1145/3394486.3403361.
- [8] F. Cartella, O. Anunciacao, Y. Funabiki, D. Yamaguchi, T. Akishita, and O. Elshocht, "Adversarial attacks for tabular data: Application to fraud detection and imbalanced data," 2021, *arXiv:2101.08030*.
- [9] S. S. Lad, I. Dept. of CSERajarambapu Institute of Technology Rajaramnagar Sangli Maharashtra, and A. C. Adamuthe, "Malware classification with improved convolutional neural network model," *Int. J. Comput. Netw. Inf. Secur.*, vol. 12, no. 6, pp. 30_43, Dec. 2021, doi: 10.5815/ijcnis.2020.06.03.
- [10] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Proc. Comput. Sci.*, vol. 165, pp. 631_641, Jan. 2019, doi: 10.1016/j.procs.2020.01.057.



PREDICTING THE IMPACT OF DISRUPTIONS TO URBAN RAIL TRANSIT SYSTEM

Kanumuri Srinivas Varma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract

Service disruptions of rail transit systems become more frequent in the past decades in urban cities like Singapore, due to various reasons such as power failures, signal errors, etc. We study and predict the impact of disruptions to transit systems and commuters. This benefits service providers in making both short and long term plans to improve their services. Specifically, we define two metrics, stay ratio and travel delay, to quantify the impact. To tackle the main challenge of abnormal data scarcity, i.e., only 6 observed disruptions in our one-year data records, we propose to format the problem into a training problem on a feature space relevant to alternative route choices of the commuters. We demonstrate the new feature space corresponds to more similar data distribution among different disruptions, which is beneficial for training more generalisable predictors for future disruptions. We implement and evaluate our approach with a real-world transit card dataset. The result clearly shows that our method outperforms a range of baseline methods.

1. INTRODUCTION

The rapid rail system is the backbone of the public transit systems (PTS) in urban cities. Malfunction of the rail system even in a small region may have ripple effects and significantly impair the PTS. According to our study on Singapore Mass Rapid Transit (MRT) rail system, major disruptions take place due to many reasons including technical faults, extreme weathers, human injuries, etc. The journey of thousands or even tens of thousands of commuters may be impaired. Many of them have to quit the PTS and resort to other transportation alternatives (e.g., taxis).

This paper aims at predicting the impact of rail system disruptions at the time of occurrence. Such knowledge not only benefits the PTS provider in understanding

the degradation of service, making better emergent plans and planning appropriate new services in PTS to improve system resilience, but also benefits commuters in preparing for the hazards brought by disruptions [1] [2]. Specifically, we define the following two metrics to assess the impact of disruptions. (1) Stay ratio indicates the percentage of rail riders who choose to stay within the PTS and take alternative rail lines and/or buses to complete their trip. (2) Travel delay indicates the extra time spent on alternative routes for those who stay within the PTS. Obviously, higher stay ratio and lower travel delay indicate smaller impact by a disruption. Although there have been efforts made to analysing the influence of abnormal conditions of railway on



commuters [3]–[5], most of them apply empirical knowledge or simplified human behaviour models to reason human choices, and based on that analyze the impact on commuters. Some exploit real transportation data to understand human behaviours, but they are often limited to normal PTS conditions. In this paper, taking a unique approach, we explore the transportation data during rail system disruptions and learn from the true human choices. We train a human behaviour model from those abnormal data and apply the model to predict the impact of future disruptions.

Being simple in rationale, our approach is especially challenged due to the scarcity of abnormal data, i.e., those from only 6-8 major disruptions per year. A direct challenge comes from the lack of training data for us to build an accurate model using supervised learning. The limited observation of disruptions makes the trained model difficult to generalize, i.e., applicable to future disruptions unseen in the training stage. The problem becomes more challenging if we consider that only the trips of regular commuters (which is a small portion of the total affected commuters) can be utilized to analyze human behaviours, extract features and label impact metrics, because for irregular commuters there is no way to infer their original travel intention and thus no confidence with regard to their choices under disruptions.

In order to address the above challenges, we propose a novel idea of domain projection to tackle the data distribution mismatch between training and testing sets especially in the situation of data scarcity. Similar but different to the situation of

canonical transfer learning, our data in both the training and testing sets is scarce and hence no big picture of the distribution can be profiled. Therefore, we claim the importance of proactively finding a feature space where the training and testing disruptions share similar distributions of extracted features. Specifically, the proposed domain projection method converts the original training problem on the feature space relevant to disruption itself to a new training problem on a different feature space relevant to alternative route choices of the commuters,

2. EXISTING SYSTEM

- ❖ Sun et al. [5] estimates the normal spatio-temporal distribution of commuters in rail system, and try to infer the number of affected commuters when there is a disruption. Sun et al. [4] try to reason commuters' travel delay based on their choices (e.g., stay or leave PTS). Yin et al. [11] define the impact as the damage to rail network efficiency, and utilize graph theory to quantify the impact of disruption. Some works predict impact based on actual mobility data measured from real world.
- ❖ Examples include Pan et al. [12] who take the average impact of similar historical incidents to predict that of future incidents, Fang et al. [3] who leverage contextual features and post-incident travel delays to predict future travel delays, and Garib et al. [13] who use statistical models based on contextual features to predict travel delay. Most existing



studies have not thoroughly investigated the ability of generalization and are not validated with real world incidents at the scale of this paper.

- ❖ Other studies focus on forecasting the traffic flow under anomalous conditions [14]–[16] taking a period of post-incident traffic flow as input. The traffic flows, however, cannot be translated to fine-grained impact to commuters. To sum up, so far there is no existing study which measures impact from real incidents, and meanwhile explores the model generalizing ability to predict the impact of a variety of future incidents.

Disadvantage

- 1) .The system doesn't have a method Analyzing Under-disruption Choices.
- 2). There is no system to analyze accurate disruptions on large data sets.

3.PROPOSED SYSTEM

- ❖ The system proposes a novel idea of domain projection to tackle the data distribution mismatch between training and testing sets especially in the situation of data scarcity. Similar but different to the situation of canonical transfer learning, our data in both the training and testing sets is scarce and hence no big picture of the distribution can be profiled. Therefore, we claim the importance of proactively finding a feature space where the training and

testing disruptions share similar distributions of extracted features.

- ❖ Specifically, the proposed domain projection method converts the original training problem on the feature space relevant to disruption itself to a new training problem on a different feature space relevant to alternative route choices of the commuters, which unifies our view of disruptions by their effect on commuter route choices. A model trained from the converted feature space can thus be generalized to arbitrary disruptions as long as the commuter route choices can be inferred from the disruptions.

Our contributions are summarized as follows:

- To the best of our knowledge, this is the first study of impact prediction of rail system disruptions that learns models from true human behaviors' in disruptions.
- The system proposes a novel domain projection method to address the challenges arising from data scarcity, with which we are able to build an accurate and more generalizable model for arbitrary disruptions.
- The system implements and experimentally evaluates our approach with the Singapore MRT ride records in year 2015 that involve 6 major disruptions. The results demonstrate that our method outperforms all the baseline methods.

Advantages



1. . The system has developed with huge amount of data sets to measure accurate disruptions.
2. . An efficient Domain projection to convert the prediction problem in the domain of disruption into that in the domain of interested alternative routes (IARs) that may be chosen by the commuters during disruptions, where we may address the challenge of data scarcity and train a generalizable model.

3. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

4. REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area

Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

4.1 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

5. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

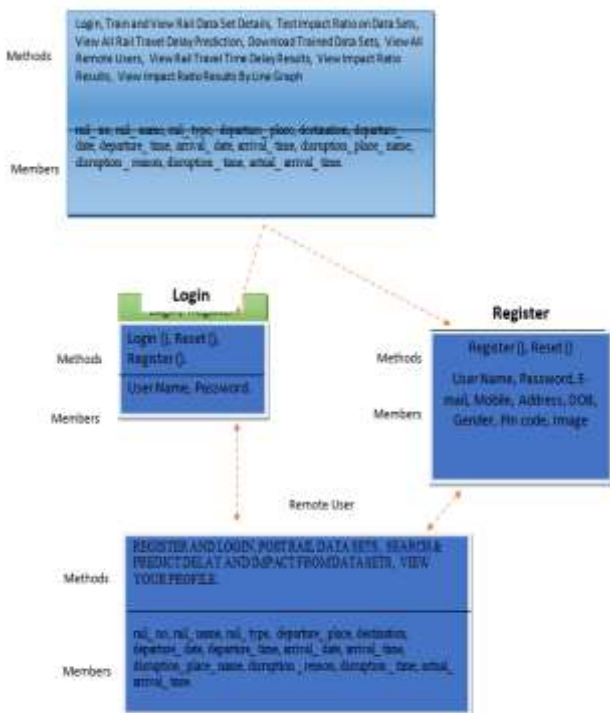
Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are

developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Class Diagram :



6. CONCLUSIONS

We propose a comprehensive solution to predict the impact of rail system disruptions, based on the real behaviors of affected commuters during disruptions. To tackle the challenge of training data scarcity, We propose to project a disruption and its affected OD into a different domain of features abstracted from commuters’ alternative route choices. The training accuracy and generalizing ability are greatly improved. Experimental results using real-world data demonstrate the effectiveness of our proposed solution.

7. REFERENCES

[1] P. Zhou, Y. Zheng, and M. Li, “How long to wait? predicting bus arrival time with mobile phone based participatory sensing,” in Proceedings of the 10th international conference on Mobile systems, applications, and services, 2012, pp. 379–392.

[2] Z. Liu, Z. Gong, J. Li, and K. Wu, “Mobility-aware dynamic taxi ridesharing,” 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, 2020, pp. 961–972.

[3] Z. Fang, Y. Yang, D. Zhang et al., “Mac: Measuring the impacts of anomalies on travel time of multiple transportation systems,” Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 3, no. 2, p. 42, 2019.

[4] H. Sun, J. Wu, L. Wu, X. Yan, and Z. Gao, “Estimating the influence of common disruptions on urban rail transit networks,” Transportation Research Part A: Policy and Practice, vol. 94, pp. 62–75, 2016.



- [5] L. Sun, D.-H. Lee, A. Erath, and X. Huang, "Using smart card data to extract passenger's spatio-temporal density and train's trajectory of mrt system," in Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, 2012, pp. 142–148.
- [6] G. Voronoi, "Nouvelles applications des param`etres continus `a la th`eorie des formes quadratiques. deuxi`eme m`emoire. recherches surles parall'ello`edres primitifs." Journal f`ur die reine und angewandte Mathematik, vol. 134, pp. 198–287, 1908.
- [7] S. Robert, "Algorithms in c, part 5: Graph algorithms," 2002.
- [8] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [9] S. L. T. A. Datamall. <https://www.mytransport.sg/content/mytransport/home.html>.
- [10] D. M. Tax and R. P. Duin, "Support vector data description," Machine learning, vol. 54, no. 1, pp. 45–66, 2004.

CONSTRUCTION SITE ACCIDENT ANALYSIS USING TEXT MINING AND NATURAL LANGUAGE PROCESSING TECHNIQUES

Kanuri Veera Venkata Someswar (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract

Workplace safety is a major concern in many countries. Among various industries, construction sector is identified as the most hazardous work place. Construction accidents not only cause human sufferings but also result in huge financial loss. To prevent reoccurrence of similar accidents in the future and make scientific risk control plans, analysis of accidents is essential. In construction industry, fatality and catastrophe investigation summary reports are available for the past accidents. In this study, text mining and natural language process (NLP) techniques are applied to analyze the construction accident reports. To be more specific, five baseline models, support vector machine (SVM), linear regression (LR), K-nearest neighbor (KNN), decision tree (DT), Naive Bayes (NB) and an ensemble model are proposed to classify the causes of the accidents. Besides, Sequential Quadratic Programming (SQP) algorithm is utilized to optimize weight of each classifier involved in the ensemble model. Experiment results show that the optimized ensemble model outperforms rest models considered in this study in terms of average weighted F1 score. The result also shows that the proposed approach is more robust to cases of low support. Moreover, an unsupervised chunking approach is proposed to extract common objects which cause the accidents based on grammar rules identified in the reports. As harmful objects are one of the major factors leading to construction accidents, identifying such objects is extremely helpful to mitigate potential risks. Certain limitations of the proposed methods are discussed and suggestions and future improvements are provided.

1. INTRODUCTION

Construction industry remains globally the most dangerous work place [1,2]. There are >2.78 million deaths every year caused by occupational accidents according to the International

Labor Organization (ILO) [3]. Among which approximately one of six fatal accidents occur in the construction sector. Construction accidents not only cause severe health issues but also lead to huge financial loss. To prevent occurrence of similar accidents and promote workplace safety, analysis of past accidents is crucial. Based on the results of cause analysis, proper actions can be taken by safety professionals to remove or reduce the identified causes. It is also noted that one major factor contributing to the risk of an accident is the presence of harmful objects [4] such as misused tools, sharp objects nearby, damaged equipment. Mitigating strategies can be made accordingly after identification of such objects. For example, raising awareness, performing mandatory regular checks before operation of the machine which went wrong and caused the accident earlier.

In construction industry, a catastrophe investigation report is generated after a fatal accident which provides a complete description of the accident, such text data can be utilized for further analysis.

Studies of text mining, NLP and ensemble techniques for the analysis of construction accidents report are rare. Motivation of this paper is to fill this research gap. In this study, text mining and NLP techniques are applied to analyze the construction site accidents using the data from Occupational Safety and Health Administration (OSHA). An ensemble model is proposed to classify the causes of accidents. While in conventional majority voting mechanism, equal weights are assigned to each base classifier involved in the ensemble model. In this study, the weight of each base classifier is optimized by Sequential Quadratic Programming (SQP) algorithm. Moreover, a rule based chunker is developed to identify common objects which cause the accidents. Neither SQP optimization nor chunker algorithm is found to be applied in this field in any existing literatures.

Major contributions of this work are:

- Various text mining and NLP techniques are explored with respect to construction site accidents analysis.
- Ensemble algorithm which has not been well studied in this field is proposed to classify the causes of accidents and SQP algorithm is utilized to search for optimal weights of the ensemble model.

- A rule based chunker is developed for dangerous objects extraction. Neither SQP optimization algorithm nor rule based chunker with regard to this field is found in the state of the art.
- Case studies are designed using OSHA dataset and effectiveness of the proposed approaches is verified by the experiment results.

2. LITERATURE REVIEW

There are several studies which utilize text mining or natural language process (NLP) approaches for occupational accidents analysis. Bertke et al. [5] developed a Naïve Bayesian model to classify the compensation claims causation due to work related injuries. The proposed model achieved an overall accuracy of approximately 90%, however the accuracy of claims belongs to minor injury categories dropped. Taylor et al. [6] applied Naïve Bayesian and Fuzzy models to categorize the injury outcome

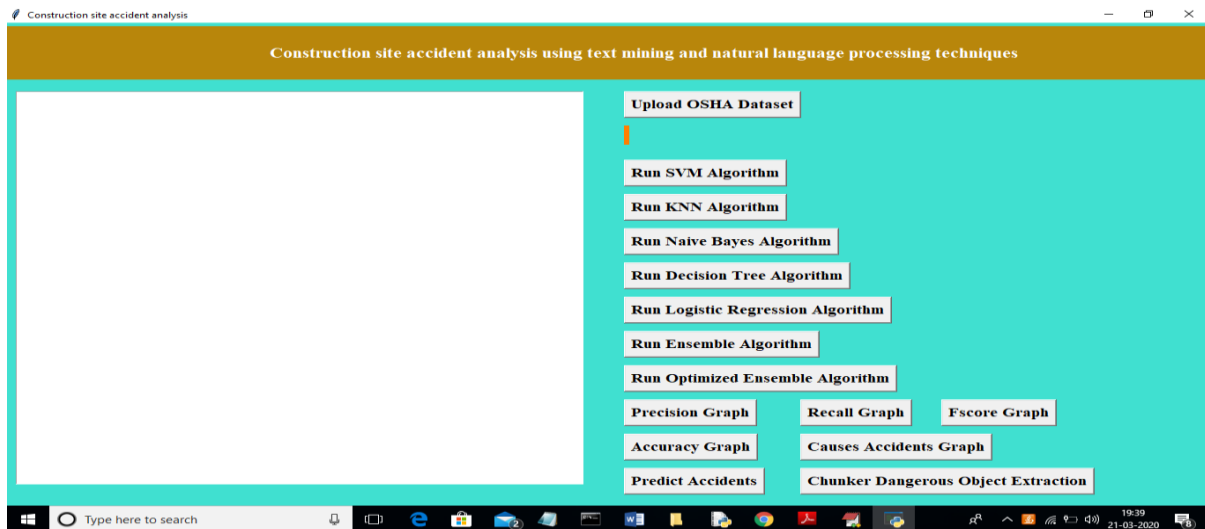
Text mining and natural language processing

Text mining, also referred to as text data mining, is defined as the process of deriving information from text data which is not previously known and not easy to be revealed [18]. It involves transforming text into numeric data which can be used in data mining algorithms then [19]. Natural language processing (NLP) involves the techniques of multiple areas in artificial intelligence, computational linguistics, mathematics and information science, it the approach to make computer understand

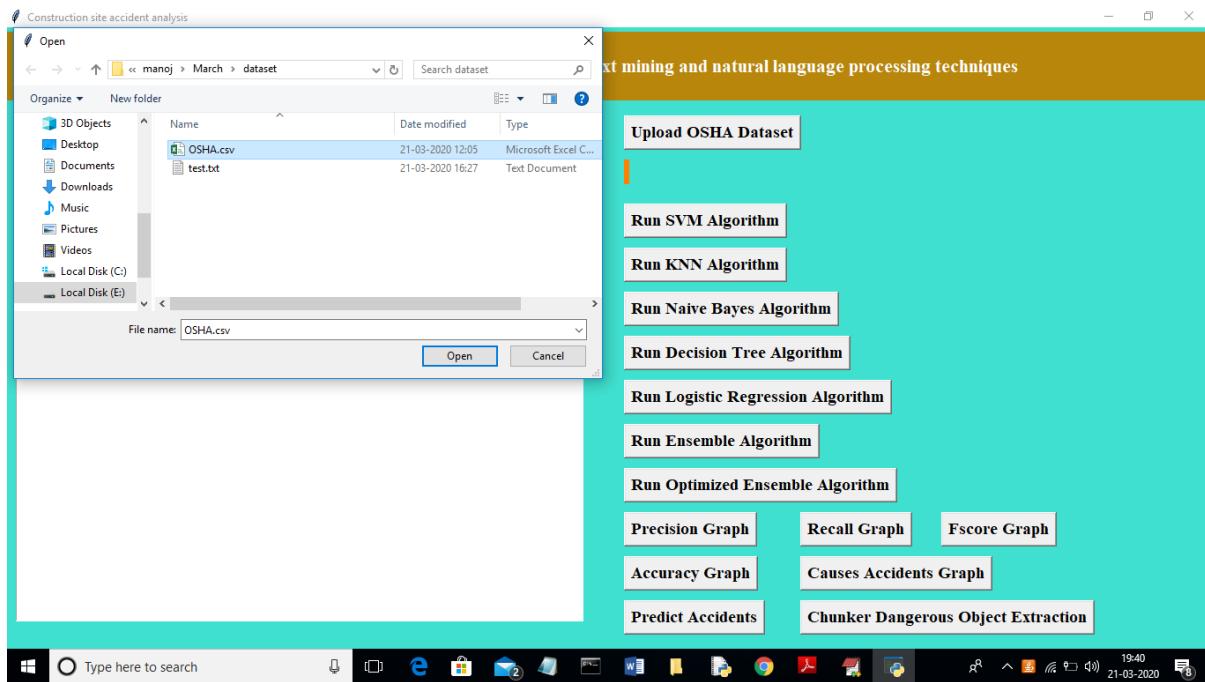
Experiment tools and data description

Two experiments are designed in this study. In the first experiment, an ensemble classifier is developed to classify the cause of construction accident while the second experiment is designed to identify common objects which cause the accident. Developing tool used is Python 2.7, main packages used for algorithms design are sklearn v0.19.1, pandas v0.22.0, nltk v3.2.5 and matplotlib v2.1.2 package for visualization. The original dataset from the Occupational Safety and Health Administration

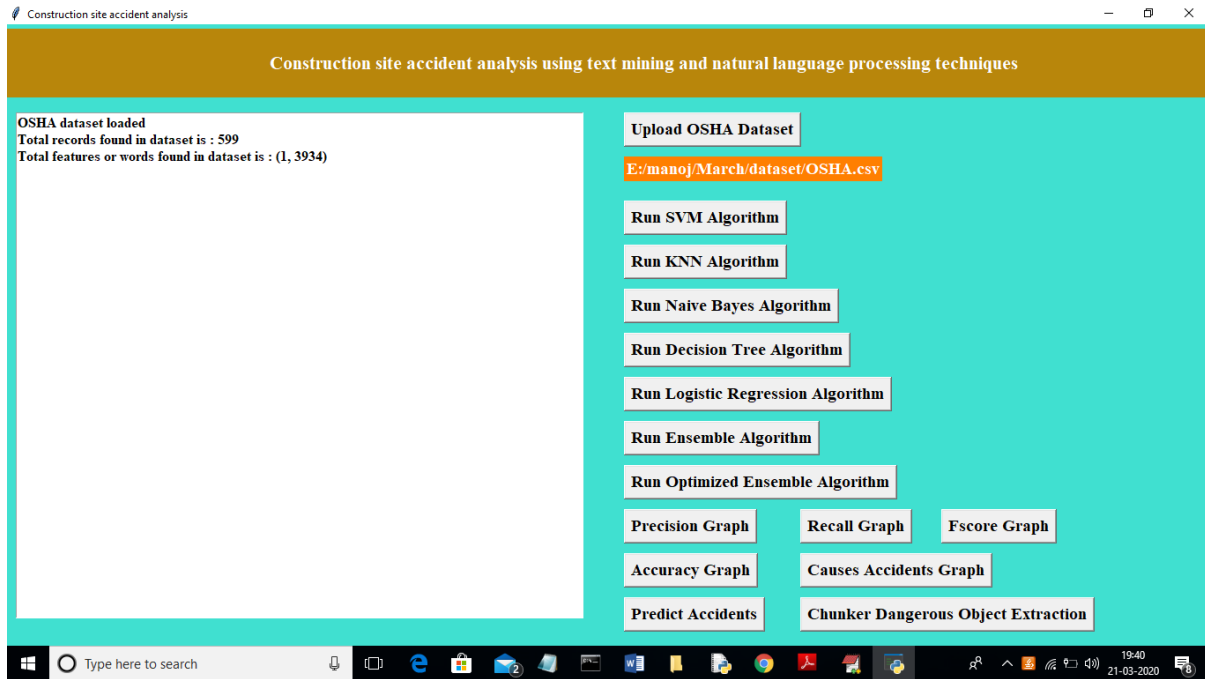
3. RESULTS



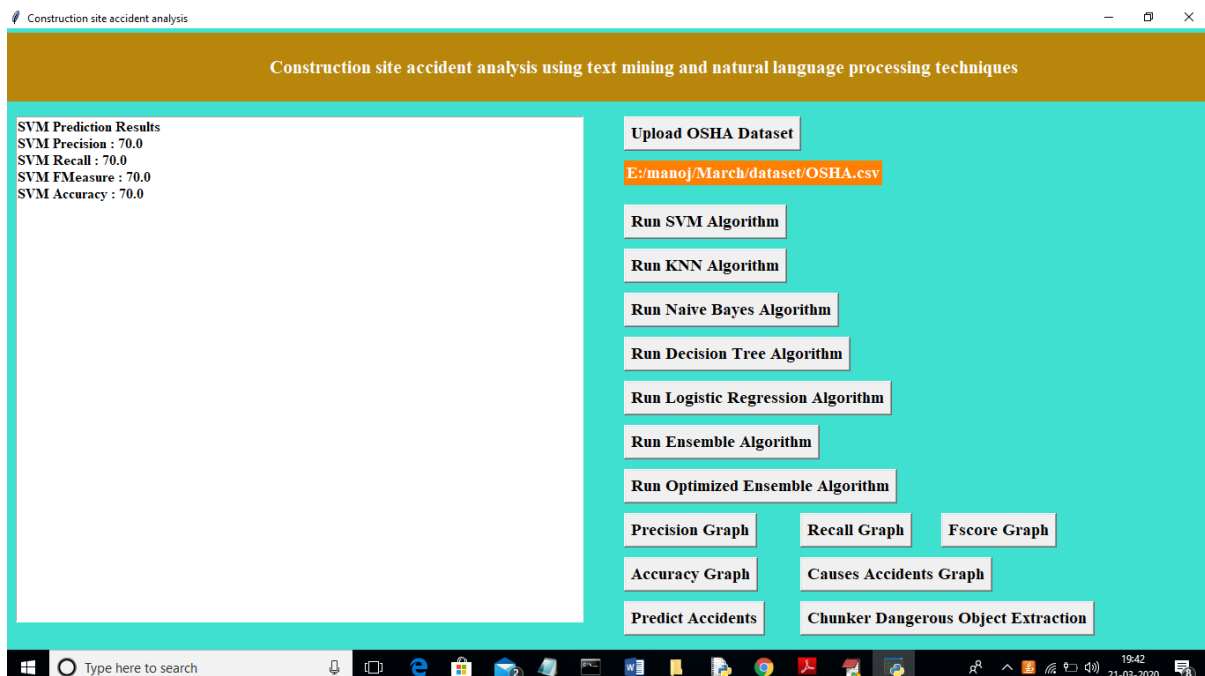
In above screen click on ‘Upload OSHA Dataset’ button and upload dataset



In above screen I am uploading ‘OSHA.csv’ dataset and after uploading dataset will get below screen



In above screen we can see dataset contains total 599 records and all records contains total 3934 words or features for vector. Now click on 'Run SVM Algorithm' button to build SVM model on uploaded dataset and calculate its prediction accuracy, precision etc.



4. CONCLUSION

We proposed an approach to automatically extract valid accident precursors from a dataset of

raw construction injury reports. Such information is highly valuable, as it can be used to better understand, predict, and prevent injury occurrence. For each of three supervised models (two of which being deep learning-based), we provided a methodology to identify (after training) the textual patterns that are, on average, the most predictive of each safety outcome. We verified that the learned precursors are valid and made several suggestions to improve the results. The proposed methods can also be used by the user to visualize and understand the models' predictions. Incidentally, while predictive skill is high for all models, we make the interesting observation that the simple TF-IDF + SVM approach is on par with (or outperforms) deep learning most of the time.

5. REFERENCES

- [1] D. Reinsel, J. Gantz, J. Rydning, Data age 2025. The digitization of the world: from edge to core, Tech. rep., Accessed 21st January 2020 (2018).
- [2] S. Grimes, Unstructured data and the 80 percent rule, Accessed 21st January 2020 (2008). URL <http://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/>
- [3] D. D. Woods, E. S. Patterson, E. M. Roth, Can we ever escape from data overload? A cognitive systems diagnosis, *Cognition, Technology & Work* 4 (1) (2002) 22–36. doi:10.1007/s101110200002.
- [4] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, G. Sethupathy, The age of analytics: competing in a data-driven world, Tech. rep., Accessed 21st January 2020 (2016). 30
- [5] D. Lukic, A. Littlejohn, A. Margaryan, A framework for learning from incidents in the workplace, *Safety Science* 50 (4) (2012) 950–957. doi:10.1016/j.ssci.2011.12.032.
- [6] J. M. Sanne, Incident reporting or storytelling? Competing schemes in a safety-critical and hazardous work setting, *Safety Science* 46 (8) (2008) 1205–1222. doi:10.1016/j.ssci.2007.06.024.
- [7] W. J. Wiatrowski, J. A. Janocha, Comparing fatal work injuries in the united states and the european union, Tech. rep., Accessed 21st January 2020 (June 2014). URL <https://www.bls.gov/opub/mlr/2014/article/comparing-fatal-work-injuries-us-eu.htm>

- [8] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016. URL <http://www.deeplearningbook.org>
- [9] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117. doi:10.1016/j.neunet.2014.09.003.
- [10] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324. doi:10.1109/5. 726791.
- [11] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of Machine Learning Research 3 (Feb) (2003) 1137–1155. doi:10.1007/ 3-540-33486-6_6.
- [12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
- [13] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint, arXiv:1301.3781. [14] Y. Kim, Convolutional neural networks for sentence classification, arXiv preprint, arXiv:1408.5882

CREDIT CARD FRAUD DETECTION USING AUTO ENCODER & DECODER

Karnati Bhargavi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y.Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract—Imbalanced data classification problem has always been a popular topic in the field of machine learning research. In order to balance the samples between majority and minority class. Oversampling algorithm is used to synthesize new minority class samples, but it could bring in noise. Pointing to the noise problems, this paper proposed a denoising autoencoder neural network (DAE) algorithm which can not only oversample minority class sample through misclassification cost, but it can denoise and classify the sampled dataset. Through experiments, compared with the denoising autoencoder neural network (DAE) with oversampling process and traditional fully connected neural networks, the results showed the proposed algorithm improves the classification accuracy of minority class of imbalanced datasets. **Keywords**—imbalanced data; oversampling; denoising autoencoder neural network; classification

1. INTRODUCTION

Credit card fraud is a growing threat with far reaching consequences in the finance industry, corporations and government. Fraud can be defined as criminal deception with intent of acquiring financial gain. As credit card became the most popular method of payment for both online and offline transaction, the fraud rate also accelerates. The main reasons for fraud is due to the lack of security, which involves the use of stolen credit card to get cash from bank through legitimate access. This results in high difficulty of preventing credit card fraud. So how to do fraud detection is very significant. A lot of researches have been proposed to the detection of such credit card fraud, which account for majority of credit card frauds. Detecting using traditional method is infeasible because of the big data. However, financial institutions have focused their attention to recent computational methodologies to handle credit card fraud problem. Classification problem is one of the key research topics in the field of machine learning. Currently available classification methods can only achieve preferable performance on balanced datasets. However, there are a large number of imbalanced datasets

in practical application. For the fraud problem, the minority class, which is the abnormal transaction, is more important [1]. For instance, when minority class accounts for less than 1 percent of the total dataset, the overall accuracy reaches more than 99% even though all the minority class has been misclassified. Minority class sampling is a common method to handle with the imbalanced data classification problem. The main purpose of oversampling is to increase the number of minority class samples so that the original classification information can get better retention. Therefore, in the fields where there is higher demand for the classification accuracy, oversampling algorithm is chosen in general. This paper seeks to implement credit card fraud detection using denoising autoencoder and oversampling. For imbalanced data, we decided use above method to achieve proper model.

2. RELATED WORKS

Data mining technique is one notable methods used in solving fraud detection problem. This is the process of identifying those transactions that are belong to frauds or not, which is based on the behaviors and habits of cardholder, many techniques have been applied to this area, artificial neural network [2], genetic algorithm, support vector machine, frequent item set mining, decision tree, migrating birds optimization algorithm, Naïve Bayes. A comparative analysis of logistic regression and Naïve Bayes is carried out in [3]. The performance of Bayesian and neural network [4] is evaluated on credit card fraud data. Decision tree, neural networks and logistic regression are tested for their applicability in fraud detections [5]. In a seminar work, [6] proposes two advanced data mining approaches, support vector machines and random forests, together with logistic regression, as part of an attempt to better detect credit card fraud while neural network and logistic regression is applied on credit card fraud detection problem [7]. A number of challenges are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legitimate ones; credit card transaction datasets are rarely available and highly imbalanced (or skewed); optimal feature (variables) selection for the models; suitable metric to evaluate performance of techniques on skewed credit card fraud data. Credit card fraud detection performance is greatly affected by type of sampling approach used, selection of variables and detection technique(s) used.

3. INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

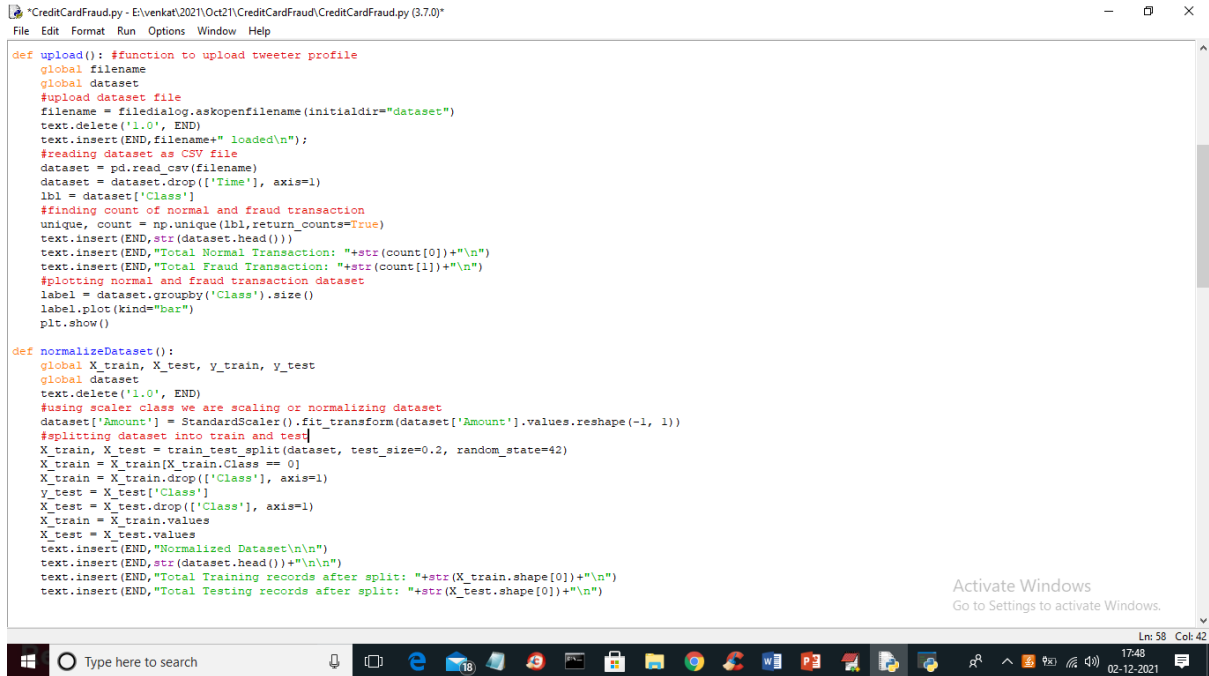
- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

4. RESULTS

Credit Card Fraud Detection using AutoEncoder & Decoder

In this project we are training Auto Encoder and Decode deep learning model on credit card dataset to predict normal and fraud transaction. To train model we have normalized the dataset and then split dataset into train and test and then by using TRAIN dataset we have trained AUTOENCODER and DECODER model.

After training model we have applied model on test data to calculate prediction accuracy and then plot MAE graph on normal and fraud transaction. Below code with red colour comments showing algorithm logic



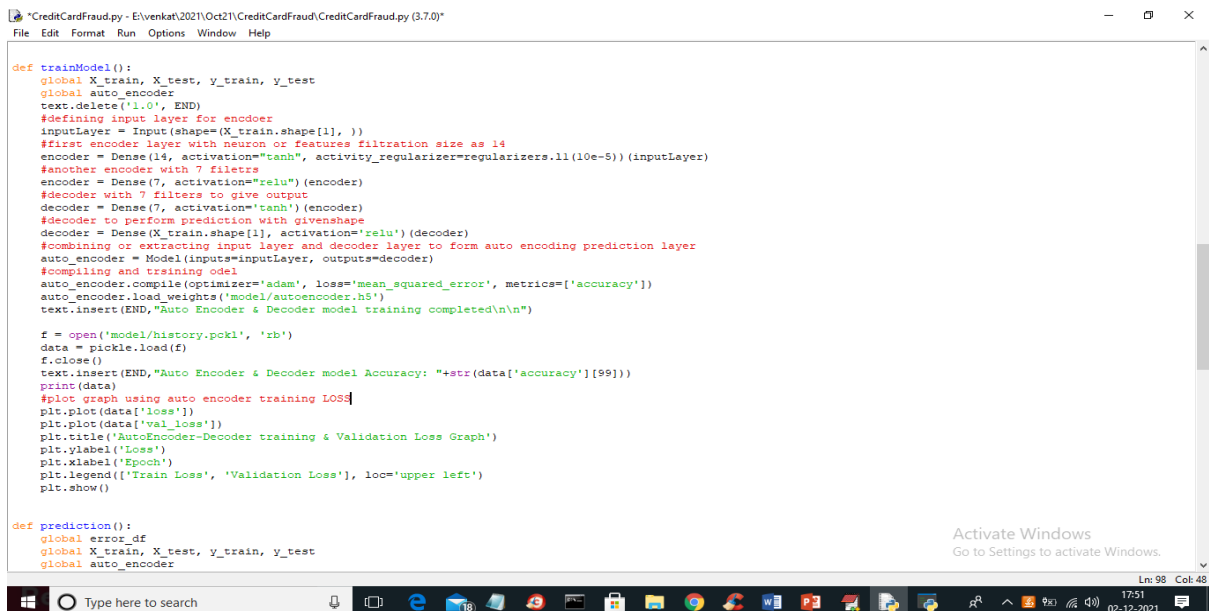
```

def upload(): #function to upload tweeter profile
    global filename
    global dataset
    #upload dataset file
    filename = filedialog.askopenfilename(initialdir="dataset")
    text.delete('1.0', END)
    text.insert(END, filename+" loaded\n");
    #reading dataset as CSV file
    dataset = pd.read_csv(filename)
    dataset = dataset.drop(['Time'], axis=1)
    lbl = dataset['Class']
    #finding count of normal and fraud transaction
    unique, count = np.unique(lbl, return_counts=True)
    text.insert(END, str(dataset.head()))
    text.insert(END, "Total Normal Transaction: "+str(count[0])+"\n")
    text.insert(END, "Total Fraud Transaction: "+str(count[1])+"\n")
    #plotting normal and fraud transaction dataset
    label = dataset.groupby('Class').size()
    label.plot(kind="bar")
    plt.show()

def normalizeDataset():
    global X_train, X_test, y_train, y_test
    global dataset
    text.delete('1.0', END)
    #using scaler class we are scaling or normalizing dataset
    dataset['Amount'] = StandardScaler().fit_transform(dataset['Amount'].values.reshape(-1, 1))
    #splitting dataset into train and test
    X_train, X_test = train_test_split(dataset, test_size=0.2, random_state=42)
    X_train = X_train[X_train.Class == 0]
    X_train = X_train.drop(['Class'], axis=1)
    y_train = X_train['Class']
    X_test = X_test.drop(['Class'], axis=1)
    X_train = X_train.values
    X_test = X_test.values
    text.insert(END, "Normalized Dataset\n\n")
    text.insert(END, str(dataset.head())+"\n\n")
    text.insert(END, "Total Training records after split: "+str(X_train.shape[0])+"\n")
    text.insert(END, "Total Testing records after split: "+str(X_test.shape[0])+"\n")

```

In above screen read red colour comments to know about logic



```

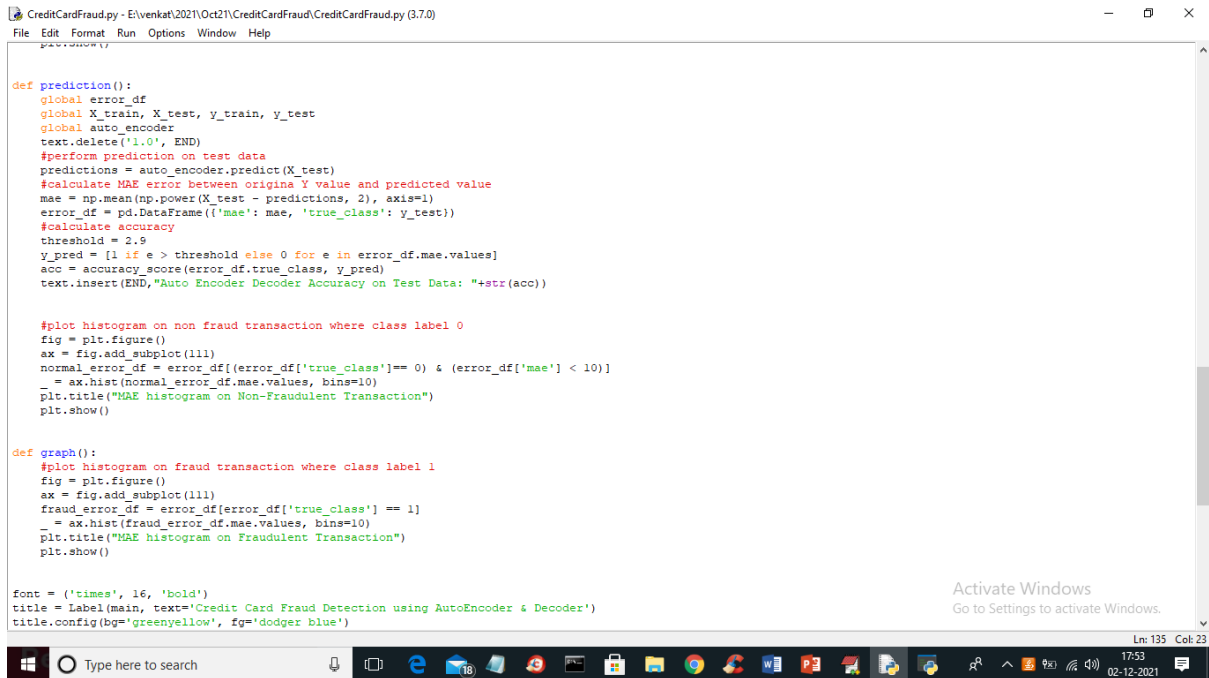
def trainModel():
    global X_train, X_test, y_train, y_test
    global auto_encoder
    text.delete('1.0', END)
    #defining input layer for encoder
    inputLayer = Input(shape=(X_train.shape[1],))
    #first encoder layer with neuron or features filtration size as 14
    encoder = Dense(14, activation="tanh", activity_regularizer=regularizers.l1(10e-5))(inputLayer)
    #another encoder with 7 filters
    encoder = Dense(7, activation="relu")(encoder)
    #decoder with 7 filters to give output
    decoder = Dense(7, activation="tanh")(encoder)
    #decoder to perform prediction with given shape
    decoder = Dense(X_train.shape[1], activation="relu")(decoder)
    #combining or extracting input layer and decoder layer to form auto encoding prediction layer
    auto_encoder = Model(inputs=inputLayer, outputs=decoder)
    #compiling and training model
    auto_encoder.compile(optimizer="adam", loss="mean_squared_error", metrics=['accuracy'])
    auto_encoder.load_weights("model/autoencoder-h5")
    text.insert(END, "Auto Encoder & Decoder model training completed\n\n")

    f = open('model/history.pkl', 'rb')
    data = pickle.load(f)
    f.close()
    text.insert(END, "Auto Encoder & Decoder model Accuracy: "+str(data['accuracy'])*100)
    print(data)
    #plot graph using auto encoder training LOSS
    plt.plot(data['loss'])
    plt.plot(data['val_loss'])
    plt.title("AutoEncoder-Decoder training & Validation Loss Graph")
    plt.ylabel('Loss')
    plt.xlabel('Epoch')
    plt.legend(['Train Loss', 'Validation Loss'], loc='upper left')
    plt.show()

def prediction():
    global error_df
    global X_train, X_test, y_train, y_test
    global auto_encoder

```

In above screen you can see code for encoder and decoder model and in below screen you can see prediction on test data and then calculating MAE values on both normal and fraud transaction



```

CreditCardFraud.py - E:\venkat\2021\Oct21\CreditCardFraud\CreditCardFraud.py (3.7.0)
File Edit Format Run Options Window Help

def prediction():
    global error_df
    global X_train, X_test, y_train, y_test
    global auto_encoder
    text.delete('1.0', END)
    #perform prediction on test data
    predictions = auto_encoder.predict(X_test)
    #calculate MAE error between origina Y value and predicted value
    mae = np.mean(np.power(X_test - predictions, 2), axis=1)
    error_df = pd.DataFrame({'mae': mae, 'true_class': y_test})
    #calculate accuracy
    threshold = 2.9
    y_pred = [1 if e > threshold else 0 for e in error_df.mae.values]
    acc = accuracy_score(error_df.true_class, y_pred)
    text.insert(END, "Auto Encoder Decoder Accuracy on Test Data: "+str(acc))

    #plot histogram on non fraud transaction where class label 0
    fig = plt.figure()
    ax = fig.add_subplot(111)
    normal_error_df = error_df[(error_df['true_class']== 0) & (error_df['mae'] < 10)]
    _ = ax.hist(normal_error_df.mae.values, bins=10)
    plt.title("MAE histogram on Non-Fraudulent Transaction")
    plt.show()

def graph():
    #plot histogram on fraud transaction where class label 1
    fig = plt.figure()
    ax = fig.add_subplot(111)
    fraud_error_df = error_df[error_df['true_class'] == 1]
    _ = ax.hist(fraud_error_df.mae.values, bins=10)
    plt.title("MAE histogram on Fraudulent Transaction")
    plt.show()

font = ('times', 16, 'bold')
title = Label(main, text='Credit Card Fraud Detection using AutoEncoder & Decoder')
title.config(bg='greenyellow', fg='dodger blue')
Ln: 135 Col: 23

```

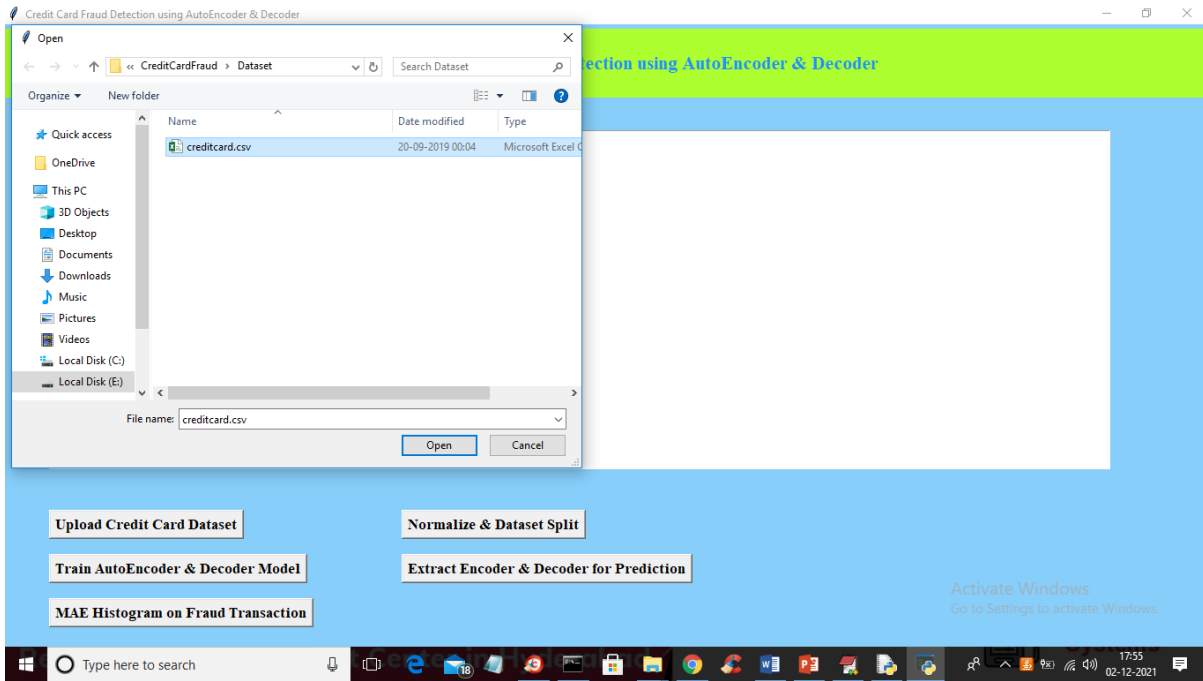
#in above screen you can see coding for prediction using encoder model and then calculate MAE and accuracy and then plot graph for both normal and fraud MAE.

SCREEN SHOTS

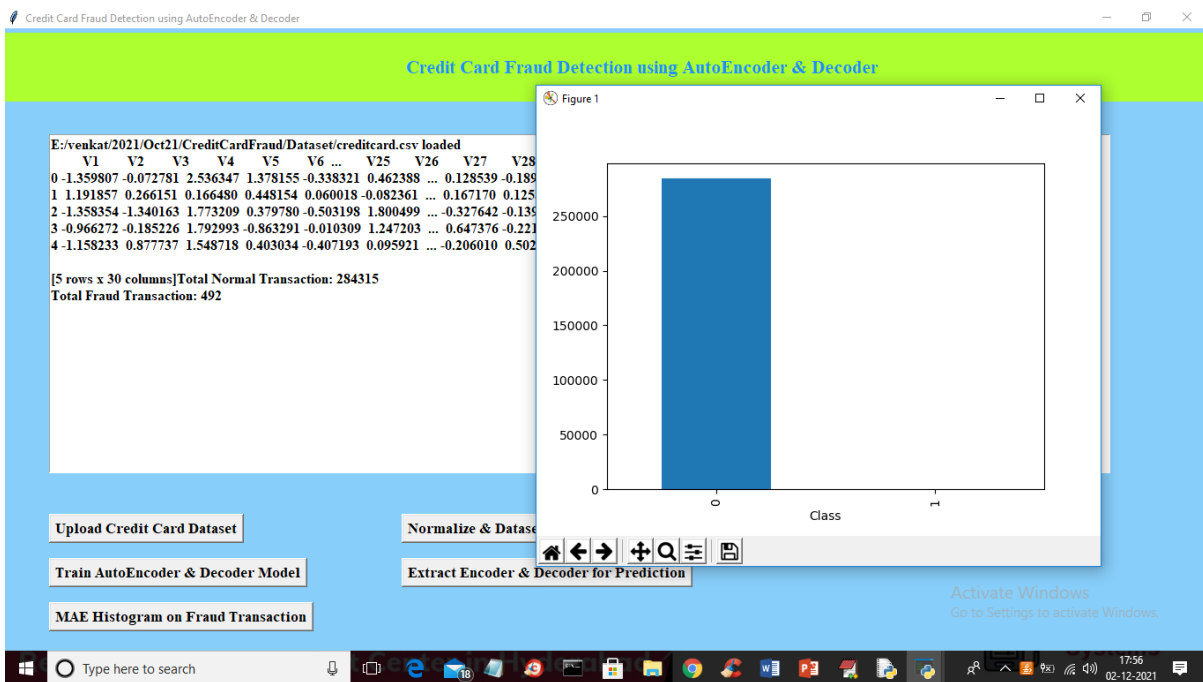
To run project double click on 'run.bat' file to get below screen



In above screen click on 'Upload Credit Card Dataset' button to upload dataset



In above screen selecting and uploading credit card dataset and then click on ‘Open’ button to load dataset and to get below screen



In above screen we can see dataset loaded and we can see total records in normal and fraud transaction and in graph also we can see very few transactions are available in fraud category and now close above graph and then click on ‘Normalize and Dataset split’ button to normalize dataset and to split into train and test

5. CONCLUSION

In machine learning area, imbalance data classification receives increasing attention as big data become popular. On account of the drawbacks of traditional method, oversampling algorithm and autoencoder can be used. This study combined stacked denoising autoencoder neural network with oversampling to build the model, which can achieve minority class sampling on the basis of misclassification cost, and denoise and classify the sampled datasets. The proposed algorithm increases classification accuracy of minority class compared to the former methods, we can achieve different accuracy by controlling the threshold. In this study, when threshold equal to 0.6, we can achieve the best performance, which is 97.93%. However, the dimensionality reduction of high-dimensional data still need to be further researched.

6. REFERENCES

- [1] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, vol. 40, pp. 5916-5923, 2013.
- [2] Ogwueleka, F. N., (2011). Data Mining Application in Credit Card Fraud Detection System, *Journal of Engineering Science and Technology*, Vol. 6, No. 3, pp. 311 – 322
- [3] Ng, A. Y., and Jordan, M. I., (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 2, 841- 848.
- [4] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit card fraud detection using Bayesian and neural networks. In *Proceedings of the 1st international nairo congress on neuro fuzzy technologies* (pp. 261-270).
- [5] Shen, A., Tong, R., & Deng, Y. (2007). Application of classification models on credit card fraud detection. In *Service Systems and Service Management, 2007 International Conference on* (pp. 1-4). IEEE.
- [6] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3), 602-613.
- [7] Sahin, Y. and Duman, E., (2011). Detecting credit card fraud by ANN and logistic regression. In *Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on* (pp. 315-319). IEEE.

[8] Autoencoder for Words, Liou, C.-Y., Cheng, C.-W., Liou, J.-W., and Liou, D.-R., *Neurocomputing*, Volume 139, 84–96 (2014), doi:10.1016/j.neucom.2013.09.055

[9] M. Koziarski and M. Woźniak, "CCR: A combined cleaning and resampling algorithm for imbalanced data classification", *International Journal of Applied Mathematics and Computer Science*, vol. 27, no. 4, 2017.



STRESS DETECTION IN IT PROFESSIONALS BY IMAGE PROCESSING AND MACHINE LEARNING

Karra M V V S Sai Pavan (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract The main motive of our project is to detect stress in the IT professionals using vivid Machine learning and Image processing techniques. Our system is an upgraded version of the old stress detection systems which excluded the live detection and the personal counseling but this system comprises of live detection and periodic analysis of employees and detecting physical as well as mental stress levels in his/her by providing them with proper remedies for managing stress by providing survey form periodically. Our system mainly focuses on managing stress and making the working environment healthy and spontaneous for the employees and to get the best out of them during working hours.

1. INTRODUCTION

Stress management systems play a significant role to detect the stress levels which disrupts our socio economic lifestyle. As World Health Organization (WHO) says, Stress is a mental health problem affecting the life of one in four citizens. Human stress leads to mental as well as socio-fiscal problems, lack of clarity in work, poor working relationship, depression and finally commitment of suicide in severe cases. This demands counselling to be provided for the stressed individuals cope up against stress. Stress avoidance is impossible but preventive actions helps to overcome the stress. Currently, only medical and physiological experts can determine whether one is under depressed state (stressed) or not. One of the traditional method to detect stress is based on questionnaire. This method completely depends on the answers given by the individuals, people will be tremulous to say whether they are stressed or normal. Automatic detection of stress minimizes the risk of health issues and improves the welfare of the society.

This paves the way for the necessity of a scientific tool, which uses physiological signals thereby automating the detection of stress levels in individuals. Stress detection is discussed in various literatures as it is a significant societal contribution that enhances the lifestyle of individuals. Ghaderi et al. analysed stress using Respiration, Heart rate (HR), facial electromyography (EMG), Galvanic skin response (GSR) foot and GSR hand data with a conclusion that, features pertaining to respiration process are substantial in stress detection. Maria Viqueira et al. describes mental stress prediction using a standalone stress sensing hardware by interfacing GSR as the only physiological sensor . David Liu et al. proposed a research to predict stress levels solely from Electrocardiogram (ECG). Multimodal sensor efficacy to detect stress of working people is experimentally discussed in . This employs the sensor data from sensors such as pressure distribution, HR, Blood Volume Pulse (BVP) and Electrodermal activity (EDA). An eye tracker sensor is also used which systematically analyses



the eye movements with the stressors like Stroop word test and information related to pickup tasks. The authors of performed perceived stress detection by a set of non-invasive sensors which collects the physiological signals such as ECG, GSR, Electroencephalography (EEG), EMG, and Saturation of peripheral oxygen (SpO₂). Continuous stress levels are estimated using the physiological sensor data such as GSR, EMG, HR, Respiration in. The stress detection is carried out effectively using Skin conductance level (SCL), HR, Facial EMG sensors by creating ICT related Stressors. Automated stress detection is made possible by several pattern recognition algorithms. Every sensor data is compared with a stress index which is a threshold value used for detecting the stress level. The authors of collected data from 16 individuals under four stressor conditions which were tested with Bayesian Network, J48 algorithm and Sequential Minimal Optimization (SMO) algorithm for predicting stress. Statistical features of heart rate, GSR, frequency domain features of heart rate and its variability (HRV), and the power spectral components of ECG were used to govern the stress levels. Various features are extracted from the commonly used physiological signals such as ECG, EMG, GSR, BVP etc., measured using appropriate sensors and selected features are grouped into clusters for further detection of anxiety levels. In, it is concluded that smaller clusters result in better balance in stress detection using the selected General Regression Neural Network (GRNN) model. This results in the fact that different combinations of the extracted features from the sensor signals provide better solutions to predict the

continuous anxiety level. Frequency domain features like LF power (low frequency power from 0.04 Hz to 0.15Hz), HF power (High frequency power from 0.15Hz to 0.4 Hz), LF/HF (ratio of LF to the HF), and time domain features like Mean, Median, standard deviation of heart signal are considered for continuous real time stress detection in. Classification using decision tree such as PLDA is performed using two stressors namely pickup task and stroop based word test wherein the authors concluded that the stressor based classification proves unsatisfactory. In 2016, Gjoreski et al. created laboratory based stress detection classifiers from ECG signal and HRV features. Features of ECG are analysed using GRNN model to measure the stress level. Heart rate variability (HRV) features and RR (cycle length variability interval length between two successive R_s) interval features are used to classify the stress level. It is noticed that Support Vector Machine (SVM) was used as the classification algorithm predominantly due to its generalization ability and sound mathematical background. Various kernels were used to develop models using SVM and it is concluded in that a linear SVM on both ECG frequency features and HRV features performed best, outperforming other model choices.

Nowadays as IT industries are setting a new peek in the market by bringing new technologies and products in the market. In this study, the stress levels in employees are also noticed to raise the bar high. Though there are many organizations who provide mental health related schemes for their employees but the issue is far from control. In this paper we try to go in the depth of this problem by trying to detect



the stress patterns in the working employee in the companies we would like to apply image processing and machine learning techniques to analyze stress patterns and to narrow down the factors that strongly determine the stress levels. Machine Learning algorithms like KNN classifiers are applied to classify stress. Image Processing is used at the initial stage for detection, the employee's image is clicked by the camera which serves as input. In order to get an enhanced image or to extract some useful information from it image processing is used by converting image into digital form and performing some operations on it. By taking input as an image from video frames and output may be image or characteristics associated with that image. Image processing basically includes the following three steps:

- Importing the image via image acquisition tools.
- Analyzing and manipulating the image.
- Output in which result is altered image or report that is based on image analysis.

System gets the ability to automatically learn and improve from self-experiences without being explicitly programmed using Machine learning which is an application of artificial intelligence (AI). Computer programs are developed by Machine Learning that can access data and use it to learn for themselves. Explicit programming to perform the task based on predictions or decisions builds a mathematical model based on "training data" by using Machine Learning. The extraction of hidden data, association of image data and additional pattern which are unclearly visible in image is done using Image Mining. It's an interrelated field that involves, Image Processing, Data

Mining, Machine Learning and Datasets. According to conservative estimates in medical books, 50- 80% of all physical diseases are caused by stress. Stress is believed to be the principal cause in cardiovascular diseases. Stress can place one at higher risk for diabetes, ulcers, asthma, migraine headaches, skin disorders, epilepsy, and sexual dysfunction. Each of these diseases, and host of others, is psychosomatic (i.e., either caused or exaggerated by mental conditions such as stress) in nature. Stress has three prong effects:

- Subjective effects of stress include feelings of guilt, shame, anxiety, aggression or frustration. Individuals also feel tired, tense, nervous, irritable, moody, or lonely.
- Visible changes in a person's behavior are represented by Behavioral effects of stress. Effects of behavioral stress are seen such as increased accidents, use of drugs or alcohol, laughter out of context, outlandish or argumentative behavior, very excitable moods, and/or eating or drinking to excess.
- Diminishing mental ability, impaired judgment, rash decisions, forgetfulness and/or hypersensitivity to criticism are some of the effects of Cognitive stress

2. LITERATURE SURVEY

1) Stress and anxiety detection using facial cues from videos

AUTHORS: G. Giannakakis, D. Manousos, F. Chiarugi

This study develops a framework for the detection and analysis of stress/anxiety emotional states through video-recorded facial cues. A thorough experimental protocol was established to induce



systematic variability in affective states (neutral, relaxed and stressed/anxious) through a variety of external and internal stressors. The analysis was focused mainly on non-voluntary and semi-voluntary facial cues in order to estimate the emotion representation more objectively. Features under investigation included eye-related events, mouth activity, head motion parameters and heart rate estimated through camera-based photoplethysmography. A feature selection procedure was employed to select the most robust features followed by classification schemes discriminating between stress/anxiety and neutral states with reference to a relaxed state in each experimental phase. In addition, a ranking transformation was proposed utilizing self reports in order to investigate the correlation of facial parameters with a participant perceived amount of stress/anxiety. The results indicated that, specific facial cues, derived from eye activity, mouth activity, head movements and camera based heart activity achieve good accuracy and are suitable as discriminative indicators of stress and anxiety.

2) Detection of Stress Using Image Processing and Machine Learning Techniques

AUTHORS: Nisha Raichur, Nidhi Lonakadi, Priyanka Mural

Stress is a part of life it is an unpleasant state of emotional arousal that people experience in situations like working for long hours in front of computer. Computers have become a way of life, much life is spent on the computers and hence we are therefore more affected by the ups and downs that they cause us. One

cannot just completely avoid their work on computers but one can at least control his/her usage when being alarmed about him being stressed at certain point of time. Monitoring the emotional status of a person who is working in front of a computer for longer duration is crucial for the safety of a person. In this work a real-time non-intrusive videos are captured, which detects the emotional status of a person by analysing the facial expression. We detect an individual emotion in each video frame and the decision on the stress level is made in sequential hours of the video captured. We employ a technique that allows us to train a model and analyze differences in predicting the features. Theano is a python framework which aims at improving both the execution time and development time of the linear regression model which is used here as a deep learning algorithm. The experimental results show that the developed system is well on data with the generic model of all ages.

3) Machine Learning Techniques for Stress Prediction in Working Employees **AUTHORS : U. S. Reddy, A. V. Thota and A. Dharun**

Stress disorders are a common issue among working IT professionals in the industry today. With changing lifestyle and work cultures, there is an increase in the risk of stress among the employees. Though many industries and corporates provide mental health related schemes and try to ease the workplace atmosphere, the issue is far from control. In this paper, we would like to apply machine learning techniques to analyze stress patterns in working adults and to narrow down the



factors that strongly determine the stress levels. Towards this, data from the OSMI mental health survey 2017 responses of working professionals within the tech-industry was considered. Various Machine Learning techniques were applied to train our model after due data cleaning and preprocessing. The accuracy of the above models was obtained and studied comparatively. Boosting had the highest accuracy among the models implemented. By using Decision Trees, prominent features that influence stress were identified as gender, family history and availability of health benefits in the workplace. With these results, industries can now narrow down their approach to reduce stress and create a much comfortable workplace for their employees.

4) Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ECG

AUTHORS : Tanev, G., Saadi, D.B., Hoppe, K., Sorensen, H.B

Chronic stress detection is an important factor in predicting and reducing the risk of cardiovascular disease. This work is a pilot study with a focus on developing a method for detecting short-term psychophysiological changes through heart rate variability (HRV) features. The purpose of this pilot study is to establish and to gain insight on a set of features that could be used to detect psychophysiological changes that occur during chronic stress. This study elicited four different types of arousal by images, sounds, mental tasks and rest, and classified them using linear and non-linear HRV features from electrocardiograms (ECG) acquired by the wireless wearable

ePatch® recorder. The highest recognition rates were acquired for the neutral stage (90%), the acute stress stage (80%) and the baseline stage (80%) by sample entropy, detrended fluctuation analysis and normalized high frequency features. Standardizing non-linear HRV features for each subject was found to be an important factor for the improvement of the classification results.

5) HealthyOffice: Mood recognition at work using smartphones and wearable sensors

AUTHORS: Zenonos, A., Khan, A., Kalogridis, G., Vatsikas, S., Lewis, T., Sooriyabandara

Stress, anxiety and depression in the workplace are detrimental to human health and productivity with significant financial implications. Recent research in this area has focused on the use of sensor technologies, including smartphones and wearables embedded with physiological and movement sensors. In this work, we explore the possibility of using such devices for mood recognition, focusing on work environments. We propose a novel mood recognition framework that is able to identify five intensity levels for eight different types of moods every two hours. We further present a smartphone app ('HealthyOffice'), designed to facilitate self-reporting in a structured manner and provide our model with the ground truth. We evaluate our system in a small-scale user study where wearable sensing data is collected in an office environment. Our experiments exhibit promising results allowing us to reliably recognize various classes of perceived moods.

3. SCREEN SHOTS

Home page:



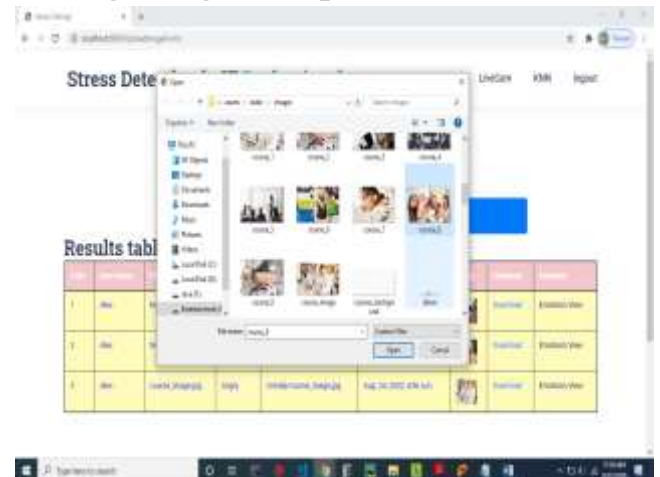
User Home Page:



User Register page:



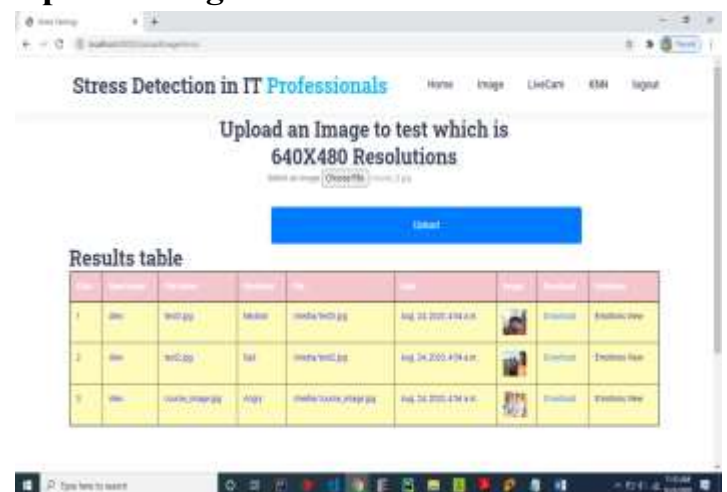
Giving Image as Input:



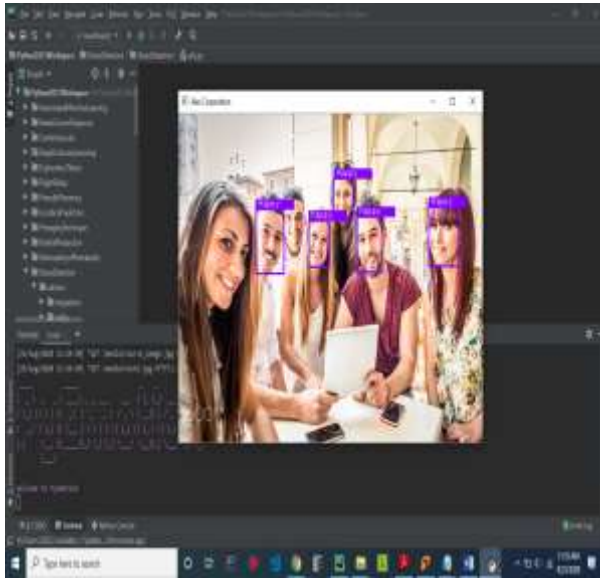
User Login Form:



Upload Image:



Response Image:

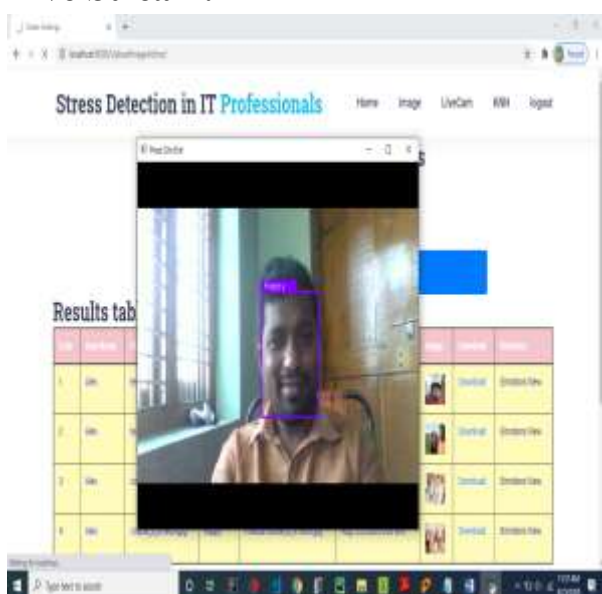


Results:



ID	Name	Gender	Age	Stress Level	Stress Type	Stress Cause
1	John	Male	35	High	Work	Work
2	Jane	Female	28	Medium	Work	Work
3	Mike	Male	42	Low	Work	Work
4	Sarah	Female	30	High	Work	Work

Live Stream:



4. CONCLUSION

Stress Detection System is designed to predict stress in the employees by monitoring captured images of authenticated users which makes the system secure. The image capturing is done automatically when the authenticate user is logged in based on some time interval. The captured images are used to detect the stress of the user based on some standard conversion and image processing mechanisms. Then the system will analyze the stress levels by using Machine Learning algorithms which generates the results that are more efficient.

5. REFERENCES

- [1] G. Giannakakis, D. Manousos, F. Chiarugi, "Stress and anxiety detection using facial cues from videos," *Biomedical Signal processing and Control*, vol. 31, pp. 89-101, January 2017.
- [2] T. Jick and R. Payne, "Stress at work," *Journal of Management Education*, vol. 5, no. 3, pp. 50-56, 1980.
- [3] Nisha Raichur, Nidhi Lonakadi, Priyanka Mural, "Detection of Stress Using Image Processing and Machine Learning Techniques", vol.9, no. 3S, July 2017.
- [4] Bhattacharyya, R., & Basu, S. (2018). Retrieved from 'The Economic Times'.
- [5] OSMI Mental Health in Tech Survey Dataset, 2017
- [6] U. S. Reddy, A. V. Thota and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICIC), Madurai, India, 2018, pp. 1-4.



- [7] <https://www.kaggle.com/qiriro/stress>
- [8] Communications, N.. World health report. 2001. URL: http://www.who.int/whr/2001/media_centre/press_release/en/.
- [9] Bakker, J., Holenderski, L., Kocielnik, R., Pechenizkiy, M., Sidorova, N.. Stess@work: From measuring stress to its understanding, prediction and handling with personalized coaching. In: Proceedings of the 2nd ACM SIGHT International health informatics symposium. ACM; 2012, p. 673–678.
- [10] Deng, Y., Wu, Z., Chu, C.H., Zhang, Q., Hsu, D.F.. Sensor feature selection and combination for stress identification using combinatorial fusion. International Journal of Advanced Robotic Systems 2013;10(8):306.

NETSPAM A NETWORK-BASED SPAM DETECTION FRAMEWORK FOR REVIEWS IN ONLINE SOCIAL MEDIA

Karri Naga Venkata Gowtham Reddy (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Y. Srinivasa Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

ABSTRACT

Nowadays, a big part of people rely on available content in social media in their decision for example, reviews and feedback on a topic or product. The possibility that anybody can leave a review provide a golden opportunity for spammers to write spam reviews about products and services for different interests.

Identifying these spammers and the spam content is a hot topic of research and although a considerable number of studies have been done recently toward this end, but so far the methodologies put forth still barely detect spam reviews, and none of them show the importance of each extracted feature type.

In this study, we propose a novel framework, named NetSpam, which utilizes spam features for modeling review datasets as heterogeneous information networks to map spam detection procedure into a classification problem in such networks.

Using the importance of spam features help us to obtain better results in terms of different metrics experimented on real-world review datasets from Yelp and Amazon websites.

The results show that NetSpam outperforms the existing methods and among four categories of features; including review-behavioral, user-behavioral, review linguistic, user-linguistic, the first type of features performs better than the other categories.

1.INTRODUCTION

Online Social Media portals play an influential role in information propagation which is considered as an important source for producers in their advertising campaigns as well as for customers in selecting products and services. In the past years, people rely a lot on the written reviews in their decision-making processes, and positive/negative reviews encouraging/discouraging them in their selection of products and services. In addition,

written reviews also help service providers to enhance the quality of their products and services. These reviews thus have become an important factor in success of a business while positive reviews can bring benefits for a company, negative reviews can potentially impact credibility and cause economic losses. The fact that anyone with any identity can leave comments as review, provides a tempting opportunity for spammers to write fake reviews designed to mislead users' opinion. These misleading reviews are then multiplied by the sharing function of social media and propagation over the web. The reviews written to change users' perception of how good a product or a service are considered as spam [11], and are often written in exchange for money. As shown in [1], 20% of the reviews in the Yelp website are actually spam reviews. On the other hand, a considerable amount of literature has been published on the techniques used to identify spam and spammers as well as different type of analysis on this topic [30], [31]. These techniques can be classified into different

categories; some using linguistic patterns in text [2], [3], [4], which are mostly based on bigram, and unigram, others are based on behavioral patterns that rely on features extracted from patterns in users' behavior which are mostly metadatabased

[34], [6], [7], [8], [9], and even some techniques using graphs and graph-based algorithms and classifiers [10], [11], [12]. Despite this great deal of efforts, many aspects have been missed or remained unsolved. One of them is a classifier that can calculate feature weights that show each feature's level of importance in determining spam reviews. The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network (HIN) [19] and to map the problem of spam detection into a HIN classification problem. In particular, we model review dataset as a HIN in which reviews are connected through different node types (such as features and users). A weighting algorithm is then employed to calculate each feature's importance (or weight). These weights are utilized to calculate the final labels for reviews using both unsupervised and supervised approaches.

To evaluate the proposed solution, we used two sample review datasets from Yelp and Amazon websites. Based on our observations, defining two views for features (review-user and behavioral-linguistic), the classified features as review behavioral

have more weights and yield better performance on spotting spam reviews in both semi-supervised and unsupervised approaches. In addition, we demonstrate that using different supervisions such as 1%, 2.5% and 5% or using an unsupervised approach, make no

noticeable variation on the performance of our approach. We observed that feature weights can be added or removed for labeling and hence time

complexity can be scaled for a specific level of accuracy. As the result of this weighting step, we can use fewer features with more weights to obtain better accuracy with less time complexity. In addition, categorizing features in four major categories (review-behavioral, user-behavioral, reviewlinguistic, user-linguistic), helps us to understand how much each category of features is contributed to spam detection. In summary, our main contributions are as follows:

(i) We propose NetSpam framework that is a novel networkbased approach which models review networks as heterogeneous information networks. The classification step uses IEEE Transactions on Information Forensics and Security, Volume:12, Issue:7, Issue Date: July.2017 2 different metapath types which are innovative in the spam detection domain.

2.EXISTING SYSTEM

The results show that NetSpam outperforms the existing methods and among four categories of features; including review-behavioral, use behavioral, review linguistic, user linguistic, the first type of features performs better than the other categories.

Despite this great deal of efforts, many aspects have been missed or remained unsolved. One of them is a classifier that can calculate feature weights that show each feature's level of importance in determining spam reviews. The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network (HIN) and to map the problem of spam detection into a HIN classification problem. In particular, we model review dataset as a HIN in which reviews are connected through different node types. The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network and to map the problem of spam detection into a HIN classification problem. In particular, we model review dataset as in which reviews are connected through different node types. **A weighting algorithm** is then employed to calculate each feature's importance. These weights are utilized to calculate the final labels for reviews using both unsupervised and supervised approaches.

DISADVANTAGE:

This utilizes spam features for modeling review datasets as heterogeneous information networks to map spam detection procedure into a classification problem in such networks.

Time Complexity.

3.PROPOSED SYSTEM

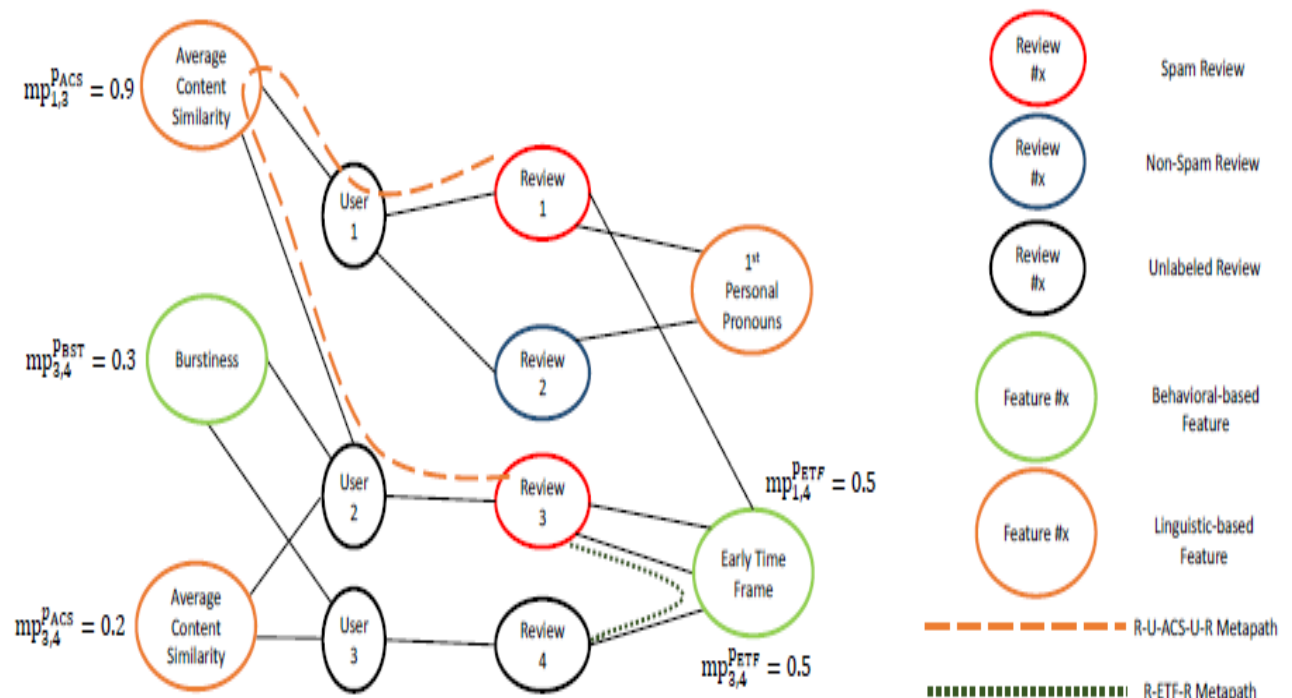
NetSpam is able to find features importance even without ground truth, and only by relying on metapath definition and based on values calculated for each review. NetSpam improves the accuracy compared to the stateof- the art in terms of time complexity, which highly depends to the number of features used to identify a spam review; hence, using features with more weights will resulted in detecting fake reviews easier with less time complexity.

A new **Content Based Algorithm** for spam features is proposed to determine the relative importance of each feature and shows how effective each of features are in identifying spams from normal reviews.

ADVANTAGE:

To identify spam and spammers as well as different type of analysis on this topic. Written reviews also help service providers to enhance the quality of their products and services.

ARCHITECTURE:



PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

Request Clarification

Feasibility Study

Request Approval

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

Operational Feasibility

Economic Feasibility

Technical Feasibility

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

4.CONCLUSION

This study introduces a novel spam detection framework namely NetSpam based on a metapath concept as well as IEEE Transactions on Information Forensics and Security, Volume:12, Issue:7, Issue Date: July.2017 10 a new graph-based method to label reviews relying on a rank-based labeling approach. The performance of the proposed framework is evaluated by using two real-world labeled datasets of Yelp and Amazon websites. Our observations show that calculated weights by using this metapath concept can be very effective in identifying spam reviews and leads to a better performance. In addition, we found that even without a train set, NetSpam can calculate the importance of each feature

and it yields better performance in the features' addition process, and performs better than previous works, with only a small number of features. Moreover, after defining four main categories for features our observations show that the reviews behavioral category performs better than other categories, in terms of AP, AUC as well as in the calculated weights. The results also confirm that using different supervisions, similar to the semi-supervised method, have no noticeable effect on determining most of the weighted features, just as in different datasets.

For future work, metapath concept can be applied to other problems in this field. For example, similar framework can be used to find spammer communities. For finding community, reviews can be connected through group spammer features (such as the proposed feature in [29]) and reviews with highest similarity based on metapath concept are known as communities. In addition, utilizing the product features is an interesting future work on this study as we used features more related to spotting spammers and spam reviews. Moreover, while single networks has received considerable attention from various disciplines for over a decade, information diffusion and content sharing in multilayer networks is still a young

5.REFERENCES

- [1] J. Donfro, A whopping 20 % of yelp reviews are fake. <http://www.businessinsider.com/20-percent-of-yelp-reviews-fake-2013-9>. Accessed: 2015-07-30.
- [2] M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In ACM WWW, 2012.
- [3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In ACL, 2011.
- [4] Ch. Xu and J. Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. In SIAM International Conference on Data Mining, 2014.
- [5] N. Jindal and B. Liu. Opinion spam and analysis. In WSDM, 2008.
- [6] F. Li, M. Huang, Y. Yang, and X. Zhu. Learning to identify review spam. Proceedings of the 22nd International Joint Conference on Artificial Intelligence; IJCAI, 2011.
- [7] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.

- [8] A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos. Trueview: Harnessing the power of multiple review sites. In ACM WWW, 2015.
- [9] B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In USENIX, 2014.
- [10] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective PU learning. In ICDM, 2014.

PREDICTING DRUG-DRUG INTERACTIONS BASED ON INTEGRATED SIMILARITY AND SEMI-SUPERVISED LEARNING

Karumuri Jayasri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

A drug-drug interaction (DDI) is defined as an association between two drugs where the pharmacological effects of a drug are influenced by another drug. Positive DDIs can usually improve the therapeutic effects of patients, but negative DDIs cause the major cause of adverse drug reactions and even result in the drug withdrawal from the market and the patient death. Therefore, identifying DDIs has become a key component of the drug development and disease treatment. In this study, we propose a novel method to predict DDIs based on the integrated similarity and semi-supervised learning (DDI-IS-SL). DDI-IS-SL integrates the drug chemical, biological and phenotype data to calculate the feature similarity of drugs with the cosine similarity method. The Gaussian Interaction Profile kernel similarity of drugs is also calculated based on known DDIs. A semi-supervised learning method (the Regularized Least Squares classifier) is used to calculate the interaction possibility scores of drug-drug pairs. In terms of the 5-fold cross validation, 10-fold cross validation and de novo drug validation, DDI-IS-SL can achieve the better prediction performance than other comparative methods. In addition, the average computation time of DDI-IS-SL is shorter than that of other comparative methods. Finally, case studies further demonstrate the performance of DDI-IS-SL in practical applications.

1. INTRODUCTION

Recently, based on machine learning models, many computational approaches have been developed to predict potential DDIs. Tatonetti *et al.* developed a signal discovery method to infer DDIs [18], main features of drugs used in this method are drug adverse event profiles. By combining drug chemical similarities, side effect similarities, proteinprotein interaction similarities and target sequence similarities, an INDI (INferring Drug Interactions) framework was developed to predict DDIs, which used two types of drug interactions (potential CYP (Cytochrome P450)-related, and non-CYP-related DDIs (NCRDs)) [19].

By the combination of crizotinib with ketoconazole or rifampin, a PBPK (physiologically based pharmacokinetic) model was developed for predicting DDIs [20]. Based on properties of the drug metabolism, the text-mining and reasoning approaches were also used to discover novel DDIs [21]. Vilar *et al.* computed the molecular fingerprint similarity and the molecular structure similarity of drugs to predict DDIs [22].

With 2D and 3D molecular structures, interaction profiles, target and side-effect similarities, Vilar *et al.* further developed a protocol applicable on a large scale data to infer novel DDIs [23]. Based on drug phenotypic, therapeutic, chemical, and genomic properties and machine learning model, Cheng *et al.* proposed a computational method to predict DDIs [24]. Based on the drug molecular similarity and phenotypic similarity, Li *et al.* developed a computational method to discover the combination efficacy of drugs with a Bayesian network model [25]. Based on a random forest model, Liu *et al.* proposed a computational method to predict DDIs by integrating chemical interactions, protein-protein interactions between targets of drugs and target enrichment of KEGG pathways [26]. This method adopted a feature selection technique to obtain the important features of drugs.

Luo *et al.* developed a computational method to predict DDIs by implementing the chemical-protein interactome, which provided as a web server (called DDICPI) [27]. Based on the framework probabilistic soft logic, Sridhar *et al.* took a PSL (Probabilistic Soft Logic) method to predict novel DDIs by integrating networks of multiple drug similarities and known DDs [13]. With 2D structural similarities of drugs, Takako *et al.* developed a logistic regression model to infer potential DDIs [28]. Its prediction performance is further improved by combining target related and enzyme-related scores.

Based on inner product-based similarity measures (IPSMs), Ferdousi *et al.* provided a computational method to predict DDIs. This method also used the drug similarity constructed with key biological elements including carriers, transporters, enzymes and targets of drugs. In addition, based on the assumption that synergistic effects with drugs are often similar and vice versa, NLLSS (Network-based Laplacian regularized Least Square Synergistic drug combination prediction) was proposed to predict hidden synergistic drug combinations, but it can not predict DDIs for new drugs [29].

Disadvantages

- ❖ The system is not implemented Regularized least squares classifier.
- ❖ In conjunction with security threats, an emerging concern on ML-based solutions is not suitable for drugs test, namely the non ml classifiers are very weak of information from the ML models to the adversaries.

2. PROPOSED SYSTEM

In this study, by integrating the chemical, biological and phenotype information of drugs, we develop a computational method (called DDI-IS-SL) to predict DDIs. This drug information includes drug chemical structures, drugtarget interactions, drug enzymes, drug transports, drug pathways, drug indications, drug side effects, drug off side effects and known DDIs.

First, based on these pieces of drug information, we construct a high-dimensional binary vector to calculate the feature similarity of drugs via the cosine similarity method. Furthermore, we also compute the Gaussian Interaction Profile (GIP) kernel similarity [38] of drugs based on known DDIs.

The final drug similarity is constructed by their feature similarity and GIP similarity.

Then a Regularized Least Squares (RLS) classifier [39] is adapted to predict DDIs. For new drugs which do not have any interactions with other drugs, we also calculate their relational initial scores via performing the node-based drug network diffusion method. Therefore, our method can predict potential DDIs not only for known drugs but also for new drugs. The prediction performance of our method and other competing methods are systematically assessed by the 5-fold cross validation, the 10-fold cross validation and the de novo validation. The AUC (area under the ROC curve) is used as the metric to evaluate the performance of computational methods. In terms of AUC, our method is superior to other competing methods.

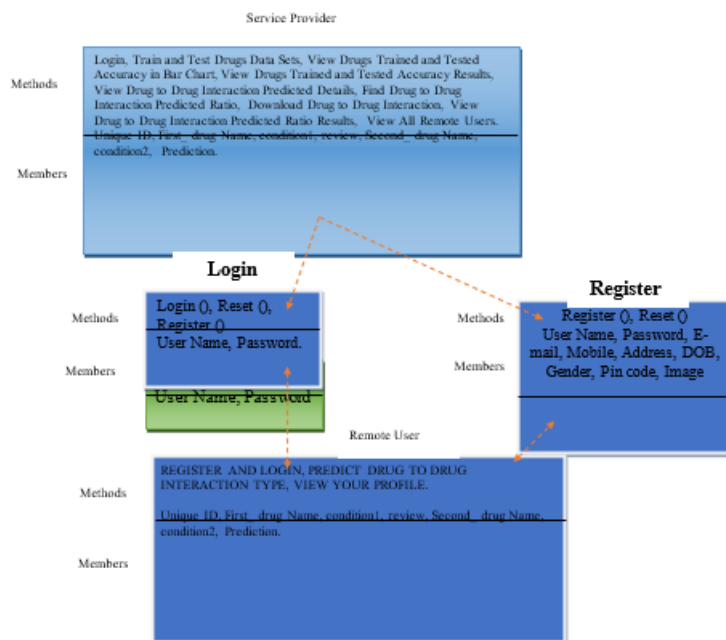
Specifically, in the 5-fold cross validation, the AUC value of our method is 0.9691, which is larger than the AUC of 0.9570 from the state-of-the-art L1E. Furthermore, in the 10-fold cross validation, the AUC value of our method reaches 0.9745, which is also larger than the best result of L1E whose AUC value is 0.9599. Our method also obtains the best prediction performance in the de novo drug validation, its AUC value is 0.9292, which is also larger than the best result of other methods (WAE (weighted average ensemble method) : 0.9073). In addition, the comparison of the average running time further improves that our method has the higher running efficiency than other competing methods. Finally, the verification results of case studies also prove the prediction ability of our method in practical applications and show that DDI-IS-SL is an effective computational method to predict new DDIs.

Advantages

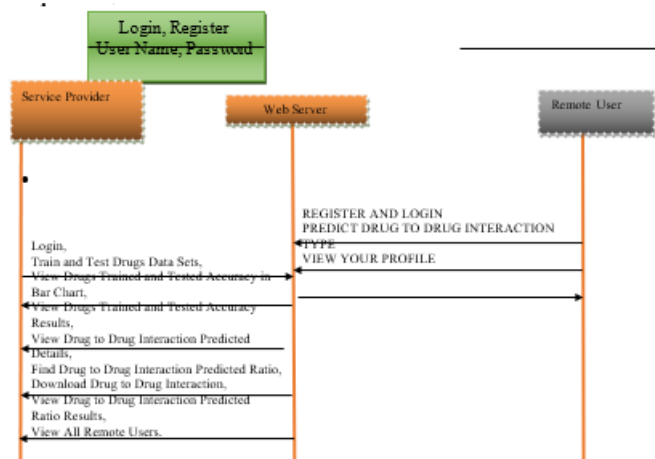
- The proposed system implemented many ml classifiers for testing and training on datasets.

- The proposed system developed a Regularized least squares classifier to find an accurate accuracy on the datasets.

3. CLASS DIAGRAM



4. SEQUENCE DIAGRAM



5. UNIT TESTING

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module’s control structure to ensure

complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

The following are the types of Integration Testing:

1. Top Down Integration

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner.

In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

2. Bottom-up Integration

This method begins the construction and testing with the modules at the lowest level in the program structure. Since the modules are integrated from the bottom up, processing required for modules subordinate to a given level is always available and the need for stubs is eliminated. The bottom up integration strategy may be implemented with the following steps:

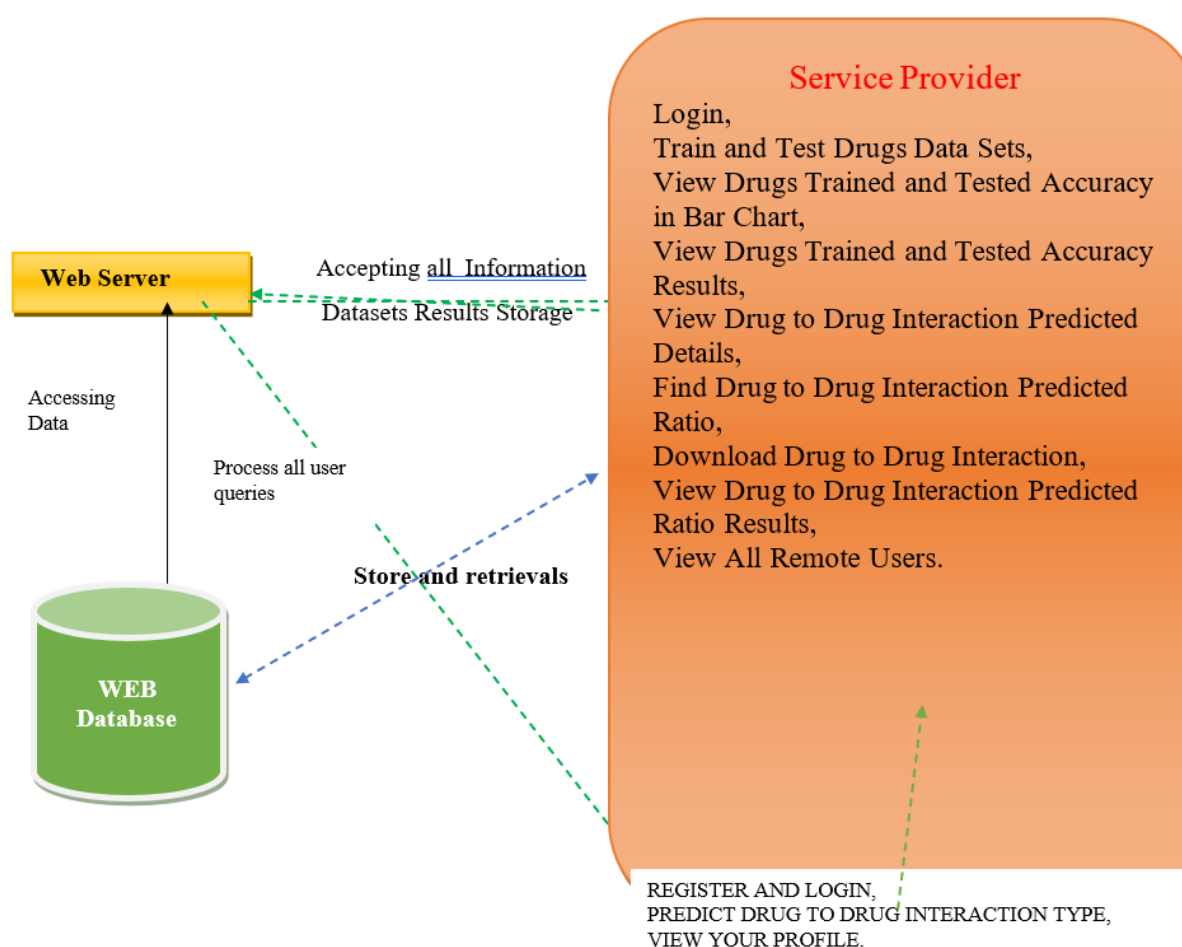
- The low-level modules are combined into clusters into clusters that perform a specific Software sub-function.
- A driver (i.e.) the control program for testing is written to coordinate test case input and output.
- The cluster is tested.
- Drivers are removed and clusters are combined moving upward in the program structure

The bottom up approaches tests each module individually and then each module is module is integrated with a main module and tested for functionality.

6. TEXT FIELD

The text field can contain only the number of characters lesser than or equal to its size. The text fields are alphanumeric in some tables and alphabetic in other tables. Incorrect entry always flashes and error message.

7. ARCHITECTURE DIAGRAM



8. CONCLUSION

Multi-drug therapies have widely been used to treat diseases, especially complex diseases such as cancer to improve the treatment effect and reduce the burden of patients. However, the adverse effects resulted from multi-drug therapies have also been observed, which may caused some serious complications and even the patient death. Therefore, identifying drug-

drug interactions is helpful in contributing to improved treatment of diseases and reducing the difficulty of drug developments. Especially, it is very necessary to develop new computational methods for identifying DDIs.

In this study, we propose a new computational method (DDI-IS SL) to infer DDIs. DDI-IS-SL integrates the drug chemical, drug biological and drug phenotypic data. The used chemical substructure information of drugs is Pub- Chem substructure which is the 2D binary fingerprints (0 and 1). The biological features of drugs contain drug target interactions, drug enzymes, drug transports and drug pathways. The phenotypic data of drugs include drug indications, drug side effects and drug-off side effects. For each drug, a high-dimensional binary feature vector is constructed with these data. Then we calculate the feature similarity of drugs with the cosine measure. We also compute the GIP similarity of drugs by known DDIs. The final similarity of drugs is calculated as the mean of drug feature similarity and drug GIP similarity. Then we use a semi-supervised learning model (RLS) to compute the probability scores of drug pairs. In the 5-fold cross validation and 10-fold cross validation, DDI-IS-SL achieves the better prediction performance than other competing methods. Furthermore, for new drugs, we also calculate the relational initial interaction scores by using the node-based drug network diffusion method. Our method also achieves the better prediction performance in de novo validation than competing methods.

Although the DDI-IS SL is an effective approach to predict the potential DDIs, there are still some areas for the improvement

9. REFERENCES

- [1] D. Quinn and R. Day, "Drug interactions of clinical importance," *Drug safety*, vol. 12, no. 6, pp. 393–452, 1995.
- [2] T. Prueksaritanont, X. Chu, C. Gibson, D. Cui, K. L. Yee, J. Ballard, T. Cabalu, and J. Hochman, "Drug–drug interaction studies: Regulatory guidance and an industry perspective," *The AAPS journal*, vol. 15, no. 3, pp. 629–645, 2013.
- [3] H. Kusuvara, "How far should we go? perspective of drug-drug interaction studies in drug development," *Drug metabolism and pharmacokinetics*, vol. 29, no. 3, pp. 227–228, 2014.
- [4] N. R. Crowther, A. M. Holbrook, R. Kenwright, and M. Kenwright, "Drug interactions among commonly used medications. Chart simplifies data from critical literature review." *Canadian Family Physician*, vol. 43, p. 1972, 1997.

- [5] R. Nahta, M.-C. Hung, and F. J. Esteva, “The her-2-targeting antibodies trastuzumab and pertuzumab synergistically inhibit the survival of breast cancer cells,” *Cancer research*, vol. 64, no. 7, pp. 2343–2346, 2004.
- [6] T.-C. Chou, “Drug combination studies and their synergy quantification using the chou-talalay method,” *Cancer research*, vol. 70, no. 2, pp. 440–446, 2010.
- [7] K. Venkatakrisnan, L. L. von Moltke, R. Obach, and D. J. Greenblatt, “Drug metabolism and drug interactions: application and clinical value of in vitro models,” *Current drug metabolism*, vol. 4, no. 5, pp. 423–459, 2003.



CROP YIELD PREDICTION USING MACHINE LEARNING ALGORITHM

Karumuri Vasudha Sri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Agriculture is one of the major and the least paid occupation in India. Machine learning can bring a boom in the agriculture field by changing the income scenario through growing the optimum crop. This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.

1.INTRODUCTION

1.1 MOTIVATION

The history of agriculture in India[1] dates back to the Indus Valley Civilization Era. India ranks second in this sector. Agriculture and allied sectors like forestry and fisheries account for 15.4 percent of the GDP (gross domestic product) with about 31 percent of the workforce. India ranks first globally with the highest net cropped area followed by US and China. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India. Due to the revolution in industrialization, the economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth.

1.2 PROBLEM DEFINITION

The problem that the Indian Agriculture sector is facing is the integration of technology to bring the desired outputs. With the advent of new technologies and overuse of non-renewable energy resources patterns of rainfall and temperature are disturbed. The inconsistent trends developed from the side effects of global warming make it cumbersome for the farmers to clearly predict the temperature and rainfall patterns thus affecting their crop yield productivity. In order to perform accurate prediction and handle inconsistent trends in temperature and rainfall various machine learning algorithms like RNN, LSTM, etc can be applied to get a pattern. It will complement the agricultural growth in India and all together augment the ease of living for farmers. In past, many researchers have applied machine learning techniques to enhance agricultural growth of the country

1.3 OBJECTIVE OF PROJECT

This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.



2.LITERATURE SURVEY

2.1 PREDICTING YIELD OF THE CROP USING MACHINE LEARNING ALGORITHM

AUTHORS: P.Priya, U.Muthaiah & M.Balamurugan

The agriculture plays a dominant role in the growth of the country's economy. Climate and other environmental changes has become a major threat in the agriculture field. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Crop Yield Prediction involves predicting yield of the crop from available historical available data like weather parameter, soil parameter and historic crop yield. This paper focus on predicting the yield of the crop based on the existing data by using Random Forest algorithm. Real data of Tami Inadu were used for building the models and the models were tested with samples. The prediction will helps to the farmer to predict the yield of the crop before cultivating onto the agriculture field. To predict the crop yield in future accurately Random Forest, a most powerful and popular supervised machine learning algorithm is used.

2.2 Applications of machine learning techniques in agricultural crop production: a review

AUTHORS: Mishra .s, Mishra .D and Santra .G. H

This paper has been prepared as an effort to reassess the research studies on the relevance of machine learning techniques in the domain of agricultural crop production. Methods/Statistical Analysis: This method is a new approach for production of agricultural crop management. Accurate and timely forecasts of crop production are necessary for important policy decisions like import-export, pricing marketing distribution etc. which are issued by the directorate of economics and statistics. However one has understand that these prior estimates are not the objective estimates as these estimate requires lots of descriptive assessment based on many different qualitative factors. Hence there is a requirement to develop statistically sound objective prediction of crop production. That development in computing and information storage has provided large amount of data. Findings: The problem has been to intricate knowledge from this raw data, this has lead to the development of new approach and techniques such as machine learning that can be used to unite the knowledge of the data with crop yield evaluation. This research has been intended to evaluate these innovative techniques such that significant relationship can be found by their applications to the various variables present in the data base. Application/Improvement: The few techniques like artificial neural networks, Information Fuzzy Network, Decision Tree, Regression Analysis, Bayesian belief network. Time series analysis, Markov chain model, k-means clustering, k nearest neighbor, and support vector machine are applied in the domain of agriculture were presented.

2.3 A Model for Prediction of Crop Yield.

AUTHORS: Manjula.E

Data Mining is emerging research field in crop yield analysis. Yield prediction is a very important issue in agricultural. Any farmer is interested in knowing how much yield he is



about to expect. In the past, yield prediction was performed by considering farmer's experience on particular field and crop. The yield prediction is a major issue that remains to be solved based on available data. Data mining techniques are the better choice for this purpose. Different Data Mining techniques are used and evaluated in agriculture for estimating the future year's crop production. This research proposes and implements a system to predict crop yield from previous data. This is achieved by applying association rule mining on agriculture data. This research focuses on creation of a prediction model which may be used to future prediction of crop yield. This paper presents a brief analysis of crop yield prediction using data mining technique based on association rules for the selected region i.e. district of Tamil Nadu in India. The experimental results shows that the proposed work efficiently predict the crop yield production.

2.4 Agricultural crop yield prediction using artificial neural network approach

AUTHORS: Dahikar, S. S, Rode and S. V.

By considering various situations of climatologically phenomena affecting local weather conditions in various parts of the world. These weather conditions have a direct effect on crop yield. Various researches have been done exploring the connections between large-scale climatologically phenomena and crop yield. Artificial neural networks have been demonstrated to be powerful tools for modeling and prediction, to increase their effectiveness. Crop prediction methodology is used to predict the suitable crop by sensing various parameter of soil and also parameter related to atmosphere. Parameters like type of soil, PH, nitrogen, phosphate, potassium, organic carbon, calcium, magnesium, sulphur, manganese, copper, iron, depth, temperature, rainfall, humidity. For that purpose we are used artificial neural network (ANN).

2.5 Predictive ability of machine learning methods for massive crop yield prediction.

AUTHORS: Gonzalez Sanchez. A, Frausto Sols. J and Ojeda Bustamante. W

An important issue for agricultural planning purposes is the accurate yield estimation for the numerous crops involved in the planning. Machine learning (ML) is an essential approach for achieving practical and effective solutions for this problem. Many comparisons of ML methods for yield prediction have been made, seeking for the most accurate technique. Generally, the number of evaluated crops and techniques is too low and does not provide enough information for agricultural planning purposes. This paper compares the predictive accuracy of ML and linear regression techniques for crop yield prediction in ten crop datasets. Multiple linear regression, M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor methods were ranked. Four accuracy metrics were used to validate the models: the root mean square error (RMS), root relative square error (RRSE), normalized mean absolute error (MAE), and correlation factor (R). Real data of an irrigation zone of Mexico were used for building the models. Models were tested with samples of two consecutive years. The results show that M5- Prime and k-nearest neighbor techniques obtain the lowest average RMSE errors (5.14 and 4.91), the lowest RRSE errors (79.46% and 79.78%), the lowest average MAE errors (18.12% and 19.42%), and the highest average correlation factors (0.41 and 0.42). Since M5-Prime

achieves the largest number of crop yield models with the lowest errors, it is a very suitable tool for massive crop yield prediction in agricultural planning.

3.SYSTEM ANALYSIS

3.1 EXISTING SYSTEM:

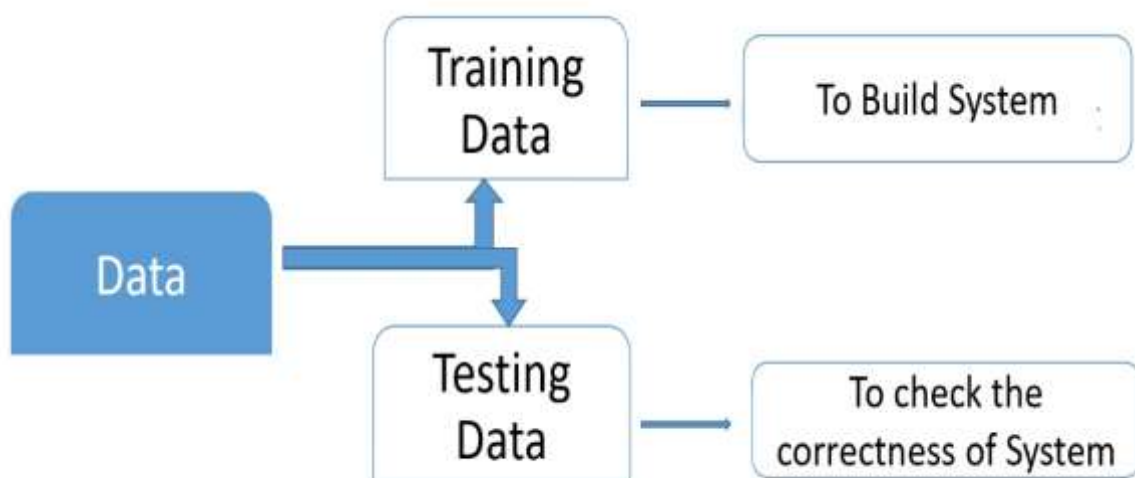
Due to the revolution in industrialization, the economic contribution of agriculture to India's GDP is steadily declining with the country's broad-based economic growth. The problem that the Indian Agriculture sector is facing is the integration of technology to bring the desired outputs. With the advent of new technologies and overuse of non-renewable energy resources patterns of rainfall and temperature are disturbed. The inconsistent trends developed from the side effects of global warming make it cumbersome for the farmers to clearly predict the temperature and rainfall patterns thus affecting their crop yield productivity. In order to perform accurate prediction and handle inconsistent trends in temperature and rainfall various machine learning algorithms like RNN, LSTM, etc can be applied to get a pattern. It will complement the agricultural growth in India and all together augment the ease of living for farmers. In past, many researchers have applied machine learning techniques to enhance agricultural growth of the country.

3.2 PROPOSED SYSTEM:

- ❖ This paper focuses on the practical application of machine learning algorithms and its quantification. The work presented here also takes into account the inconsistent data from rainfall and temperature datasets to get a consistent trend. Crop yield prediction is determined by considering all the features in contrast with the usual trend of determining the prediction considering one feature at a time.

4.SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE:



4.2 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

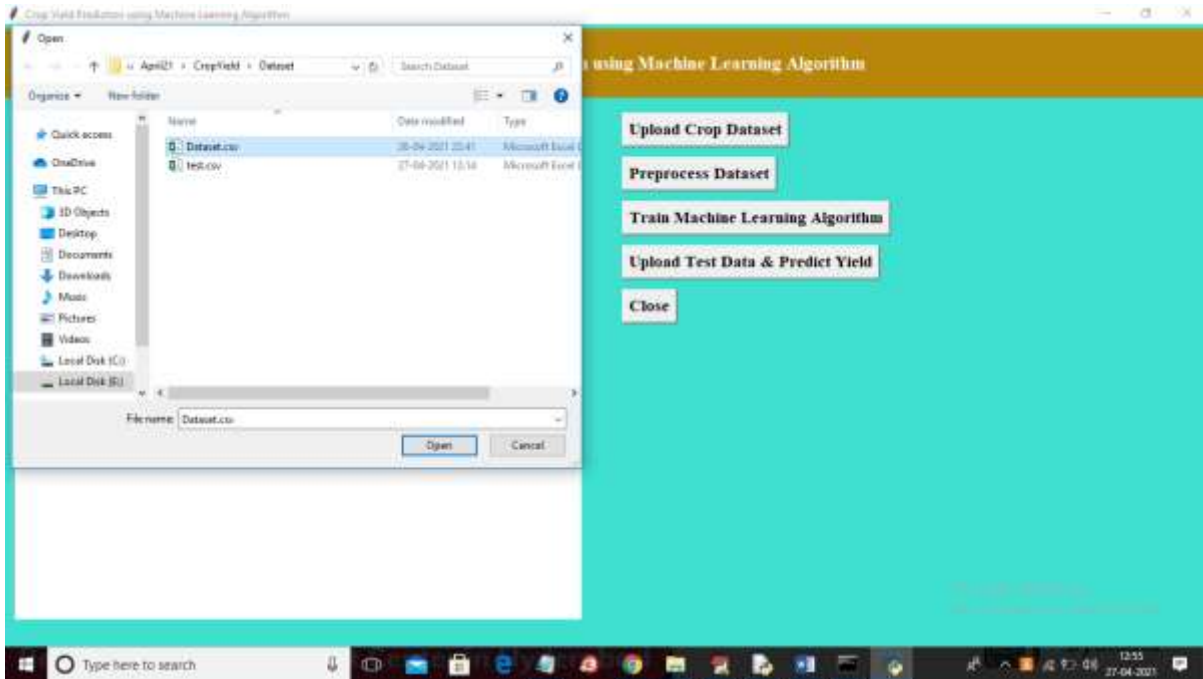
1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

5. SCREENSHOTS

To run project double click on 'run.bat' file to get below screen



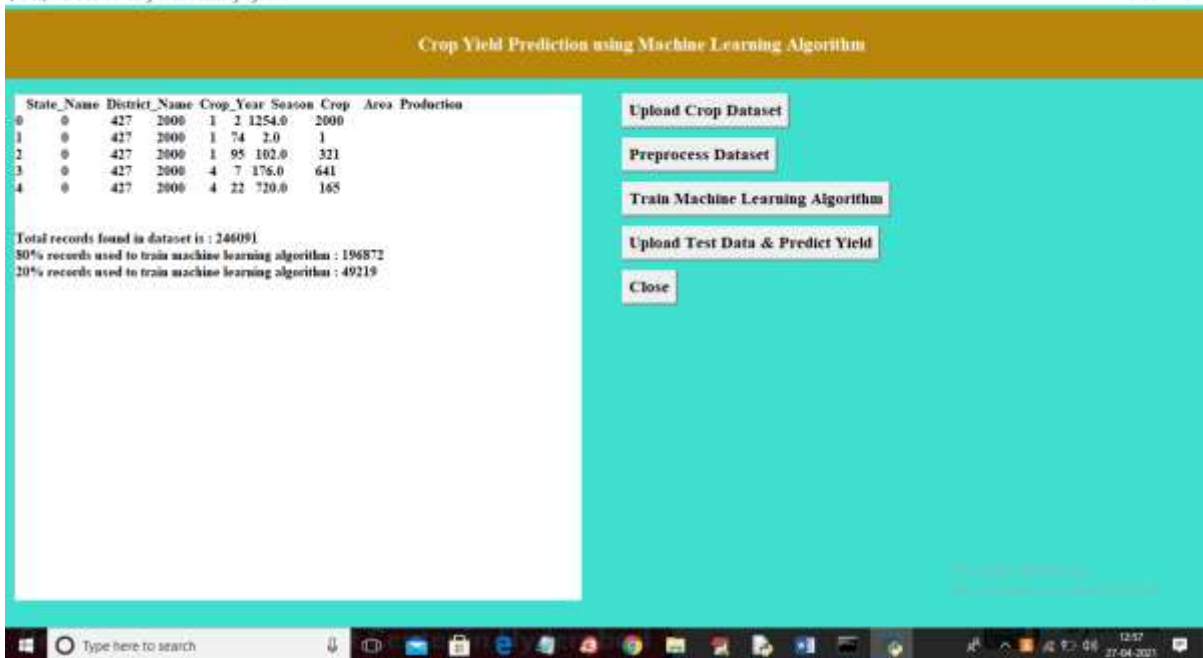
In above screen click on 'Upload Crop Dataset' button to upload dataset



In above screen selecting and uploading 'Dataset.csv' file and then click on 'Open' button to load dataset and to get below screen



In above screen dataset loaded and we can see dataset contains some non-numeric values and ML will not take non-numeric values so we need to preprocess dataset to convert non-numeric values to numeric values by assigning ID to each non-numeric value. So click on 'Preprocess Dataset' button to process dataset



In above screen all non-numeric values converted to numeric format and in below lines we can see dataset contains total 246091 records and application using (80%) 196872 records to train ML and using (20%) 49219 records to test ML prediction error rate (RMSE (root mean square error)). Now click on 'Train Machine Learning Algorithm' button to train Decision Tree Machine learning algorithm on above dataset and then calculate prediction error rate



6. CONCLUSION

The paper presented the various machine learning algorithms for predicting the yield of the crop on the basis of temperature, rainfall, season and area. Experiments were conducted on Indian government dataset and it has been established that Random Forest Regressor gives



the highest yield prediction accuracy. Sequential model that is Simple Recurrent Neural Network performs better on rainfall prediction while LSTM is good for temperature prediction. By combining rainfall, temperature along with other parameters like season and area, yield prediction for a certain district can be made. Results reveals that Random Forest is the best classifier when all parameters are combined. This will not only help farmers in choosing the right crop to grow in the next season but also bridge the gap between technology and the agriculture sector.

7. REFERENCES

1. Agriculture Role on Indian Economy Madhusudhan L - <https://www.omicsonline.org/open-access/agriculture-role-on-indianeconomy-2151-6219-1000176.php?aid=62176>
2. Priya, P., Muthaiah, U., Balamurugan, M. International Journal of Engineering Sciences Research Technology Predicting Yield of the Crop Using Machine Learning Algorithm.
3. Mishra, S., Mishra, D., Santra, G. H. (2016). Applications of machine learning techniques in agricultural crop production: a review paper. Indian J. Sci. Technol, 9(38), 1-14.
4. Manjula, E., Djodiltachoumy, S. (2017). A Model for Prediction of Crop Yield. International Journal of Computational Intelligence and Informatics, 6(4), 2349-6363.
5. Dahikar, S. S., Rode, S. V. (2014). Agricultural crop yield prediction using artificial neural network approach. International journal of innovative research in electrical, electronics, instrumentation and control engineering, 2(1), 683-686.
6. Gonzalez Sanchez, A., Frausto Sols, J., Ojeda Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction.
7. Mandic, D. P., Chambers, J. (2001). Recurrent neural networks for prediction: learning algorithms, architectures and stability. John Wiley Sons, Inc..
8. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
9. Sak, H., Senior, A., Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Fifteenth annual conference of the international speech communication association.
10. Liaw, A., Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

T-CREO A TWITTER CREDICATABILITY ANALYSIS FRMAEWORK

Katakamsetti Saikiran (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract: Social media and other platforms on Internet are commonly used to communicate and generate information. In many cases, this information is not validated, which makes it difficult to use and analyze. Although there exist studies focused on information validation, most of them are limited to specific scenarios. Thus, a more general and flexible architecture is needed, that can be adapted to user/developer requirements and be independent of the social media platform. We propose a framework to automatically and in real-time perform credibility analysis of posts on social media, based on three levels of credibility: Text, User, and Social. The general architecture of our framework is composed of a front-end, a light client proposed as a web plug-in for any browser; a back-end that implements the logic of the credibility model; and a third-party services module. We develop a first version of the proposed system, called TCREo (Twitter Credibility analysis framework) and evaluate its performance and scalability. In summary, the main contributions of this work are: the general framework design; a credibility model adaptable to various social networks, integrated into the framework; and T-CREo as a proof of concept that demonstrates the framework applicability and allows evaluating its performance for unstructured information sources; results show that T-CREo qualifies as a highly scalable real-time service. The future work includes the improvement of T-CREo implementation, to provide a robust architecture for the development of third party applications, as well as the extension of the credibility model for considering bots detection, semantic analysis and multimedia analysis.

1. INTRODUCTION

NOWADAYS, social media generates an immense amount of information, since they are what people mostly use to share and read about a wide variety of topics. In this way, information is shared in free environments that can be used in several contexts, ranging from everyday life, global and local news, to the development of new technologies [1]–[3]. Social media and other platforms on the Internet, which allow users to communicate, share, and generate information

without formal references to sources, became popular in the early 1990s, producing such a vast amount of information that fits into the Big Data category. However, in many cases, this information is not documented or validated, which makes it tough to use and analyze. Hence, the concept of credibility, as the level of belief that is perceived about (how credible it is) a person, object, or process [4], has become essential in various disciplines and from different perspectives, such as information engineering, business administration, communications management, journalism, information retrieval, human-computer interaction [5], [6].

However, existing works are limited to be applicable to analysis of credibility on specific scenarios (e.g., for a specific social platform, for a particular application). These works differ in the characteristics taken into account to calculate credibility (e.g., attributes of the posts or of users who posted them, the text of the posts, user social impact) and in the extraction techniques used to gather the information to feed the credibility models (i.e., web scraping¹ or API). Thus, a more general and flexible architecture is needed, that can be adapted to user/developer's requirements and be independent of the social media platform.

To overcome these limitations, we propose a framework to automatically and in real-time perform credibility analysis of posts on social media. The framework instantiates a credibility model proposed in our previous work [4], which consists of the credibility analysis of publications on information sources, adaptable to various social networks. The credibility model is based on three aspects: Text Credibility (based on text analysis), User Credibility (based on attributes about the user's account, such as creation date, verified account), and Social Credibility (based on attributes that reflect social impact, such as followers and following). In this work, we describe the general architecture of the framework and demonstrate its applicability for unstructured information sources, taking as reference Twitter, which is one of the most used among social media networks.

The characteristics of our proposed framework architecture, that make it different from the existing works, are mainly the following:

- _ It provides two approaches for accessing the information needed for the credibility model: web scraping and social media API; users/developers can configure the system to base the information gathering only with web scraping or combining it with the use of the available API;
- _ It performs credibility analysis automatically and in real-time;
- _ It consists of a front-end, which is proposed as a web plug-in to be incorporated on any browser, and a decoupled back-end which executes the credibility analysis;

It is light-decoupled from external components; as a consequence, it is extensible and flexible; thus, it can be adapted to any social media platform and the credibility model can be extended by replacing or integrating other measures to calculate different credibility levels.

We develop a first version of the proposed system, called TCREO (Twitter credibility analysis framework) as a proof of concept. As a Google Chrome Extension, TCREO perform the credibility analysis of tweets, in real-time. According to the study presented in [7], Twitter statistics indicate that around 500 millions of tweets are published every day. Thus, credibility analysis in such as platform has become a trending topic in the last decades [8]–[11]. There exist many studies proposing Twitter credibility models [4], [8], [11], [12] and more complete studies, which also propose frameworks to perform the credibility analysis automatically and in real-time [13]–[18]. We qualitatively compare our proposal with the state-of-the-art and we show the performance evaluation of T-CREO in various scenarios, with different variables, such as number of requests and number of concurrent clients/connections. Results show that the performance of T-CREO qualifies it as a real-time and highly scalable service.

In summary, the main contributions of this work are: (i) the design of a framework to perform credibility analysis on social networks, automatically and in real-time; (ii) a credibility model adaptable to various social networks, integrated into the framework; and (iii) T-CREO as a proof of concept that demonstrates the framework applicability and allows a comparative evaluation with existing systems and an evaluation of its performance.

2. EXISTING SYSTEM

- ❖ Text Credibility (Post Credibility): measures the level of relevance and accuracy of the text, independent of the referenced topic [8] or with respect to a certain topic [11]. It is calculated through text analysis techniques, such as Natural Language Processing (NLP).
- ❖ User Credibility: calculates the user account credibility based on attributes that describe it. It can be calculated based on, for example, the account creation date, if the account is verified, user's age.
- ❖ Social Credibility: calculates the credibility of a publication, related or not to a topic, based on the available metadata that describe the social impact of the user account and

the post itself, with respect to other users. It is calculated based on data such as number of followers, number of following, retweets.

- ❖ Topic-level Credibility: measures the level of acceptance of the topic or event referenced in the text. It consists of identifying if the text refers to a specific topic or not, usually through NLP and sentiment analysis techniques.

Disadvantages

- 1). The system doesn't have technique to find a concept in which credibility measures attribute a global credibility level of a publication in a social information source.
- 2). There is no technique to detect Source Of Fake News (SOFNs), by analysing credibility of tweets based on graph Machine Learning

3. PROPOSED SYSTEM

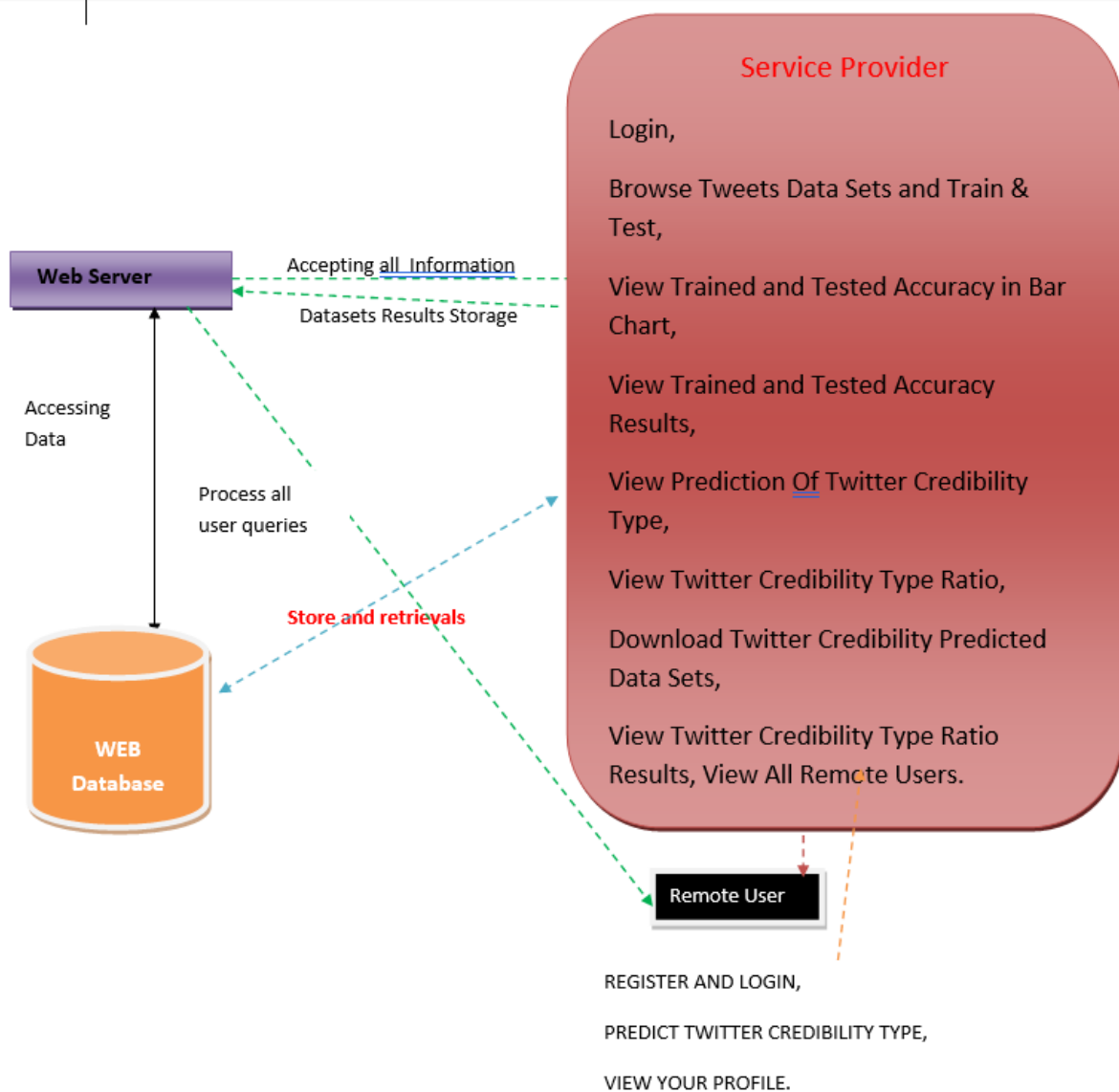
- ❖ The system proposes a framework to automatically and in real-time perform credibility analysis of posts on social media. The framework instantiates a credibility model proposed in our previous work [4], which consists of the credibility analysis of publications on information sources, adaptable to various social networks.
- ❖ The credibility model is based on three aspects: Text Credibility (based on text analysis), User Credibility (based on attributes about the user's account, such as creation date, verified account), and Social Credibility (based on attributes that reflect social impact, such as followers and following).
- ❖ ¹In this work, we describe the general architecture of the framework and demonstrate its applicability for unstructured information sources, taking as reference Twitter, which is one of the most used among social media networks.

Advantages

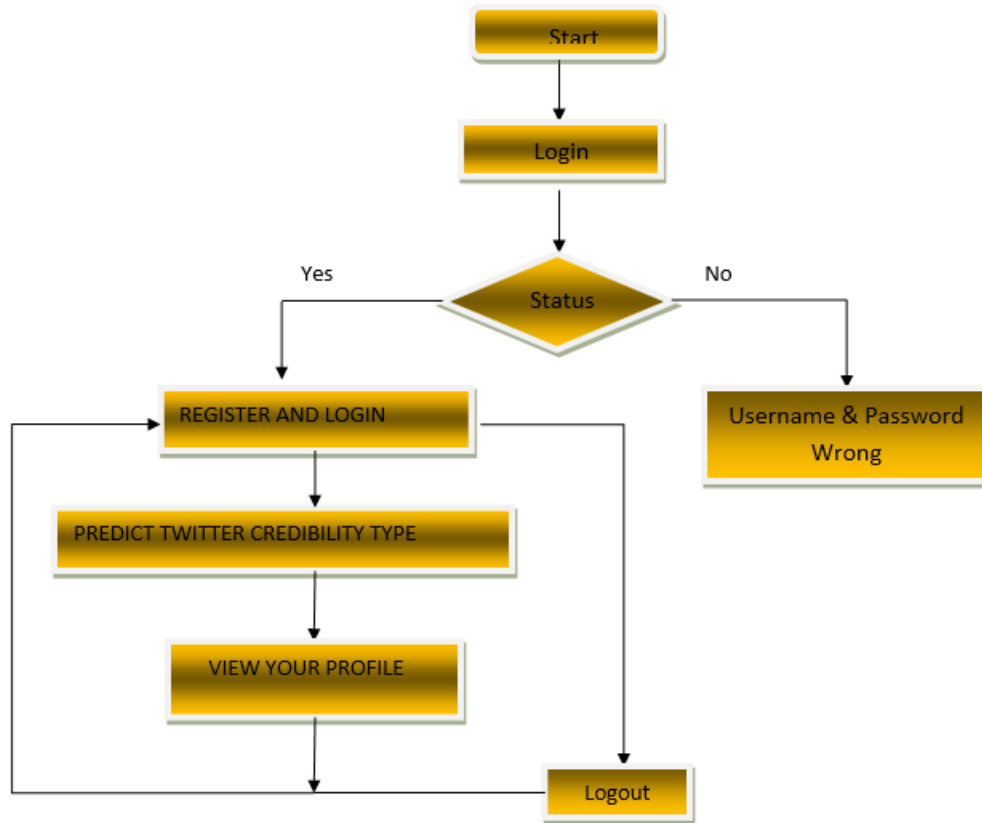
- It provides two approaches for accessing the information needed for the credibility model: web scraping and social media API; users/developers can configure the system to base the information gathering only with web scraping or combining it with the use of the available API;

- It performs credibility analysis automatically and in real-time;
- It consists of a front-end, which is proposed as a web plug-in to be incorporated on any browser, and a decoupled back-end which executes the credibility analysis;
- It is light-decoupled from external components; as a consequence, it is extensible and flexible; thus, it can be adapted to any social media platform and the credibility model can be extended by replacing or integrating other measures to calculate different credibility levels.

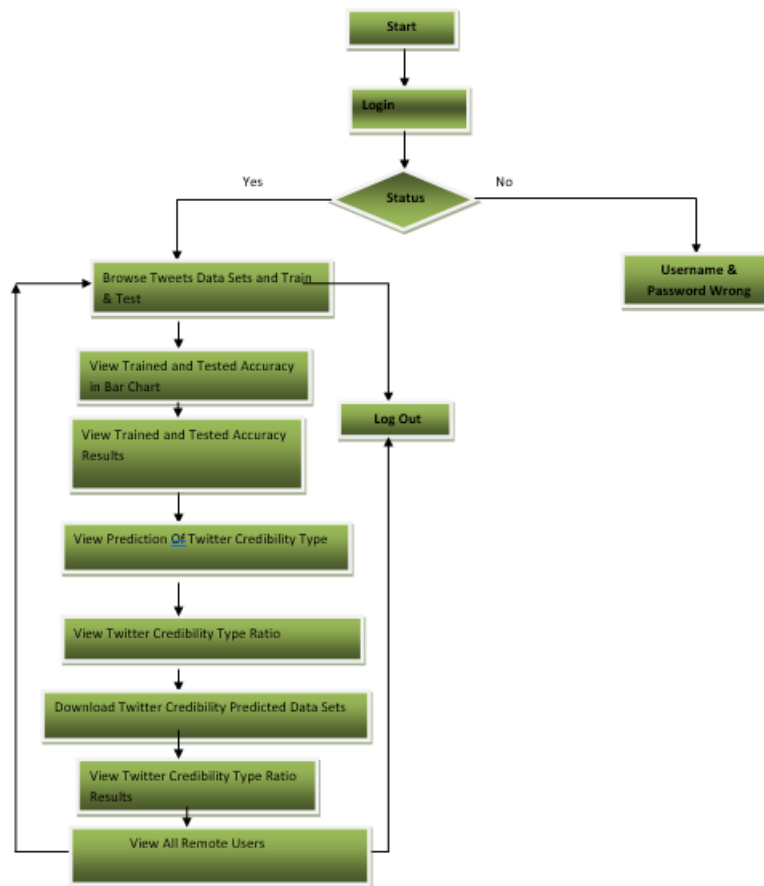
4. SYSTEM ARCHITECTURE



5. FLOWCHART



Remote user



Service provider

6. CONCLUSION

In this work, we propose a general architecture of a framework for credibility analysis in social media based on a general credibility model. The framework is capable of calculating credibility on any social media in real-time, combining web-scraping and social media APIs to gather the parameters needed to instantiate the credibility model. A proof of concept, for a specific use case of Twitter and to show the feasibility of the proposed architecture, named TCREO (Twitter credibility analysis framework), is developed and tested to evaluate its performance. Results show that our proposed framework can be implemented as a real time service and the scalability is ensured by increasing the level of concurrency. This experience allows outlining some suggestions to improve overall performance for high capacity servers. The modularity and simplicity of T-CREO, and the use of the credibility model, enable the creation of a real-time service; however, the connection time (latency) can be a determining factor, that might be considered in the deployment of the system.

Our future research is focused on the improvement of TCREO, starting with the suggestions from such as the implementation of several instances or multi-threaded versions of the back-end to improve the performance, keep an external database of posts to overcome API limitations, and incorporate credibility analysis in other social platforms, to provide a robust architecture to the community for the development of third-party applications. We also plan to extend the credibility model by considering bots detection, semantic analysis of the text, and multimedia data analysis.

7. REFERENCES

- [1] D. Westerman, P. R. Spence, B. Van Der Heide, Social media as information source: Recency of updates and credibility of information, *J. of Computer-Mediated Comm.* 19 (2) (2014) 171–183. doi:10.1111/jcc4.12041.
- [2] Y. Kammerer, E. Kalbfell, P. Gerjets, Is this information source commercially biased? how contradictions between web pages stimulate the consideration of source information, *Discourse Processes* 53 (5-6) (2016) 430–456. doi:10.1080/0163853X.2016.1169968.
- [3] J. Slomian, O. Bruyère, J.-Y. Reginster, P. Emonts, The internet as a source of information used by women after childbirth to meet their need for information: A web-based survey, *Midwifery* 48 (2017) 46–52. doi:10.1016/j.midw.2017.03.005.
- [4] I. Dongo, Y. Cardinale, A. Aguilera, Credibility analysis for available information sources on the web: A review and a contribution, in: 2019 4th Internat. Conf. on System Reliability and Safety (ICSRS), IEEE, 2019, pp. 116–125. doi:10.1109/ICSRS48664.2019.8987623.
- [5] S. Y. Rieh, D. R. Danielson, Credibility: A multidisciplinary framework, *Annual Rev. Info. Sci & Technol.* 41 (1) (2007) 307–364. doi:10.1002/aris.144.v41:1.
- [6] T. J. Johnson, B. K. Kaye, Reasons to believe: Influence of credibility on motivations for using social networks, *Computers in human behavior* 50 (2015) 544–555. doi:10.1016/j.chb.2015.04.002.
- [7] Omnicore, Twitter by the Numbers: Stats, Demographics & Fun Facts, <https://www.omnicoreagency.com/twitter-statistics> (2020).
- [8] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proc. of Internat. Conf. on WWW, 2011, pp. 675–684. doi:10.1145/1963405.1963500.

[9] B. Kang, T. Höllerer, J. O'Donovan, Believe it or not? analyzing information credibility in microblogs, in: Proc. of Internat. Conf. on Advances in Social Networks Analysis and Mining, 2015, pp. 611–616. doi:10.1145/2808797.2809379.

[10] S. M. Shariff, X. Zhang, M. Sanderson, On the credibility perception of news on twitter: Readers, topics and features, Computers in Human Behavior 75 (2017) 785–796. doi:10.1016/j.chb.2017.06.026.

[11] M. Alrubaian, M. Al-Qurishi, A. Alamri, M. Al-Rakhami, M. M. Hassan, G. Fortino, Credibility in online social networks: A survey, IEEE Access 7 (2019) 2828–2855. doi:10.1109/ACCESS.2018.2886314.

CLOUD RAID DETECTING DISTRIBUTED CONCURRENCY BUGS VIA LOG MINING AND ENHANCEMENT

Katari Naga Raju (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Cloud systems suffer from distributed concurrency bugs, which often lead to data loss and service outage. This paper presents CLOUDRAID, a new automatic tool for finding distributed concurrency bugs efficiently and effectively. Distributed concurrency bugs are notoriously difficult to find as they are triggered by untimely interaction among nodes, i.e., unexpected message orderings. To detect concurrency bugs in cloud systems efficiently and effectively, CLOUDRAID analyzes and tests automatically only the message orderings that are likely to expose errors. Specifically, CLOUDRAID mines the logs from previous executions to uncover the message orderings that are feasible but inadequately tested. In addition, we also propose a log enhancing technique to introduce new logs automatically in the system being tested. These extra logs added improve further the effectiveness of CLOUDRAID without introducing any noticeable performance overhead. Our log-based approach makes it well-suited for live systems. We have applied CLOUDRAID to analyze six representative distributed systems: Hadoop2/Yarn, HBase, HDFS, Cassandra, Zookeeper, and Flink. CLOUDRAID has succeeded in testing 60 different versions of these six systems (10 versions per system) in 35 hours, uncovering 31 concurrency bugs, including nine new bugs that have never been reported before. For these nine new bugs detected, which have all been confirmed by their original developers, three are critical and have already been fixed.

1. INTRODUCTION

Distributed systems, such as scale-out computing frameworks [1], [2], distributed key-value stores [3], [4], scalable file systems [3], [4] and cluster management services [2], are the fundamental building blocks of modern cloud applications. As cloud applications provide 24/7 online services to users, high reliability of their underlying distributed systems becomes crucial. However, distributed systems are notoriously difficult to get right. There are widely

existing software bugs in real-world distributed systems, which often cause data loss and cloud outage, costing service providers millions of dollars per outage [5], [6].

Among all types of bugs in distributed systems, distributed concurrency bugs are among the most troublesome [7], [8]. These bugs are triggered by complex inter leavings of messages, i.e., unexpected orderings of communication events. It is difficult for programmers to correctly reason about and handle concurrent executions on multiple machines. This fact has motivated a large body of research on distributed system model checkers [9], [10], [11], [12], which detect hard-to-find bugs by exercising all possible message orderings systematically. Theoretically, these model checkers can guarantee reliability when running the same workload verified earlier. However, distributed system model checkers face the state-space explosion problem [9]. Despite recent advances [9], it is still difficult to scale them to many large real-world applications. For example, in our experiments for running the WordCount workload on Hadoop2/Yarn, 5,495 messages are involved. Even in such a simple case, it becomes impractical to test exhaustively all possible message orderings in a timely manner.

This paper proposes a novel strategy for detecting distributed concurrency bugs. Instead of trying all possible message orderings exhaustively, we test selectively only those message orderings that are likely to expose bugs. Which message orderings are likely to trigger errors then? We address this key question based on two observations:

2. EXISTING SYSTEM

Liu et al. have recently extended race detection techniques for multi-threaded programs] to detect race conditions in distributed systems. Their approach instruments memory accesses and communication events in a system to collect runtime traces at run time. An offline analysis is performed to analyze the happen-before relation among the memory accesses, by using a happenbefore model customized to distributed systems. Concurrent memory accesses that may trigger exceptions are regarded as harmful data races. A trigger is employed to further verify the detected race conditions. In , its approach mines logs to recover runtime traces without instrumentation, by restricting itself to message orderings involving only two messages. In this paper, we have improved the effectiveness of this earlier approach with two significant extensions. First, we introduce a new log enhancement technique, which allows us to detect bugs that would otherwise be missed. Second, we are now capable of detecting bugs that manifest themselves in message orderings involving an arbitrary number of messages.

With these two extensions, we have provided experimental evidence that our framework can find more bugs in new applications.

Fault injection techniques are commonly used to test the resilience of distributed systems. However, they focus on how to inject faults at different system states to expose bugs in the fault handlers. CLOUDRAID can be applied together to detect fault-related concurrency bugs more effectively.

Xu et al. mine console logs from a system and apply machine learning techniques to detect anomaly executions. Mined information such as logged values and logging frequencies is visualized to help users diagnose anomaly behaviors. DISTALYZER compares logs from abnormal and normal executions to infer the strongest association between system components and performance. Iprof extracts request IDs and timing information from logs to profile request latency. Stitch organizes log instances into tasks and sub-tasks, by analyzing relations among the logged ID variables to profile different components in the entire distributed software stack. In contrast, CLOUDRAID mines logs to uncover insufficiently exercised message orderings to detect concurrency bugs effectively.

CRASHTUNER applies a similar log analysis to infer some system meta-info, e.g., the running nodes and tasks/resources associated to each node. This tool makes use of the meta-info to detect crash-recovery bugs, which are triggered by crashing a node where its associated meta-info is being accessed. In contrast, CLOUDRAID applies log analysis to uncover the orderings between communication events for the purposes of detecting distributed concurrency bugs.

Disadvantages

An existing methodology doesn't implement a novel strategy for detecting distributed concurrency bugs.

The system is not aiming at CLOUDRAID leverages the run-time logs of live systems and avoids unnecessary repetitive tests.

3. PROPOSED SYSTEM

We propose a new approach, CLOUDRAID, for detecting concurrency bugs in distributed systems efficiently and effectively. CLOUDRAID leverages the run-time logs of live systems

and avoids unnecessary repetitive tests, thereby drastically improving the efficiency and effectiveness of our approach.

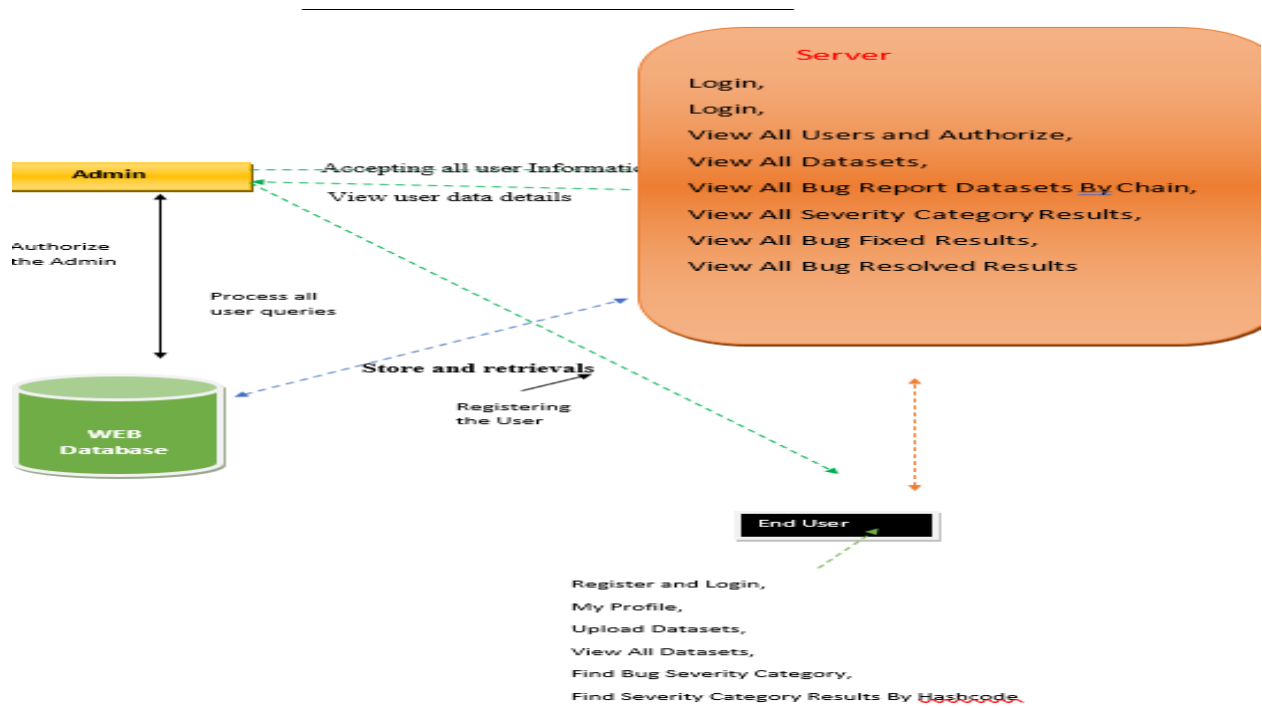
We describe a new log enhancing technique for improving log quality automatically. This enables us to log key communication events in a system automatically without introducing any noticeable performance penalty. The enhanced logs can further improve the overall effectiveness of our approach.

We have evaluated extensively CLOUDRAID using six representative distributed systems: Hadoop2/Yarn, HBase, HDFS, Cassandra, Zookeeper, and Flink. CLOUDRAID can test 60 different versions of these six systems (with six workloads in total) in 35 hours, and detect successfully 31 concurrency bugs. Among them, there are nine new bugs, including three critical ones, which have been fixed by their original developers.

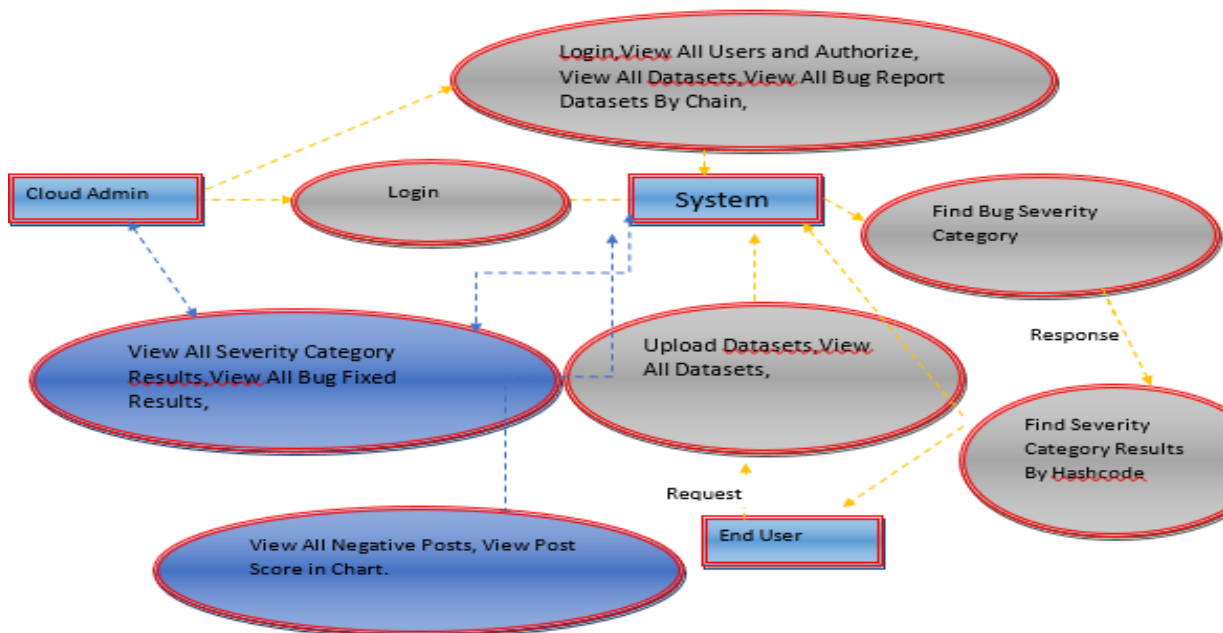
Advantages

The proposed approach focuses on detecting the bugs caused by order violation, i.e., the bugs which manifest themselves whenever a message arrives at a wrong order with respect to another event. The majority of these bugs can be exposed by reordering a pair of messages, as suggested previously. However, relatively few but critical bugs still occur when more than two messages are involved. These bugs can only be exposed under special timing conditions, involving, for example, some specific messages or events (e.g., node crashes or reboots). To detect such errors, we have empowered our approach with the capability of reordering an arbitrary number of messages for an application.

4. ARCHITECTURE DIAGRAM



5. DATA FLOW DIAGRAM



6. CONCLUSION

We present CLOUDRAID, a simple yet effective tool for detecting distributed concurrency bugs. CLOUDRAID achieves its efficiency and effectiveness by analyzing message orderings that are likely to expose errors from existing logs. Our evaluation shows that

CLOUDRAID is simple to deploy and effective in detecting bugs. In particular, CLOUDRAID can test 60 versions of six representative systems in 35 hours, finding successfully 31 bugs, including 9 new bugs that have never been reported before.

7. REFERENCES

- [1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, Jan. 2008. [Online]. Available: <http://doi.acm.org/10.1145/1327452.1327492>
- [2] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth, B. Saha, C. Curino, O. O'Malley, S. Radia, B. Reed, and E. Baldeschwieler, "Apache hadoop yarn: Yet another resource negotiator," in *Proceedings of the 4th Annual Symposium on Cloud Computing*, ser. SOCC '13. New York, NY, USA: ACM, 2013, pp. 5:1–5:16. [Online]. Available: <http://doi.acm.org/10.1145/2523616.2523633>
- [3] L. George, *HBase: the definitive guide: random access to your planet-size data*. "O'Reilly Media, Inc.", 2011.
- [4] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010.
- [5] Z. Guo, S. McDirmid, M. Yang, L. Zhuang, P. Zhang, Y. Luo, T. Bergan, P. Bodik, M. Musuvathi, Z. Zhang, and L. Zhou, "Failure recovery: When the cure is worse than the disease," in *Proceedings of the 14th USENIX Conference on Hot Topics in Operating Systems*, ser. HotOS'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 8–8. [Online]. Available: <http://dl.acm.org/citation.cfm?id=2490483.2490491>
- [6] D. Yuan, Y. Luo, X. Zhuang, G. R. Rodrigues, X. Zhao, Y. Zhang, P. U. Jain, and M. Stumm,

Certificate of Publication



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN:2320 - 2882

An International Open Access , Peer-reviewed, Refereed Journal

The Board of
International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

KATRAJUMOUNIKA

In recognition of the publication of the paper entitled
**INTELLIGENT AGENT BASED JOB SEARCH SYSTEM USING
DJANGO.**

Published In IJCRT(www.ijert.org) & 7.97 Impact Factor by Google Scholar
Volume 8 Issue 8 , Date of Publication: August 2020 2020-08-04 08:01:54




EDITOR IN CHIEF

Registration ID :197593

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS IJCRT
An International scholar, open Access, Multi-disciplinary, Indexed Journal
Website:www.ijert.org | Email:editor@ijert.org | ESTD:2013

IJCRT | ISSN: 2320-2882 | IJCRT.ORG



GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES

Ketali Soujanya (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

In this paper Wikipedia cloning involves designing, writing, and coding a website in a way that helps to improve the volume and quality of traffic to your website from people using informative website. This website will be an informative website which gives any information which is to be needed by the users exactly like a Wikipedia. Informative websites are built for the purpose of providing information. They can include anything like News website, Science Websites, Encyclopaedia etc.

1. INTRODUCTION

The sequence-to-sequence framework has demonstrated success in natural-language sequence transduction tasks such as machine translation. More recently, neural techniques have been applied to do single-document, abstractive (paraphrasing) text summarization of news articles (Rush et al. (2015), Nallapati et al. (2016)). In this prior work, the input to supervised models ranged from the first sentence to the entire text of an article, and they are trained end-to-end to predict reference summaries. Doing this end-to-end requires a significant number of parallel article-summary pairs since language understanding is a pre-requisite to generate fluent summaries. In contrast, we consider the task of multi-document summarization, where the input is a collection of related documents from which a summary is distilled. Prior work has focused on extractive summarization, which select sentences or phrases from the input to form the summaries, rather than generating new text. There has been limited application of abstractive neural methods and one possible reason is the paucity of

large, labeled datasets. In this work, we consider English Wikipedia as a supervised machine learning task for multidocument summarization where the input is comprised of a Wikipedia topic (title of article) and a collection of non-Wikipedia reference documents, and the target is the Wikipedia article text. We describe the first attempt to abstractively generate the first section, or lead, of Wikipedia articles conditioned on reference text. In addition to running strong baseline models on the task, we modify the Transformer architecture (Vaswani et al., 2017) to only consist of a decoder, which performs better in the case of longer input sequences compared to recurrent neural network (RNN) and Transformer encoder-decoder models. Finally we show our modeling improvements allow us to generate entire Wikipedia articles.

The existence of an abundance of dynamic and heterogeneous information on the Web has offered many new opportunities for users to advance their knowledge discovery. As the amount of information on the Web has increased substantially in



the past decade, it is difficult for users to find information through a simple sequential inspection of web pages or recall previously accessed URLs. Consequently, the service from a search engine becomes indispensable for users to navigate around the Web in an effective manner.

PROPOSED SYSTEM

Here we propose to create an exact replica of Wikipedia website which is a Informative website which is very helpful in getting any kind information present in the whole world which is free and more informative for the users. This website are updated by admins or by users as well. Admin only checks the authentication of the website content. This is an effective way to learn for all like students, Businessmen's, Researchers, Politicians, and Actors etc.

ADVANTAGES

anyone can edit

easy to use and learn

Wikis are instantaneous so there is no need to wait for a publisher to create a new edition or update information

people located in different parts of the world can work on the same document

the wiki software keeps track of every edit made and it's a simple process to revert back to a previous version of an article

widens access to the power of web publishing to non-technical users

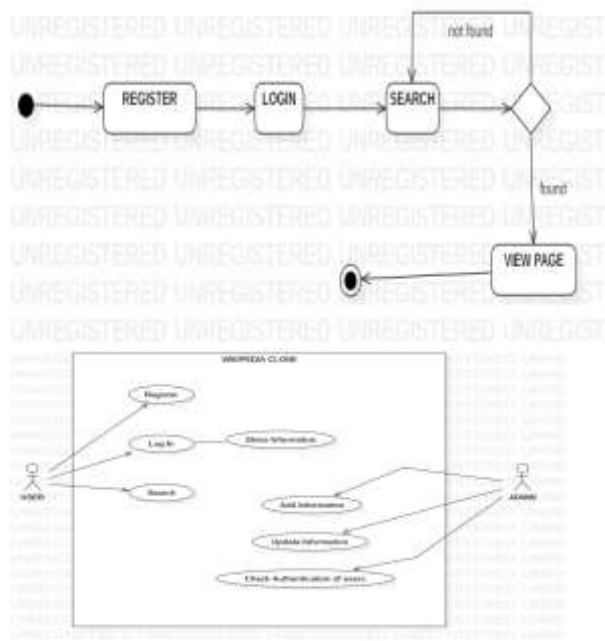
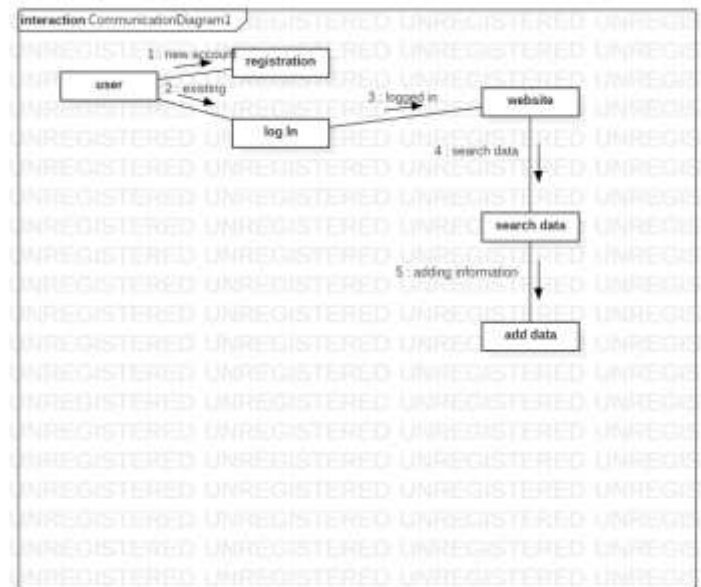
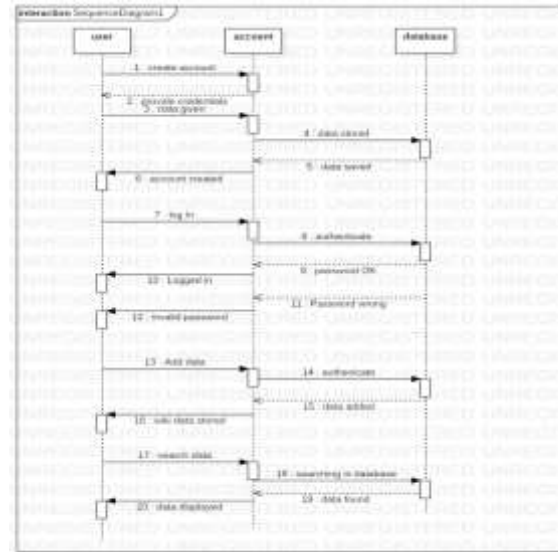
Wikipedia has no predetermined structure – consequently it is a flexible tool which can be used for a wide range of applications

There are a wide range of open source software wiki's to choose from so licensing costs shouldn't be a barrier to installing an institutional Wikipedia.

2. RELATED WORK

2.1 OTHER DATASETS USED IN NEURAL ABSTRACTIVE SUMMARIZATION Neural abstractive summarization was pioneered in Rush et al. (2015), where they train headline generation models using the English Gigaword corpus (Graff & Cieri, 2003), consisting of news articles from number of publishers. However, the task is more akin to sentence paraphrasing than summarization as only the first sentence of an article is used to predict the headline, another sentence. RNN-based encoder-decoder models with attention (seq2seq) perform very well on this task in both ROUGE (Lin, 2004), an automatic metric often used in summarization, and human evaluation (Chopra et al., 2016). In Nallapati et al. (2016), an abstractive summarization dataset is proposed by modifying a questionanswering dataset of news articles paired with story highlights from Daily Mail and CNN. This task is more difficult than headline-generation because the information used in the highlights may come from many parts of the article and not only the first sentence. One downside of the dataset is that it has an order-of-magnitude fewer parallel examples (310k vs. 3.8M) to learn from. Standard seq2seq models with attention do less well, and a number of techniques are used to augment performance. Another downside is that it is unclear what the guidelines are for creating story highlights and it is obvious that there are significant stylistic differences between the two news publishers. In our work we also train neural abstractive models, but in the multi-document regime with Wikipedia. As can be seen in Table 1, the input and output

text are generally much larger, with significant variance depending on the article. The summaries (Wikipedia lead) are multiple sentences and sometimes multiple paragraphs, written in a fairly uniform style as encouraged by the Wikipedia Manual of Style¹. However, the input documents may consist of documents of arbitrary style originating from arbitrary sources. We also show in Table 1 the ROUGE-1 recall scores of the output given the input, which is the proportion of unigrams/words in the output co-occurring in the input. A higher score corresponds to a dataset more amenable to extractive summarization. In particular, if the output is completely embedded somewhere in the input (e.g. a wiki-clone), the score would be 100. Given a score of only 59.2 compared to 76.1 and 78.7 for other summarization datasets shows that ours is the least amenable to purely extractive methods.



3. CONCLUSION

We have shown that generating Wikipedia can be approached as a multi-document summarization problem with a large, parallel dataset, and demonstrated a two-stage extractive-abstractive framework for carrying it out. The coarse extraction method used in the first stage appears to have a significant effect on final performance, suggesting further research on improving it would be fruitful. We introduce a new, decoder-only sequence transduction model for the abstractive stage, capable of handling very long input-



output examples. This model significantly outperforms traditional encoder decoder architectures on long sequences, allowing us to condition on many reference documents and to generate coherent and informative Wikipedia articles

4. REFERENCES

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.
- Sumit Chopra, Michael Auli, and Alexander M Rush. Abstractive sentence summarization with attentive recurrent neural networks. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–98, 2016.
- Hoa Trang Dang. Overview of duc 2005. In Proceedings of the document understanding conference, volume 2005, pp. 1–12, 2005.
- David Graff and Christopher Cieri. English gigaword 2003. Linguistic Data Consortium, Philadelphia, 2003.
- Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. Wikireading: A novel large-scale language understanding task over wikipedia. arXiv preprint arXiv:1608.03542, 2016.
- Rémi Lebre, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pp. 1203–1213, 2016. URL <http://aclweb.org/anthology/D/D16/D16-1128.pdf>.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia—a largescale, multilingual knowledge base extracted from wikipedia. *SemanticWeb*, 6(2):167–195, 2015.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop, volume 8. Barcelona, Spain, 2004.
- Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing, 2004.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, C, a glar Gulc,ehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016*, pp. 280, 2016.
- Ani Nenkova and Lucy Vanderwende. The impact of frequency on summarization. Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005, 101, 2005.



International Journal For Advanced Research In Science & Technology

A peer reviewed international journal

www.ijarst.in

ISSN: 2457-0362

IJARST

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking:

Bringing order to the web. Technical report, Stanford InfoLab, 1999.

CRYPTO CURRENCY PRICES WITH AI

Koduri Divya Meghana (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Cryptocurrency is playing an increasingly important role in reshaping the financial system due to its growing popular appeal and merchant acceptance. While many people are making investments in Cryptocurrency, the dynamical features, uncertainty, the predictability of Cryptocurrency are still mostly unknown, which dramatically risk the investments. It is a matter to try to understand the factors that influence the value formation. In this study, we use advanced artificial intelligence frameworks of fully connected Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) Recurrent Neural Network to analyze the price dynamics of Bitcoin, Ethereum, and Ripple. We find that ANN tends to rely more on long-term history while LSTM tends to rely more on short-term dynamics, which indicate the efficiency of LSTM to utilize useful information hidden in historical memory is stronger than ANN. However, given enough historical information ANN can achieve a similar accuracy, compared with LSTM. This study provides a unique demonstration that Cryptocurrency market price is predictable. However, the explanation of the predictability could vary depending on the nature of the involved machine-learning model.

1.INTRODUCTION

Cryptocurrency is the peer-to-peer digital money and payment system that exist online via a controlled algorithm. When a miner cracks an algorithm to record a block of transactions to public ledger named blockchain and the cryptocurrency is created when the block is added to the blockchain. It allows people to store and transfer through encryption protocol and distributed network. Mining is a necessary and competitive component of the cryptocurrency system. The miner with more computational power has a better chance of finding a new coin than that of less. Bitcoin is the first and one of the leading digital currencies (its market capitalisation had more than \$ 7 billion in 2014, and then it increased significantly to \$ 29 billion in 2017) which was first introduced by Satoshi Nakamoto in 2008. Among many features of bitcoin, the most impressive one is decentralisation that it can remove the

involvement of traditional financial sectors and monetary authorities effectively due to its blockchain network features . In addition, the electronic payment system of Bitcoin is based on cryptographic proof rather than the trust between each other as its transaction history cannot be changed unless redoing all proof of work of all blockchain, which play a critical role of being a trust intermediary and this can be widely used in reality such as recording charitable contribution to avoid corruption. Moreover, bitcoin has introduced the controllable anonymity scheme, and this enhances users' safety and anonymity by using this technology, for instance, we can take advantage of this property of blockchain to make identification cards, and it not only can protect our privacy but verify our identity. Nowadays, investing in cryptocurrencies, like Bitcoin, is one of the efficient ways of earning money. For example, the rate of Bitcoin significant rises in 2017, from a relatively low point 963 USD on January 1ST 2017, to its peak 19186 USD on December 17th 2017, and it closed with 9475 USD at the end of the year. Consequently, the rate of return of bitcoin investment for 2017 was over 880%, which is an impressive and surprising scenery for most investors. While an increasing number of people are making investments in Cryptocurrency, the majority of investors cannot get such profit for being inconsiderable to cryptocurrencies' dynamics and the critical factors that influence the trends of bitcoins. Therefore, raising people's awareness of vital factors can help us to be wise investors. Although market prediction is demanding for its complex nature the dynamics are predictable and understandable to some degree. For example, when there is a shortage of the bitcoin, its price will be increased by their sellers as investors who regard bitcoin as a profitable investment opportunity will have a strong desire to pay for bitcoin. Furthermore, the price of bitcoin may be easily influenced by some influential external factors such as political factors . Although existing efforts on Cryptocurrency analysis and prediction is limited, a few studies have been aiming to understand the Cryptocurrency time series and build statistical models to reproduce and predict price dynamics. For example, Madan et al. collected bitcoins price with the time interval of 0.5, 1 and 2 hours, and combined it with the blockchain network, the underlying technology of bitcoin. Their predictive model leveraging random forests and binomial logistic regression classifiers , and the precision of the model is around 55% in predicting bitcoin's price. Shah et al. used Bayesian regression and took advantages of high frequency (10-second) prices data of Bitcoin to improve investment strategy of bitcoin . Their models had also achieved great success. In an Multi-Layer Perceptron (MLP) based prediction model was presented to forecast the next day price of bitcoin by using two sets of input: the first type of inputs: the opening, minimum, maximum and closing price and the second set of inputs: Moving

Average of both short (5,10,20 days) and long (100, 200 days) windows. During validation, their model was proved to be accurate at the 95% level. There has been many academic researches looking at exchange rate forecasting, for example, the monetary and portfolio balance models examined by Meese and Rogoff (1983, 1988) . Significant efforts have been made to analyse and predict the trends of traditional financial markets especially the stock market however, predicting cryptocurrencies market prices is still at an early stage. Compared to these stock price prediction models, traditional time series methods are not very useful as cryptocurrencies are not precisely the same with stocks but can be deemed as a complementary good of existing currency system with sharp fluctuations features. Therefore, it is urgently needed to understand the dynamics of cryptocurrencies better and establish a suitable predictive modelling framework. In this study, we hypothesise that time series of cryptocurrencies exhibits a clear internal memory, which could be used to help the memory-based time series model to works more appropriately if the length of internal memory could be quantified. We aim to use two artificial intelligence modelling frameworks to understand and predict the most popular cryptocurrencies price dynamics, including Bitcoin, Ethereum, and Ripple.

2.EXISTING SYSTEM

Although existing efforts on Cryptocurrency analysis and prediction is limited, a few studies have been aiming to understand the Cryptocurrency time series and build statistical models to reproduce and predict price dynamics. While an increasing number of people are making investments in Cryptocurrency, the majority of investors cannot get such profit for being inconsiderable to cryptocurrencies' dynamics and the critical factors that influence the trends of bitcoins.

DISADVANTAGES OF EXISTING SYSTEM:

Therefore, raising people's awareness of vital factors can help us to be wise investors. Although market prediction is demanding for its complex nature, the dynamics are predictable and understandable to some degree.

3.PROPOSED SYSTEM

Among many features of bitcoin, the most impressive one is decentralisation that it can remove the involvement of traditional financial sectors and monetary authorities effectively due to its blockchain network features. In addition, the electronic payment system of Bitcoin

is based on cryptographic proof rather than the trust between each other as its transaction history cannot be changed unless redoing all proof of work of all blockchain, which play a critical role of being a trust intermediary and this can be widely used in reality such as recording charitable contribution to avoid corruption.

ADVANTAGES OF PROPOSED SYSTEM:

The bitcoin has introduced the controllable anonymity scheme, and this enhances users' safety and anonymity by using this technology, for instance, we can take advantage of this property of blockchain to make identification cards, and it not only can protect our privacy but verify our identity.

4.LITERATURE SURVEY

1) Using the Bitcoin Transaction Graph to Predict the Price of Bitcoin

AUTHORS: Greaves, A., & Au, B.

with different objectives. A pre-defined set of minimum qualification levels should be distributed between the crew members with minimum training time differences, training expenses or a maximum of the training level with a limitation of the budget. First, a description of the cosmonaut training process is given. Then four models are considered for the volume planning problem. The objective of the first model is to minimize the differences between the total time of the preparation of all crew members, the objective of the second one is to minimize the training expenses with a limitation of the training level, and the objective of the third one is to maximize the training level with a limited budget. The fourth model considers the problem as an n -partition problem. Then two models are considered for the calendar planning problem.

We consider the problem of planning the ISS cosmonaut training with different objectives. A pre-defined set of minimum qualification levels should be distributed between the crew members with minimum training time differences, training expenses or a maximum of the training level with a limitation of the budget. First, a description of the cosmonaut training process is given. Then four models are considered for the volume planning problem. The objective of the first model is to minimize the differences between the total time of the preparation of all crew members, the objective of the second one is to minimize the training expenses with a limitation of the training level, and the objective of the third one is to

maximize the training level with a limited budget. The fourth model considers the problem as an n -partition problem. Then two models are considered for the calendar planning problem.

Bitcoin is the world's leading cryptocurrency, allowing users to make transactions securely and anonymously over the Internet. In recent years, The Bitcoin the ecosystem has gained the attention of consumers, businesses, investors and speculators alike. While there has been significant research done to analyze the network topology of the Bitcoin network, limited research has been performed to analyze the network's influence on overall Bitcoin price. In this paper, we investigate the predictive power of blockchain network-based features on the future price of Bitcoin. As a result of blockchain-networkbased feature engineering and machine learning optimization, we obtain up-down Bitcoin price movement classification accuracy of roughly 55%. We consider the problem of planning the ISS cosmonaut training with different objectives. A pre-defined set of minimum qualification levels should be distributed between the crew members with minimum training time differences, training expenses or a maximum of the training level with a limitation of the budget. First, a description of the cosmonaut training process is given. Then four models are considered for the volume planning problem. The objective of the first model is to minimize the differences between the total time of the preparation of all crew members, the objective of the second one is to minimize the training expenses with a limitation of the training level, and the objective of the third one is to maximize the training level with a limited budget. The fourth model considers the problem as an n -partition problem. Then two models are considered for the calendar planning problem.

For the volume planning problem, two algorithms are presented. The first one is a heuristic with a complexity of (n) operations. The second one consists of a heuristic and exact parts, and it is based on the n -partition problem appro

AUTHORS: Hayes, A. S.

This paper aims to identify the likely source(s) of value that cryptocurrencies exhibit in the marketplace using cross sectional empirical data examining 66 of the most used such 'coins'. A regression model was estimated that points to three main drivers of cryptocurrency value: the difficulty in 'mining 'for coins; the rate of unit production; and the cryptographic algorithm employed. These amount to relative differences in the cost of production of one coin over another at the margin, holding all else equal. Bitcoin-denominated relative prices were used, avoiding much of the price volatility associated with the dollar exchange rate. The

resulting regression model can be used to better understand the drivers of relative value observed in the emergent area of cryptocurrencies. Using the above analysis, a cost of production model is proposed for valuing bitcoin, where the primary input is electricity. This theoretical model produces useful results for both an individual producer, by setting breakeven points to start and stop production, and for the bitcoin exchange rate on a macro level. Bitcoin production seems to resemble a competitive commodity market; in theory miners will produce until their marginal costs equal their marginal product.

3. Economic prediction using neural networks: the case of IBM daily stock returns

AUTHORS: H. White

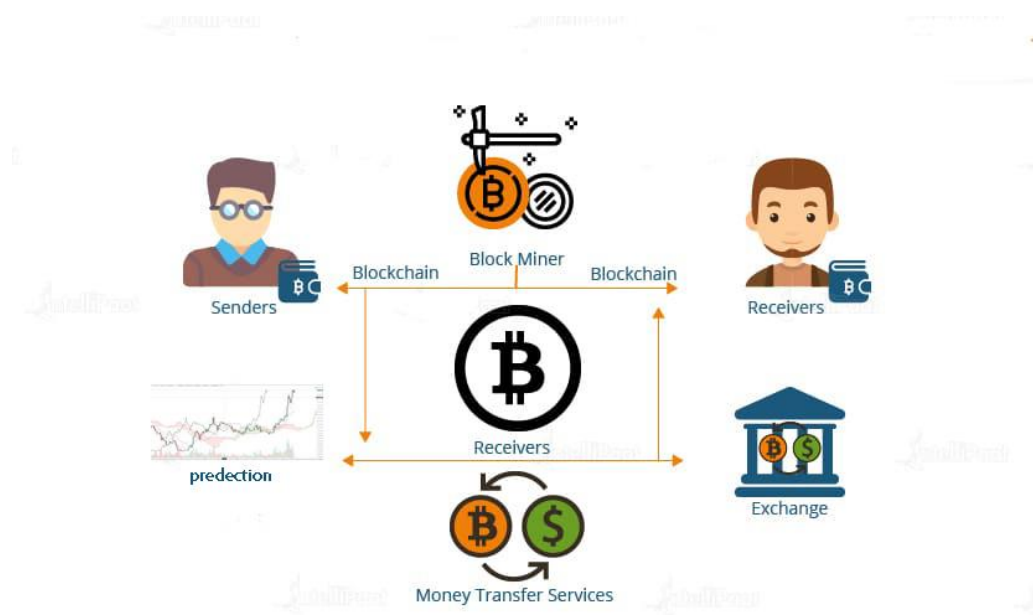
A report is presented of some results of an ongoing project using neural-network modeling and learning techniques to search for and decode nonlinear regularities in asset price movements. The author focuses on the case of IBM common stock daily returns. Having to deal with the salient features of economic data highlights the role to be played by statistical inference and requires modifications to standard learning techniques which may prove useful in other contexts

2). Designing a neural network for forecasting financial and economic time series

AUTHORS: Kaastra and M. Boyd

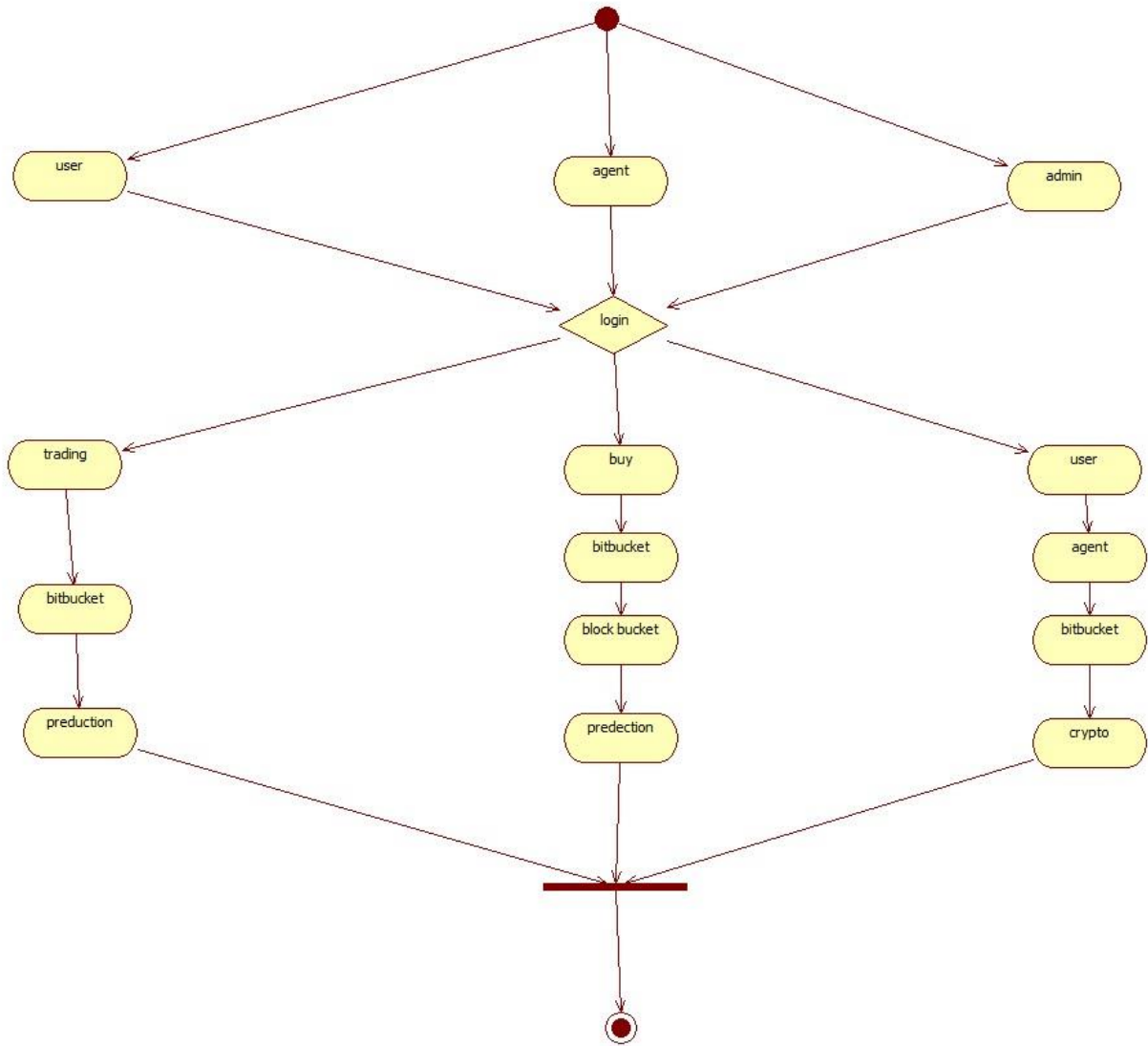
Artificial neural networks are universal and highly flexible function approximators first used in the fields of cognitive science and engineering. In recent years, neural network applications in finance for such tasks as pattern recognition, classification, and time series forecasting have dramatically increased. However, the large number of parameters that must be selected to develop a neural network forecasting model have meant that the design process still involves much trial and error. The objective of this paper is to provide a practical introductory guide in the design of a neural network for forecasting economic time series data. An eight-step procedure to design a neural network forecasting model is explained including a discussion of tradeoffs in parameter selection, some common pitfalls, and points of disagreement among practitioners.

SYSTEM ARCHITECTURE:



ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.



5.SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the

proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

6.CONCLUSION

Cryptocurrency, such as Bitcoin, has established itself as the leading role of decentralisation. There are a large number of cryptocurrencies sprang up after Bitcoin such as Ethereum and Ripple. Because of the significant uncertainty in its prices, many people hold them as a means of speculation. Therefore, it is critically important to understand the internal features

and predictability of those cryptocurrencies. In this study, we use two distinct artificial intelligence frameworks, namely, fully-connected Artificial Neural Network (ANN) and Long-Short-Term-Memory (LSTM) to analyse and predict the price dynamics of Bitcoin, Ethereum, and Ripple. We showed that the ANN and LSTM models are comparable and both reasonably well enough in price prediction, although the internal structures are different. Then we further analyse the influence of historical memory on model prediction. We find that ANN tends to rely more on long-term history while LSTM tends to rely more on short-term dynamics, which indicate the efficiency of LSTM to utilise useful information hidden in historical memory is stronger than ANN. However, given enough historical information ANN can achieve a similar accuracy, compared with LSTM. This study provides a unique demonstration that Cryptocurrency market price is predictable. However, the explanation of the predictability could vary depending on the nature of the involved machine-learning model.

7.REFERENCES

- [1] Greaves, A., & Au, B. (2015). Using the bitcoin transaction graph to predict the price of bitcoin. No Data.
- [2] Hayes, A. S. (2017). Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin. *Telematics and Informatics*, 34(7), 1308-1321.
- [3] Shah, D., & Zhang, K. (2014, September). Bayesian regression and Bitcoin. In *Communication, Control, and Computing (Allerton)*, 2014 52nd Annual Allerton Conference on (pp. 409-414). IEEE.
- [4] Indra N I, Yassin I M, Zabidi A, Rizman Z I. Non-linear autoregressive with exogenous input (mrx) bitcoin price prediction model using so-optimized parameters and moving average technical indicators. *J. Fundam. Appl. Sci.*, 2017, 9(3S), 791-808`
- [5] Adebisi AA, Ayo C K, Adebisi MO, Otokiti SO. Stock price prediction using a neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, 2012, 3(1):1-9
- [6] Adebisi AA, Ayo C K, Adebisi MO, Otokiti SO. Stock price prediction using a neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, 2012, 3(1):1-9

[7] Ariyo AA, Adewumi AO, Ayo CK. Stock price prediction using the ARIMA model. In UKSim-AMSS 16th IEEE International Conference on Computer Modelling and Simulation (UKSim), 2014, pp. 106-112

[8] Ron, D., & Shamir, A. (2013, April). Quantitative analysis of the full bitcoin transaction graph. In International Conference on Financial Cryptography and Data Security (pp. 6-24). Springer, Berlin, Heidelberg.

INTERNET FINANCIAL FRAUD DETECTION BASED ON A DISTRIBUTED BIG DATA APPROACH WITH NODE2VEC

Kokala Viswash (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The rapid development of information technologies like Internet of Things, Big Data, Artificial Intelligence, Blockchain, etc., has profoundly affected people's consumption behaviors and changed the development model of the financial industry. The financial services on Internet and IoT with new technologies has provided convenience and efficiency for consumers, but new hidden fraud risks are generated also. Fraud, arbitrage, vicious collection, etc., have caused bad effects and huge losses to the development of finance on Internet and IoT. However, as the scale of financial data continues to increase dramatically, it is more and more difficult for existing rule-based expert systems and traditional machine learning model systems to detect financial frauds from large-scale historical data. In the meantime, as the degree of specialization of financial fraud continues to increase, fraudsters can evade fraud detection by frequently changing their fraud methods. In this article, an intelligent and distributed Big Data approach for Internet financial fraud detections is proposed to implement graph embedding algorithm Node2Vec to learn and represent the topological features in the financial network graph into low-dimensional dense vectors, so as to intelligently and efficiently classify and predict the data samples of the large-scale dataset with the deep neural network. The approach is distributedly performed on the clusters of Apache Spark GraphX and Hadoop to process the large dataset in parallel. The groups of experimental results demonstrate that the proposed approach can improve the efficiency of Internet financial fraud detections with better precision rate, recall rate, F1-Score and F2-Score.

1. INTRODUCTION

With the rapid development of the information technologies like Internet of Things, Big Data, Artificial Intelligence, Blockchain, etc., the digital life led by financial technology has profoundly affected people's consumption behaviors and changed the development model of the traditional financial industry to a certain extent [1]. In particular, technical products such as mobile payment, IOT financial services and Internet financial wealth management have penetrated into lots of aspects of economic and social activities. From 2014 to the present, the development momentum of China's Internet consumer finance industry has been good, and various mobile e-commerce companies have entered the consumer finance field through installment payments and small loans, which has promoted the development of related industries. Internet financial services based on consumer credits in China, such as Huabei launched by Ant Financial and Alipay of Alibaba Group, Jingdong Baitiao operated by JD.com, WeiLiDai launched by WeBank of Tencent, etc., have enabled consumers to enjoy the online shopping experience of "consumption first, pay later", and covered the e-commerce installment shopping, cash borrowing and other businesses. Especially in 2020, the COVID-19 pandemic [2] has caused a surge in online transaction volume and brought a large number of online customers to online service providers. It has cultivated the habit of more groups of users to make online purchases and payments through mobile phones and IOT devices, which brings continuous impetus to the development of the Internet financial industry.

The rapid development of mobile and IOT financial payment services has not only provided convenience and efficiency for consumers, but also brought more hidden fraud risks. Due to the concealment of the complex network, there could be a breeding ground for fraudulent activities by criminals. The control of fraud risks is becoming more and more difficult and fraud cases occur frequently, which causes the fraud losses to commercial banks and financial institutions are also increasing. The continuous happening of Internet financial fraudulent problems, such as the agreement cash-out incident of Huabei and Taobao merchants, and "Baitiao" multiple fraud incidents, have not only damaged the legitimate rights and interests of the service platform, but also caused consumers to question the company's account security and risk identification capabilities.

A large number of violations are beyond the scope of the industry's existing laws and regulations, and industry regulation has always lagged behind the innovative development of Internet consumer finance, which makes the regulatory laws and regulations

are often in a state of absence so that it impossible to deal with industry violations in a timely manner. Fraud, arbitrage, vicious collection and other phenomena are becoming more and more rampant in online financial service platforms, which has caused bad effects and huge losses to the development of consumer finance on Internet and IOT. Fraud is an illegal or criminal deception aimed at obtaining financial or personal benefits.

Fraud generally has the attributes of abnormal or unfair transactions. Due to the inconsistency with previous fund operation rules or other normal behaviors, fraudulent behavior presents various abnormal characteristics, including abnormal transaction amount, abnormal transaction time, abnormal transaction account, abnormal transaction IP, or abnormal personal credit rating.

2. EXISTING SYSTEM

Allen et al. find that there are many credit channels in the United States and based on the research of American household credit models, and that household consumption, household income, credit banks and credit scale are obviously related [7]. Kregel studies the development trend of consumer finance and finds that the development of Internet consumer finance companies must fully consider the current market legal environment, financial market and consumer behavior factors, etc. Internet consumer finance is directly related to the current development of the national financial system [8]. Momparler et al. take the American Internet consumer finance company as the research object, study the risks and advantages of the Internet consumer finance platform, and design a related risk management model [9].

Allen et al. find that there are many credit channels in the United States and based on the research of American household credit models, and that household consumption, household income, credit banks and credit scale are obviously related [7]. Kregel studies the development trend of consumer finance and finds that the development of Internet consumer finance companies must fully consider the current market legal environment, financial market and consumer behavior factors, etc. Internet consumer finance is directly related to the current development of the national financial system [8]. Momparler et al. take the American Internet consumer finance company as the research object, study the risks and advantages of the Internet consumer finance platform, and design a related risk management model [9].

Ficawoyi et al. analyze the positive relationship between Internet exposure levels and credit card default through surveys on consumer finance and income nodes [14]. The research

points out that Internet access, low income, and male families are more likely to cause credit card defaults. Giudici et al. propose how to improve credit risk accuracy of P2P Internet financial platforms and of those who lend to small and medium enterprises [15]. The augmented traditional credit scoring methods are put forward with “alternative data” that consist of centrality measures derived from similarity networks among borrowers and deduced from their financial ratios. The experimental findings suggest that the proposed approach improves predictive accuracy as well as model explainability.

Disadvantages

- 1) The system doesn't support Resilient Distributed Datasets.
- 2) There is no Directed Acyclic Graph method to find fraud accurately.

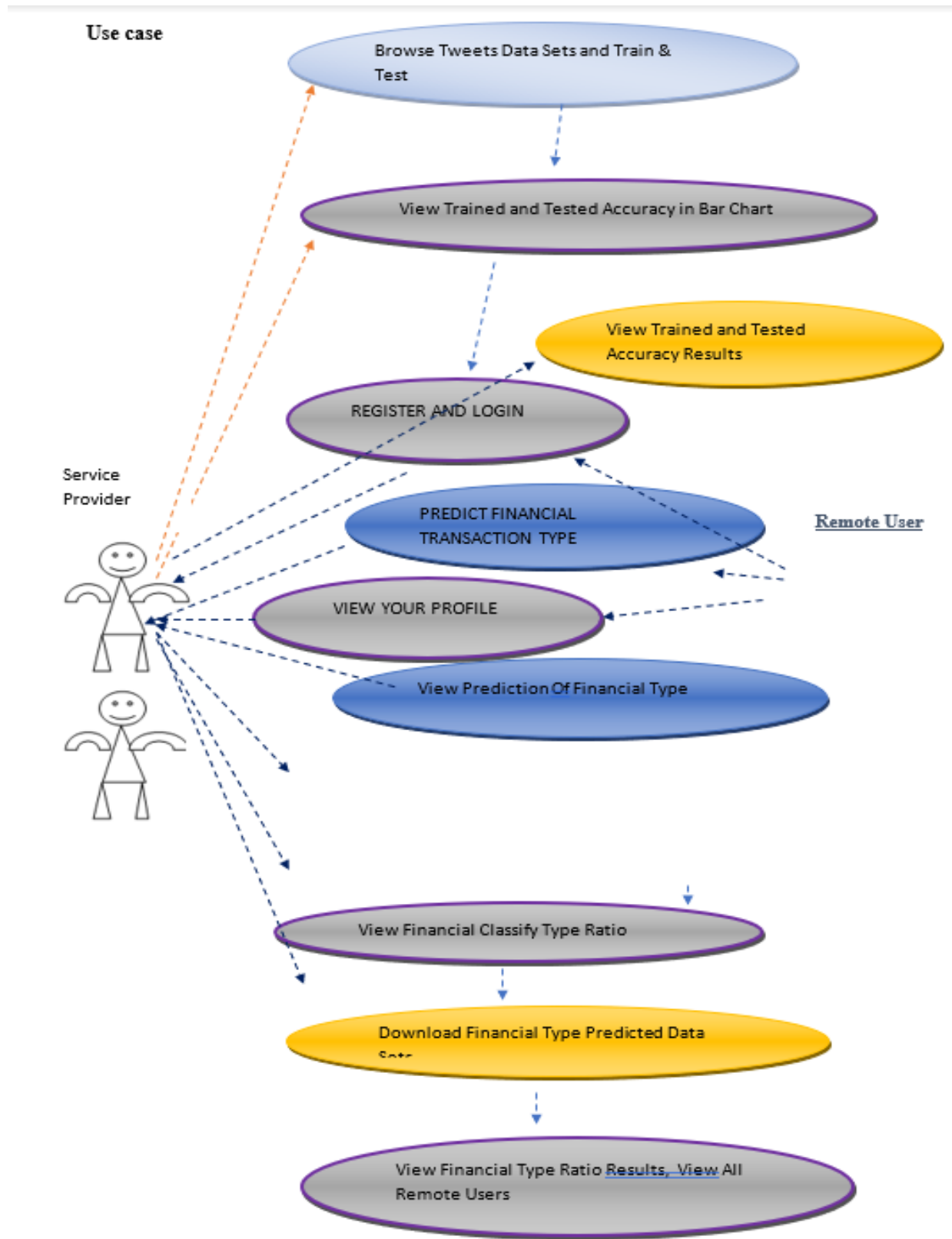
3. PROPOSED SYSTEM

Through studying a large number of Internet financial fraud cases, two important characteristics are found: (1) The pattern of Internet financial fraud continues to evolve and develop over time, not just repeating the existing individual behavior patterns appeared in historical cases; (2) With the advancement of anti-fraud technology, it is getting harder for individuals to commit Internet financial fraud. It needs to be organized and conducted through related and connected groups. A graph is an abstract graph formed by a number of nodes and the edges connecting each node [31], [32]. It is usually used to describe a specific relationship between things. A relational network graph refers to a graph-based data structure composed of nodes and edges. Each node represents an entity, and each edge is the relationship between an entity and the other connected entity. The relationship network graph connects different entities together according to their relationships; thus it could provide the ability to analyze problems from the perspective of "relationship".

In anti-fraud applications, entities in the network graph, such as people, equipment, mailboxes, card numbers, etc., can be represented by nodes, and the relationships between these nodes in the business can be represented by edges. Through continuous construction and reproduction of the associated relationships hidden covertly in Internet financial frauds, fraud characteristics can be detected and corresponding risk control strategies can be designed. The graph algorithms can characterize various high-risk features in the Internet finance, such as batch attacks, intermediary participation, etc., which is more effective to identify abnormal group frauds from normal behaviors.

Advantages

- 1) Node2Vec is a graph embedding algorithm that introduces two biased random walk methods---BFS (Breadth First Search) and DFS (Depth First Search) on the basis of Deep Walk.
- 2) An intelligent and distributed big data approach for internet financial fraud detection.



FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

4. CONCLUSIONS

The occurrences of Internet financial fraud cases have caused huge losses to commercial banks or financial institutions. In order to enhance the efficiency of financial fraud detections, an intelligent and distributed Big Data approach is proposed in this article. The approach mainly includes four modules: data preprocessing module, normal data feature module, graph embedding module, prediction module. The graph embedding algorithm Node2Vec is implemented on Spark GraphX and Hadoop to learn and represent the topological features of each vertex in the network graph into a low- dimensional dense vector, so as to improve the classification effectiveness of deep neural network and predict the fraudulent samples of the dataset. The experiments evaluate the indicators of precision rate, recall rate, F1-Score and F2- Score, and the results show that due to the Node2Vec properties of structural equivalence and homophily, the features of samples can be better learned and represented and the proposed approach is better than the comparative methods. In future work, the inductive graph embedding network algorithms, such as GraphSage, PinSage, etc., would be improved and implemented to effectively learn the features of newly generated vertices in a dynamic network graph, so as to achieve the better effect of financial fraud detection.

5. REFERENCES

- [1] U. Paschen, C. Pitt, and J. Kietzmann, "Artificial intelligence: building blocks and an innovation typology," *Business Horizons*, vol. 63, no. 2, pp. 147-155, 2020.
- [2] P. Yu, Z. Xia, J. Fei, and S. K. Jha, "An application review of artificial intelligence in prevention and cure of COVID-19 pandemic," *CMC-Computers Materials & Continua*, vol. 65, no. 1, pp. 743-760, 2020.
- [3] L. Shen, X. Chen, Z. Pan, K. Fan, F. Li, and J. Lei, "No-reference stereoscopic image quality assessment based on global and local content characteristics," *Neurocomputing*, vol. 424, no. 2, pp. 132- 142, 2021.

- [4] H. Beck, "Banking is essential, banks are not, the future of financial intermediation in the age of the Internet," *Netnomics*, vol. 3, no. 1, pp. 7-22, 2001.
- [5] G. N. Weiss, K. Pelger, and A. Horsch, "Mitigating adverse selection in p2p lending—Empirical evidence from prosper.com," *SSRN Electronic Journal*, vol. 19, no. 7, pp. 65-93, 2010.
- [6] Y. Houston, C. Jongrong, J. H. Cliff, and H. Y. Chih, "E-commerce, R&D, and productivity: firm-level evidence from Taiwan," *Information Economics and Policy*, vol. 18, no. 5, pp. 561-569, 2013.
- [7] F. Allen, J. Mcandrews, and P. Strahan, "E-finance: an introduction," *Center for Financial Institutions Working Papers*, vol. 22, no. 1, pp. 25-27, 2012.
- [8] J. A. Kregel, "Margins of safety and weight of the argument in generating financial fragility," *Journal of Economics Issues*, vol. 6, no. 31, pp. 543-548, 2016.
- [9] A. Momparler, C. Lassala, and D. Ribeiro, "Efficiency in banking services: a comparative analysis of Internet-primary and branching banks in the US," *Service Business*, vol. 7, no. 4, pp. 641-663, 2013.
- [10] V. Jambulapati and J. Stavins, "Credit card act of 2009: what did banks do?," *Banking & Finance*, vol. 46, no. 9, pp. 21-30, 2014.
- [11] H. Shefrin and C. M. Nicols, "Credit card behavior, financial styles and heuristics," *Business Research*, vol. 67, no. 8, pp. 1679-1687, 2014. [12] C. B. Hem and D. A. Ficawoyi, "Internet consumer spending and credit card



DETECTION OF MALICIOUS SOCIAL BOTS USING LEARNING AUTOMATA WITH URL FEATURES IN TWITTER NETWORK

Koppula Vijaya Ramya Sri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Malicious social bots generate fake tweets and automate their social relationships either by pretending like a follower or by creating multiple fake accounts with malicious activities. Moreover, malicious social bots post shortened malicious URLs in the tweet in order to redirect the requests of online social networking participants to some malicious servers. Hence, distinguishing malicious social bots from legitimate users is one of the most important tasks in the Twitter network. To detect malicious social bots, extracting URL-based features (such as URL redirection, frequency of shared URLs, and spam content in URL) consumes less amount of time in comparison with social graph-based features (which rely on the social interactions of users). Furthermore, malicious social bots cannot easily manipulate URL redirection chains. In this article, a learning automata-based malicious social bot detection (LA-MSBD) algorithm is proposed by integrating a trust computation model with URL-based features for identifying trustworthy participants (users) in the Twitter network. The proposed trust computation model contains two parameters, namely, direct trust and indirect trust. Moreover, the direct trust is derived from Bayes' theorem, and the indirect trust is derived from the Dempster–Shafer theory (DST) to determine the trustworthiness of each participant accurately. Experimentation has been performed on two Twitter data sets, and the results illustrate that the proposed algorithm achieves improvement in precision, recall, F-measure, and accuracy compared with existing approaches for MSBD.

1. INTRODUCTION

MALICIOUS social bot is a software program that pretends to be a real user in online social networks (OSNs) [1], [2]. Moreover, malicious social bots perform several malicious attacks, such as spread social spam content, generate fake identities, manipulate online ratings, and perform phishing attacks [1]. In Twitter, when a participant (user) wants to share a tweet containing URL(s) with the neighboring participants (i.e., followers or followers), the participant adapts URL shortened service (i.e., bit.ly [3]) in order

to reduce the length of URL (because a tweet is restricted up to 140 characters). Moreover, a malicious social bot may post shortened phishing URLs in the tweet [4]. As shown in Fig. 1, when participant clicks on a shortened phishing URL, the participant's request will be redirected to intermediate URLs associated with malicious servers that, in turn, redirect the user to malicious web pages. Then, the legitimate participant is exposed to an attacker. This leads to Twitter network suffering from several vulnerabilities (such as phishing attack). Several approaches



have been proposed to detect spam in the Twitter network [5]–[8]. These approaches are based on tweet-content features, social relationship features, and user profile features.

However, the malicious social bots can manipulate profile features, such as hashtag ratio, follower ratio, URL ratio, and the number of re tweets. The malicious social bots can also manipulate tweet-content features, such as sentimental words, emoticons, and most frequent words used in the tweets, by manipulating the content of each tweet [9]. The social relationship-based features are highly robust because the malicious social bots cannot easily manipulate the social interactions of users in the Twitter network. However, extracting social relationship-based features consumes a huge amount of time due to the massive volume of social network graph [10]. Therefore, identifying the malicious social bots from the legitimate participants is a challenging task in the Twitter network. The existing malicious URL detection approaches [11], [12] are based on DNS information and lexical properties of URLs. The malicious social bots use URL redirections in order to avoid detection [13]. However, for detectors, identification of all malicious social bots is an issue because malicious social bots do not post malicious URLs directly in the tweets. Thus, it is important to identify malicious URLs (i.e., harmful URLs) posted by malicious social bots in Twitter. Most of the existing approaches [14], [15] are based on supervised learning algorithms, where the model is trained with the labeled data in order to detect malicious bots in

OSNs. However, these approaches rely on statistical features instead of analyzing the social behavior of users [16].

Moreover, these approaches are not highly robust in detecting the temporal data patterns with noisy data (i.e., where the data is biased with untrustworthy or fake information) because the behavior of malicious bots changes over time in order to avoid detection [17], [18]. This motivated us to consider one of the reinforcement learning techniques (such as the learning automata (LA) model) to handle temporal data patterns. In this work, we design an LA model to detect malicious social bots with improved precision and recall. In this article, the malicious behavior of participants is analyzed by considering features extracted from the posted URLs (in the tweets), such as URL redirection, frequency of shared URLs, and spam content in URL, to distinguish between legitimate and malicious tweets. To protect against the malicious social bot attacks, our proposed LA-based malicious social bot detection (LA-MSBD) algorithm integrates a trust computational model with a set of URL-based features for the detection of malicious social bots. The proposed trust computational model contains two parameters, namely, direct trust and indirect trust. The direct trust value is derived from the Bayesian learning [19] (by considering URL-based features) to determine the trustworthiness of tweets posted by each participant. In addition to the direct trust, belief values (i.e., indicators for determining indirect trust) are collected from multiple neighbors of a participant. This is due to



the fact that in case the neighbors of a participant are trustworthy, the participant is likely to be trustworthy. Furthermore, Dempster's combination rule [20] aggregates the belief values provided by multiple one-hop neighboring participants in order to evaluate the indirect trust value of participants in the Twitter network.

2. EXISTING SYSTEM

Besel et al. [23] analyzed social botnet attack on Twitter. The authors have presented that social bots use URL shortening services and URL redirection in order to redirect users to malicious web pages. Echeverria and Zhou [24] presented methods to detect, retrieve, and analyze botnet over thousands of users to observe the social behavior of bots. In [25], a social bot hunter model has been presented based on the user behavioral features, such as follower ratio, the number of URLs, and reputation score. In [26], a trust model has been designed to detect malicious activities in an OSN. The authors analyzed that the low trust value of a user indicates that the information spread by the user is considered as untrustworthy.

In [1], an MSBD approach has been proposed by considering user behavioral features, such as commenting, liking, and sharing. Madisetty and Desarkar [5] have developed five different convolutional neural network models by considering tweet features. In [27], a social botnet detection algorithm is proposed by considering spam content in tweets and trust to identify social bots. Gupta et al. [7] designed a framework for detecting spammers in the Twitter network using different machine learning algorithms. In this article, we focus to detect malicious social bots (who perform phishing attacks)

by considering various URL-based features using an LA model.

Several spam-detection approaches have been proposed in the Twitter network to distinguish non spam accounts and spam accounts [5]–[8]. Moreover, these studies consider user profile features, which can easily be modified by malicious bots. To avoid feature manipulation, Yang et al. [28] considered social relationships between malicious users and with their neighboring users based on closeness centrality. Moreover, profile features and social interaction features may not help in detecting malicious URLs that are posted by the participants.

To address the above-mentioned problem, Janabi et al. [29] considered URL-based features (such as URL length, Http-302 status code, and disabling right click) to distinguish legitimate URLs from suspicious URLs. In [30], a URL-based approach is proposed to detect spam tweets in Twitter based on the tweet content and URL redirection chains. Patil and Patil [31] used decision tree classifiers with statistical features in order to detect malicious URLs. Moreover, social bots may use malicious URL redirections in order to avoid detection.

Disadvantages

In the existing work, the system considers user profile features, which can easily be modified by malicious bots.

This system aims to profile features and social interaction features which may not help in detecting malicious URLs that are posted by the participants..

3. PROPOSED SYSTEM

The proposed LA-MSBD algorithm helps to detect malicious social bots accurately

(in terms of precision, recall, F-measure, and accuracy) in Twitter. The major contributions are as follows.

Analyze the malicious behavior of a participant by considering URL-based features, such as URL redirection, the relative position of URL, frequency of shared URLs, and spam content in URL.

Evaluate the trustworthiness of tweets (posted by each participant) by using the Bayesian learning and Dempster-Shafer theory (DST).

Design of an LA-MSBD algorithm by integrating a trust model with a set of URL-based features.

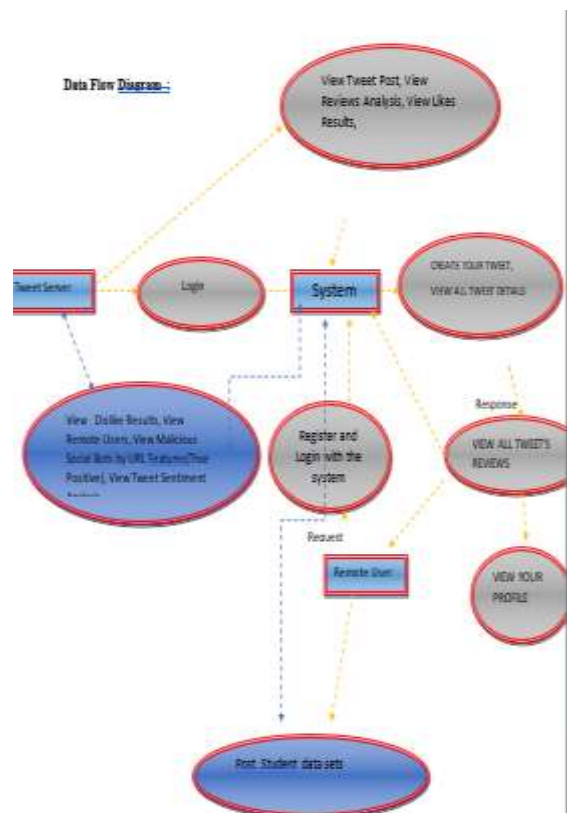
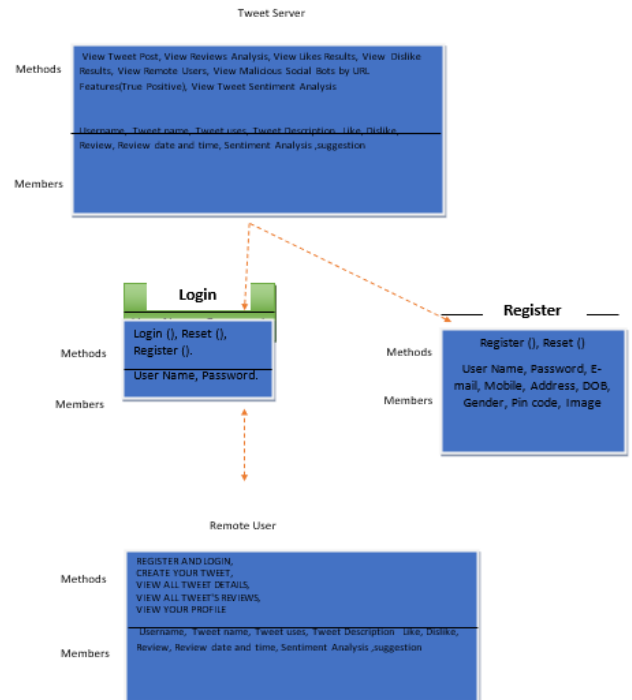
Performance evaluation of the proposed LA-MSBD algorithm using two Twitter data sets, namely, The Fake Project data set [21] and Social Honeypot data set [22] in terms of precision, recall, F-measure, and accuracy for MSBD in the Twitter network.

Advantages

The malicious behavior of participants is analyzed effectively by considering features extracted from the posted URLs (in the tweets), such as URL redirection, frequency of shared URLs, and spam content in URL, to distinguish between legitimate and malicious tweets.

To protect against the malicious social bot attacks, our proposed LA-based malicious social bot detection (LA-MSBD) algorithm integrates a trust computational model with a set of URL-based features for the detection of malicious social bots.

Class Diagram :





4. CONCLUSIONS

This article presents an LA-MSBD algorithm by integrating a trust computational model with a set of URL-based features for MSBD. In addition, we evaluate the trustworthiness of tweets (posted by each participant) by using the Bayesian learning and DST. Moreover, the proposed LA-MSBD algorithm executes a finite set of learning actions to update action probability value (i.e., probability of a participant posting malicious URLs in the tweets). The proposed LA-MSBD algorithm achieves the advantages of incremental learning. Two Twitter data sets are used to evaluate the performance of our proposed LA-MSBD algorithm. The experimental results show that the proposed LA-MSBD algorithm achieves up to 7% improvement of accuracy compared with other existing algorithms. For The Fake Project and Social HoneyPot data sets, the proposed LA-MSBD algorithm has achieved precisions of 95.37% and 91.77% for MSBD, respectively. Furthermore, as a future research challenge, we would like to investigate the dependence among the features and its impact on MSBD.

5. REFERENCES

- [1] J.P. Shi, Z. Zhang, and K.-K.-R. Choo, "Detecting malicious social bots based on clickstream sequences," *IEEE Access*, vol. 7, pp. 28855–28862, 2019.
- [2] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Appl. Intell.*, vol. 49, no. 11, pp. 3947–3964, Nov. 2019.
- [3] D. Choi, J. Han, S. Chun, E. Rappos, S. Robert, and T. T. Kwon, "Bit.ly/practice: Uncovering content publishing and sharing through URL shortening services," *Telematics Inform.*, vol. 35, no. 5, pp. 1310–1323, 2018.
- [4] S. Lee and J. Kim, "Fluxing botnet command and control channels with URL shortening services," *Comput. Commun.*, vol. 36, no. 3, pp. 320–332, Feb. 2013.
- [5] S. Madisetty and M. S. Desarkar, "A neural network-based ensemble approach for spam detection in Twitter," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 4, pp. 973–984, Dec. 2018.
- [6] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1166–1177, Feb. 2015.
- [7] H. Gupta, M. S. Jamal, S. Madisetty, and M. S. Desarkar, "A framework for real-time spam detection in Twitter," in *Proc. 10th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2018, pp. 380–383.
- [8] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proc. Australas. Comput. Sci. Week Multiconf. (ACSW)*, 2017, p. 3.
- [9] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Key challenges in defending against malicious socialbots," Presented at the 5th USENIX Workshop Large-Scale Exploits Emergent Threats, 2012, pp. 1–4.
- [10] G. Yan, "Peri-watchdog: Hunting for hidden botnets in the periphery of online social networks," *Comput. Netw.*, vol. 57, no. 2, pp. 540–555, Feb. 2013.
- [11] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: A fast filter for the large-scale detection of malicious Web pages," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, 2011, pp. 197–206.



[12] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," *J. Ambient Intell.*

ANDROID MALWARE DETECTION USING GENETIC ALGORITHM BASED OPTIMIZED FEATURE SELECTION AND MACHINE LEARNING

Korasikha Lakshmi Prasanna (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Android platform due to open source characteristic and Google backing has the largest global market share. Being the world's most popular operating system, it has drawn the attention of cyber criminals operating particularly through wide distribution of malicious applications. This paper proposes an effectual machine-learning based approach for Android Malware Detection making use of evolutionary Genetic algorithm for discriminatory feature selection. Selected features from Genetic algorithm are used to train machine learning classifiers and their capability in identification of Malware before and after feature selection is compared. The experimentation results validate that Genetic algorithm gives most optimized feature subset helping in reduction of feature dimension to less than half of the original feature-set. Classification accuracy of more than 94% is maintained post feature selection for the machine learning based classifiers, while working on much reduced feature dimension, thereby, having a positive impact on computational complexity of learning classifiers.

1.INTRODUCTION

Android Apps are freely available on Google Playstore, the official Android app store as well as third-party app stores for users to download. Due to its open source nature and popularity, malware writers are increasingly focusing on developing malicious applications for Android operating system. In spite of various attempts by Google Playstore to protect against malicious apps, they still find their way to mass market and cause harm to users by misusing personal information related to their phone book, mail accounts, GPS location information and others for misuse by third parties or else take control of the phones remotely. Therefore, there is need to

perform malware analysis or reverse-engineering of such malicious applications which pose serious threat to Android platforms. Broadly speaking, Android Malware analysis is of two types: Static Analysis and Dynamic Analysis. Static analysis basically involves analyzing the code structure without executing it while dynamic analysis is examination of the runtime behavior of Android Apps in constrained environment. Given in to the ever-increasing variants of Android Malware posing zero-day threats, an efficient mechanism for detection of Android malwares is required. In contrast to signature-based approach which requires regular update of signature database.

2.EXISTINGSYSTEM

The main contribution of the work is reduction of feature dimension to less than half of original feature-set using Genetic Algorithm such that it can be fed as input to machine learning classifiers for training with reduced complexity while maintaining their accuracy in malware classification. In contrast to exhaustive method of feature selection which requires testing for 2^N different combinations, where N is the number of features, Genetic Algorithm, a heuristic searching approach based on fitness function has been used for feature selection. The optimized feature set obtained using Genetic algorithm is used to train two machine learning algorithms: Support Vector Machine and Neural Network. It is observed that a decent classification accuracy of more than 94% is maintained while working on a much lower feature dimension, thereby, reducing the training time complexity of classifiers.

3.PROPOSEDSYSTEM

Two set of Android Apps or APKs: Malware/Goodware are reverse engineered to extract features such as permissions and count of App Components such as Activity, Services, Content Providers, etc. These features are used as featurevector with class labels as Malware and Goodware represented by 0 and 1 respectively in CSV format. To reducedimensionality of feature-set, the CSV is fed to Genetic Algorithm to select the most optimized set of features. The optimized set of features obtained is used for training two machine learning classifiers: Support Vector Machine and Neural Network.

In the proposed methodology, static features are obtained from AndroidManifest.xml which contains all the important information needed by any Android platform about the Apps. Androguard tool has been used for disassembling of the APKs and getting the static features.

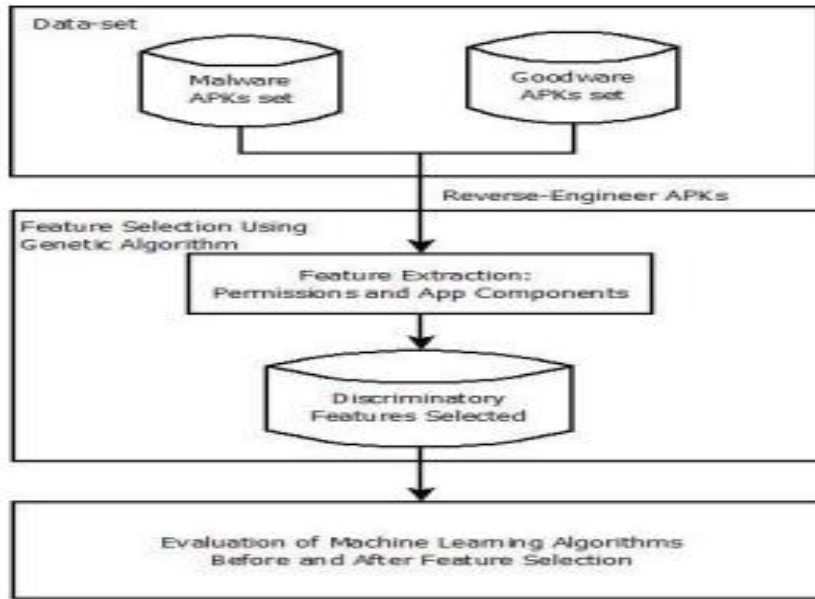
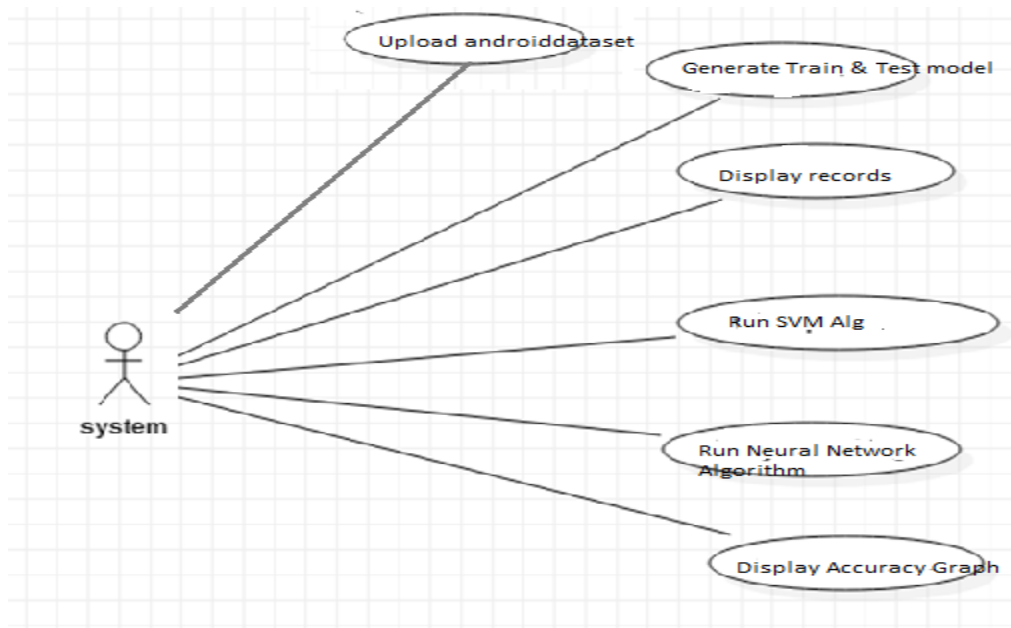
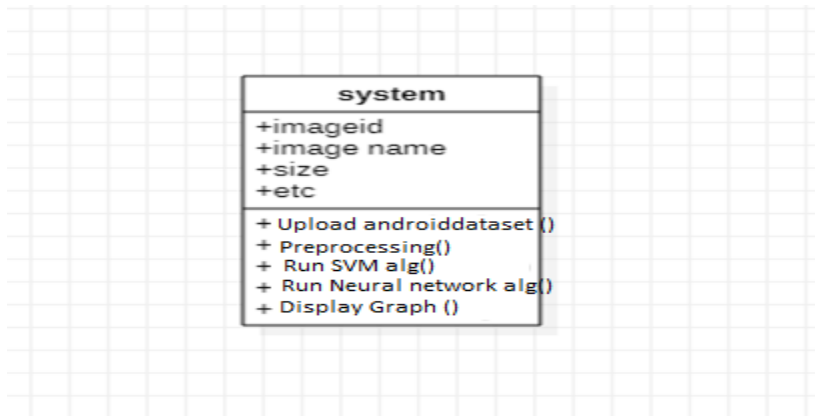


Fig. 1. Proposed Methodology

UML DIAGRAMS



Class Diagram :-**4.TEST RESULTS**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS**4.1Unit testing**

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application

.it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

4.2 Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

4.3 Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted. **Invalid Input** : identified classes of invalid input must be rejected. **Functions** : identified functions must be exercised.

Output : identified classes of application outputs must be exercised. **Systems/Procedures**: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

4.4 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

4.5 White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

4.6 Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format

- No duplicate entries should be allowed
- All links should take the user to the correct page.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

5. CONCLUSION

As the number of threats posed to Android platforms is increasing day to day, spreading mainly through malicious applications or malwares, therefore it is very important to design a framework which can detect such malwares with accurate results. Where signature-based approach fails to detect new variants of malware posing zero-day threats, machine learning based approaches are being used. The proposed methodology attempts to make use of evolutionary Genetic Algorithm to get most optimized feature subset which can be used to train machine learning algorithms in most efficient way.

6. REFERENCES

1. D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, and K. Rieck, "Drebin: Effective and Explainable Detection of Android Malware in Your Pocket," in Proceedings 2014 Network and Distributed System Security Symposium, 2014.
- [2] N. Milosevic, A. Dehghantanha, and K. K. R. Choo, "Machine learning aided Android malware classification," *Comput. Electr. Eng.*, vol. 61, pp. 266–274, 2017.
- [3] J. Li, L. Sun, Q. Yan, Z. Li, W. Srisa-An, and H. Ye, "Significant Permission Identification for Machine-Learning-Based Android Malware Detection," *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3216–3225, 2018.
- [4] A. Saracino, D. Sgandurra, G. Dini, and F. Martinelli, "MADAM: Effective and Efficient Behavior-based Android Malware Detection and Prevention," *IEEE Trans. Dependable Secur. Comput.*, vol. 15, no. 1, pp. 83–97, 2018.

[5] S. Arshad, M. A. Shah, A. Wahid, A. Mehmood, H. Song, and H. Yu, “SAMADroid: A Novel 3-Level Hybrid Malware Detection Model for Android Operating System,” *IEEE Access*, vol. 6, pp. 4321–4339, 2020

AN INTELLIGENT DATA-DRIVEN MODEL TO SECURE INTRAVEHICLE COMMUNICATIONS BASED ON MACHINE LEARNING

Korlepara Sairam Gupta (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

The high relying of electric vehicles on either in vehicle or between-vehicle communications can cause big issues in the system. This paper is going to mainly address the cyber attack in electric vehicles and propose a secured and reliable intelligent framework to avoid hackers from penetration into the vehicles. The proposed model is constructed based on an improved support vector machine model for anomaly detection based on the controller area network (CAN) bus protocol. In order to improve the capabilities of the model for fast malicious attack detection and avoidance, a new optimization algorithm based on social spider (SSO) algorithm is developed which will reinforce the training process at offline. Also, a two-stage modification method is proposed to increase the search ability of the algorithm and avoid premature convergence. Last but not least, the simulation results on the real data sets reveal the high performance, reliability and security of the proposed model against denial-of-service (DoS) hacking in the electric vehicles.

1. INTRODUCTION

Technically, vehicles are composed of many hardware modules namely called electronic control units (ECUs) being controlled by different software tools. All sensors installed in a vehicle will send their data to the ECU, where this data are processed and the requiring orders are sent to the relevant actuators [1]. Such a highly complex hardware software data transfer process may happen through the use of different network protocols such as CAN, LIN, Flex Ray or MOST [2]. Among these protocols, CAN bus is the most popular one not only in vehicles, but also in medical apparatuses, agriculture, etc due to its high capability and promising characteristics. Some of the main advantages of the CAN bus standard may be briefly named as allowing up to 1Mbps data rate transfer, reducing the wiring in the device

saving cost and time due to the simple wiring, auto retransmission of lost messages and error detection capability [3]. Unfortunately, since CAN bus protocol was devised at a time where vehicles were almost isolated, this standard suffers from some security issues in the new dynamic environment of smart grids. This will motivate the hackers to attack the electric vehicles through the ECU and inject malicious messages into their systems. In [4], some cyber intrusion scenarios are modelled and applied on the electric vehicles to assess their vulnerabilities and possible side effects getting finally into the power grid. In [5], a new classification method is developed for cyber intrusion detection in vehicles. In [6], a data intrusion detection system is developed which can detect the cyber attack based on the CAN bus message frequency increase or CAN message ID misuse. This will help the driver to detect that an attack has happened so to stop the vehicle immediately. In [7], authors suggest that all CAN messages should pass a data management system to avoid any cyber intrusion. In [8], an algorithmic solution is used to stop attacks of types of denial-of-service or error flag in the vehicle. In [9], it is suggested to assign an ECU as the master ECU in the manufacturing stage of the vehicle so to run an attestation process in the system. In [10], a firewall is introduced for the vehicle to sit between the CAN bus and the communicating system and stop the cyber attack commands to the CAN bus. In [11], an intrusion detection system based on the traffic entropy of in-vehicle network communication system of the CAN bus is suggested. In [12], an anomaly detection approach is developed which is capable of detecting faults of known and unknown type without requiring the setting of expert parameters.

This paper aims to propose an intelligent and highly secure method to equip the electric vehicles with a powerful anomaly detection and avoidance mechanism. The proposed method is constructed based on support vector machine and the concept of one-class detection system to avoid any malicious behavior in the vehicle [13]. Here the experimental CAN bus data are used to let the support vector machine learn the normal frequency of the different message frames at different commands. In order to get into the maximum capability of the model, a new optimization algorithm based on social spider optimization (SSO) algorithm is proposed to adjust the SVR setting parameters, properly [14]. Due to the high complexity and nonlinearity of the electric vehicle CAN bus dataset, a new two-stage modification method based on crossover and mutation operators of genetic algorithm is developed which can increase the algorithm population diversity and at the same time avoid premature

convergence. The feasibility and satisfying performance of the proposed model are examined using the real datasets gathered from an electric vehicle.

2. EXISTING SYSTEM

In electric vehicles, CAN standard is the most widely used protocol by automakers for communications with low cost in the units with a high number of components, up to 500 million chips. In the vehicle industry, the CAN resiliency and noise resistance level is acceptable owing to its structure. Unfortunately, CAN bus protocols do not offer confidentiality and authentication to CAN data frames so making it possible for hackers to enter the vehicle system, either on a wired or wireless approach. In the wired approach, one can communicate with the CAN bus through the OBD-II maintenance port located under the steering in most vehicles. Although the main idea behind this port is to be used for diagnostics of engine and vehicle maintenance, but it will let hackers take the CAN packets using a simple scanning tool. From this point, it is easy to read and write traffic in the CAN bus with the use of ECOM API such as CAN Receive Message and CAN Transmit Message [10]. In the wireless attack, the cyber interfere is the same by targeting ECU except that the penetration point is not OBD-II. While the penetration points in the wireless hacking can be different, but in most of them it is required for the car to be connected to a malicious WIFI hotspot. Also, the security mechanism of the transponder can be reverse engineered in the keyless vehicles. The research reveals several weaknesses in the design of the cipher, the authentication protocol and also in their implementation. Some of the other wireless entry points to vehicles can be named as “Wireless connection between sensors and ECUs such as TPMS system”, “Add-on technologies, entertainment system (gaming), smart key” and “Internet, smart infrastructures”.

Disadvantages of existing system

- 1) Less accuracy
- 2) low Efficiency

3. PROPOSED SYSTEM

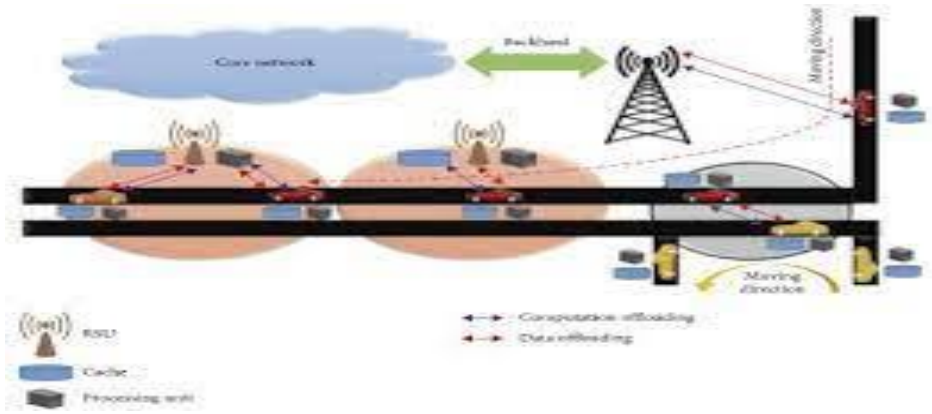
The last sections were mainly focusing on the proposed model, the theories and backgrounds. In this section, the performance of the proposed model is examined using the experimental data gathered from an electric car. This paper assesses the DoS attack since it is focusing on the vehicle intra-communication within which DoS has a high significance among different

attacks. In the DoS attack, the hacker attempts to prevent legitimate users (driver) from accessing the service. Considering the fact that vehicles are mobile devices, DoS attack is so dangerous (and thus important) in vehicles since it can make severe car crash or losses. Examples of hackings achieved through the DoS attack in the vehicles are activating the brakes while the vehicle is in motion, turning the steering wheel to the left/right suddenly, turning off the engine, unlocking a door, etc. According to the analysis from the recorded CAN traffic during a normal driving time of 10-minute, each message frame with a specific ID has some unique frequencies which can be learned by the proposed anomaly detection model. In Table II, a set of CAN bus identifiers and frequencies is shown. In order to make sure that our model is learning all possible ID numbers, we had a complete trace analysis. Therefore, after capturing the traffic log and implementing the trace analysis, it was realized that a 10-min driving scenario would capture the majority of the messages that are occurring commonly, owing to the fact that most of the CAN messages are periodic. Therefore, the model developed can be regarded as the proof-of-concept that shows the proposed anomaly detection model can learn the existing pattern in the CAN messages to distinguish between normal and anomalous behaviors in the testing phase. In order to make a realistic condition for the driving test, the CAN traffic file covers the following conditions: the engine ignition was turned on and the vehicle remained at a standstill for a few seconds and then the gear was engaged to “D” mode. Then, the vehicle is driven for about 8 minutes at a public street. For several times, the brake pedal is also pressed during the drive. The car is then stopped and the gear mode is changed to “R” to drive backwards a bit and make a parking maneuver. Finally, the gear moves to “P” mode so the vehicle would remain at the standstill for a few seconds and then the engine is turned off.

Advantages Of Proposed System

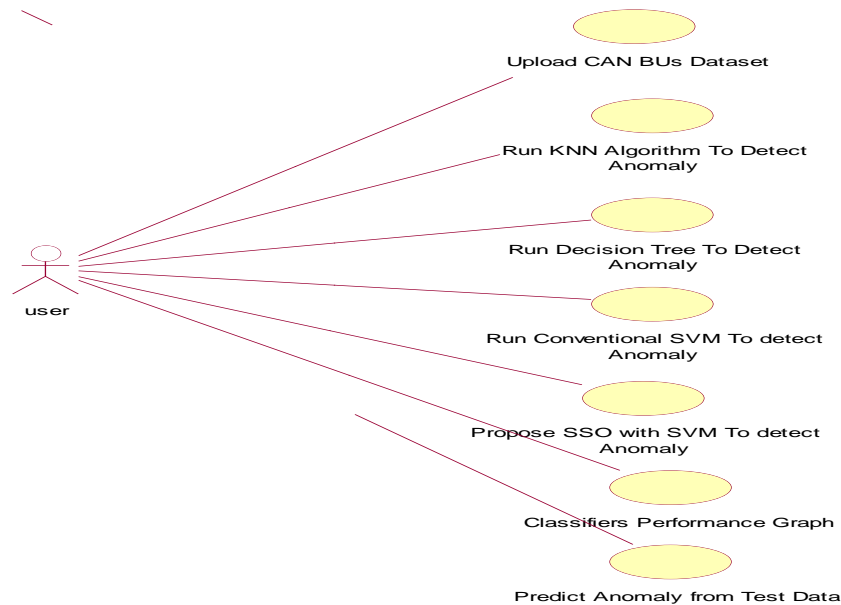
- 1) High accuracy
- 2) High efficiency

4. SYSTEM ARCHITECTURE



5. USE CASE DIAGRAM

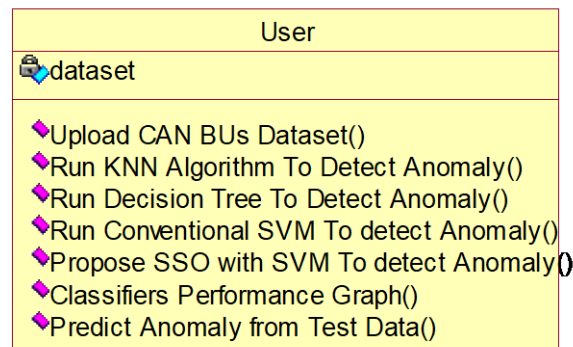
A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



6. CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's

classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



7. CONCLUSION

This paper proposed a novel intelligent and secured anomaly detection model for cyber attack detection and avoidance in the electric vehicles. The proposed model is constructed based on an improved support vector machine model reinforced by the MSSO algorithm. From the cyber security point of view, the proposed model could successfully detect malicious behaviors while letting the trusted message frames broadcast in the CAN protocol. The high HR% and FR% indices prove the true positive and true negative decisions made by the proposed model. Regarding the MR% and CR% indices, the very low values which most of them are around the upper and lower bounds of the message frame frequency, show the highly trustable performance of this model. The authors will assess the effect of other cyberattacks on the performance of different anomaly detection models in the future works.

8. REFERENCES

- [1] A. Monot ; N. Navet ; B. Bavoux ; F. Simonot-Lion, “Multisource Software on Multicore Automotive ECUs—Combining Runnable Sequencing With Task Scheduling”, IEEE Trans. Industrial Electronics, vol. 59, no. 10. Pp. 3934-3942, 2012.
- [2] T.Y. Moon; S.H. Seo; J.H. Kim; S.H. Hwang; J. Wook Jeon, “Gateway system with diagnostic function for LIN, CAN and FlexRay”, 2007 International Conference on Control, Automation and Systems, pp. 2844 – 2849, 2007.
- [3] B. Groza; S. Murvay, “Efficient Protocols for Secure Broadcast in Controller Area Networks”, IEEE Trans. Industrial Informatics, vol. 9, no. 4, pp. 2034-2042, 2013.

- [4] B. Mohandes, R. Al Hammadi, W. Sanusi, T. Mezher, S. El Khatib, “Advancing cyber–physical sustainability through integrated analysis of smart power systems: A case study on electric vehicles”, *International Journal of Critical Infrastructure Protection*, vol. 23, pp. 33–48, 2018.
- [5] G. Loukas, E. Karapistoli, E. Panaousis, P. Sarigiannidis, T. Vuong, A taxonomy and survey of cyber-physical intrusion detection approaches for vehicles, *Ad Hoc Networks*, vol. 84, pp. 124-147, 2019.
- [6] Hoppe T, Kiltz S, Dittmann J. Security threats to automotive can networks. practical examples and selected short-term countermeasures. *Reliab Eng Syst Saf* vol. 96, no. 1, pp. 11–25, 2011.
- [7] Schulze S, Pukall M, Saake G, Hoppe T, Dittmann J. On the need of data management in automotive systems. In: *BTW*, vol. 144; pp. 217–26, 2009.
- [8] Ling C, Feng D. An algorithm for detection of malicious messages on can buses. 2012 national conference on information technology and computer science. Atlantis Press; 2012.
- [9] Oguma H, Yoshioka X, Nishikawa M, Shigetomi R, Otsuka A, Imai H. New attestation based security architecture for in-vehicle communication. In: *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE. IEEE*; pp. 1–6, 2008.
- [10] L. Pan, X. Zheng, H. X. Chen, T. Luan, L. Batten, “Cyber security attacks to modern vehicular systems”, *Journal of Information Security and Applications*, vol. 36, pp. 90-100, October 2017.
- [11] Kang, M. J., & Kang, J. W., “Intrusion detection system using deep neural network for in-vehicle network security”, *PloS one*, vol. 11, no. 6, e0155781, 2016.
- [12] Theissler, A., “Detecting known and unknown faults in automotive systems using ensemble-based anomaly detection”, *Knowledge-Based Systems*, vol. 123, pp. 163-173.

DRUG RECOMMENDATION SYSTEM BASED ON SENTIMENT ANALYSIS OF DRUG REVIEWS USING MACHINE LEARNING

Lakamsani Haritha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G.Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Since coronavirus has shown up, inaccessibility of legitimate clinical resources is at its peak, like the shortage of specialists and healthcare workers, lack of proper equipment and medicines etc. The entire medical fraternity is in distress, which results in numerous individual's demise. Due to unavailability, individuals started taking medication independently without appropriate consultation, making the health condition worse than usual. As of late, machine learning has been valuable in numerous applications, and there is an increase in innovative work for automation. This paper intends to present a drug recommender system that can drastically reduce specialists heap. In this research, we build a medicine recommendation system that uses patient reviews to predict the sentiment using various vectorization processes like Bow, TF-IDF, Word2Vec, and Manual Feature Analysis, which can help recommend the top drug for a given disease by different classification algorithms. The predicted sentiments were evaluated by precision, recall, f1score, accuracy, and AUC score. The results show that classifier LinearSVC using TF-IDF vectorization outperforms all other models with 93% accuracy.

1. INTRODUCTION

With the number of corona virus cases growing exponentially, the nations are facing a shortage of doctors, particularly in rural areas where the quantity of specialists is less compared to urban areas. A doctor takes roughly 6 to 12 years to procure the necessary qualifications. Thus, the number of doctors can't be expanded quickly in a short time frame. A Telemedicine framework ought to be energized as far as possible in this difficult time [1].

Clinical blunders are very regular nowadays. Over 200 thousand individuals in China and 100 thousand in the USA are affected every year because of prescription mistakes. Over 40% medicine, specialists make mistakes while prescribing since specialists compose the solution as referenced by their knowledge, which is very restricted [2][3]. Choosing the toplevel medication is significant for patients who need specialists that know wide based information about microscopic organisms, antibacterial medications, and patients [6]. Every day a new study comes up with accompanying more drugs, tests, accessible for clinical staff every day. Accordingly, it turns out to be progressively challenging for doctors to choose which treatment or medications to give to a patient based on indications, past clinical history.

With the exponential development of the web and the web-based business industry, item reviews have become an imperative and integral factor for acquiring items worldwide. Individuals worldwide become adjusted to analyze reviews and websites first before settling on a choice to buy a thing. While most of past exploration zeroed in on rating



expectation and proposals on the E-Commerce field, the territory of medical care or clinical therapies has been infrequently taken care of. There has been an expansion in the number of individuals worried about their well-being and finding a diagnosis online. As demonstrated in a Pew American Research center survey directed in 2013 [5], roughly 60% of grown-ups searched online for health related subjects, and around 35% of users looked for diagnosing health conditions on the web. A medication recommender framework is truly vital with the goal that it can assist specialists and help patients to build their knowledge of drugs on specific health conditions.

A recommender framework is a customary system that proposes an item to the user, dependent on their advantage and necessity. These frameworks employ the customers' surveys to break down their sentiment and suggest a recommendation for their exact need. In the drug recommender system, medicine is offered on a specific condition dependent on patient reviews using sentiment analysis and feature engineering. Sentiment analysis is a progression of strategies, methods, and tools for distinguishing and extracting emotional data, such as opinion and attitudes, from language [7]. On the other hand, Featuring engineering is the process of making more features from the existing ones; it improves the performance of models.

2. EXISTING SYSTEM

The study [9] presents GalenOWL, a semantic-empowered online framework, to help specialists discover details on the medications. The paper depicts a framework that suggests drugs for a patient based on the patient's infection, sensitivities, and drug interactions. For empowering GalenOWL, clinical data and terminology first converted to ontological terms utilizing worldwide standards, such as ICD-10 and UNII, and then correctly combined with the clinical information. Leilei Sun [10] examined large scale treatment records to locate the best treatment prescription for patients. The idea was to use an efficient semantic clustering algorithm estimating the similarities between treatment records. Likewise, the author created a framework to assess the adequacy of the suggested treatment. This structure can prescribe the best treatment regimens to new patients as per their demographic locations and medical complications. An Electronic Medical Record (EMR) of patients gathered from numerous clinics for testing.

The result shows that this framework improves the cure rate. In this research [11], multilingual sentiment analysis was performed using Naive Bayes and Recurrent Neural Network (RNN). Google translator API was used to convert multilingual tweets into the English language. The results exhibit that RNN with 95.34% outperformed Naive Bayes, 77.21%.

The study [12] is based on the fact that the recommended drug should depend upon the patient's capacity. For example, if the patient's immunity is low, at that point, reliable medicines ought to be recommended. Proposed a risk level classification method to identify the patient's immunity. For example, in excess of 60 risk factors, hypertension, liquor addiction, and so forth have been adopted, which decide the patient's capacity to shield himself from infection. A web-based prototype system was also created, which uses a



decision support system that helps doctors select first-line drugs. Xiaohong Jiang et al. [13] examined three distinct algorithms,

decision tree algorithm, support vector machine (SVM), and backpropagation neural network on treatment data. SVM was picked for the medication proposal module as it performed truly well in each of the three unique boundaries - model exactness, model proficiency, model versatility. Additionally, proposed the mistake check system to ensure analysis, precision and administration quality. Mohammad Mehedi

Hassan et al. [14] developed a cloudassisted drug proposal (CADRE). As per patients' side effects, CADRE can suggest drugs with top-N related prescriptions.

This proposed framework was initially founded on collaborative filtering techniques in which the medications are initially bunched into clusters as indicated by the functional description data. However, after considering its weaknesses like computationally costly, cold start, and information sparsity, the model is shifted to a cloud-helped approach using tensor decomposition for advancing the quality of experience of medication suggestion.

Disadvantages

In the existing work, the system did not implement an exact sentiment analysis for large data sets.

This system is less performance due to lack Data Classification and Data Fragmentation technique.

3. PROPOSED SYSTEM

A recommender framework is a customary system that proposes an item to the user, dependent on their advantage and necessity. These frameworks employ the customers' surveys to break down their sentiment and suggest a recommendation for their exact need. In the drug recommender system, medicine is offered on a specific condition dependent on patient reviews using sentiment analysis and feature engineering. Sentiment analysis is a progression of strategies, methods, and tools for

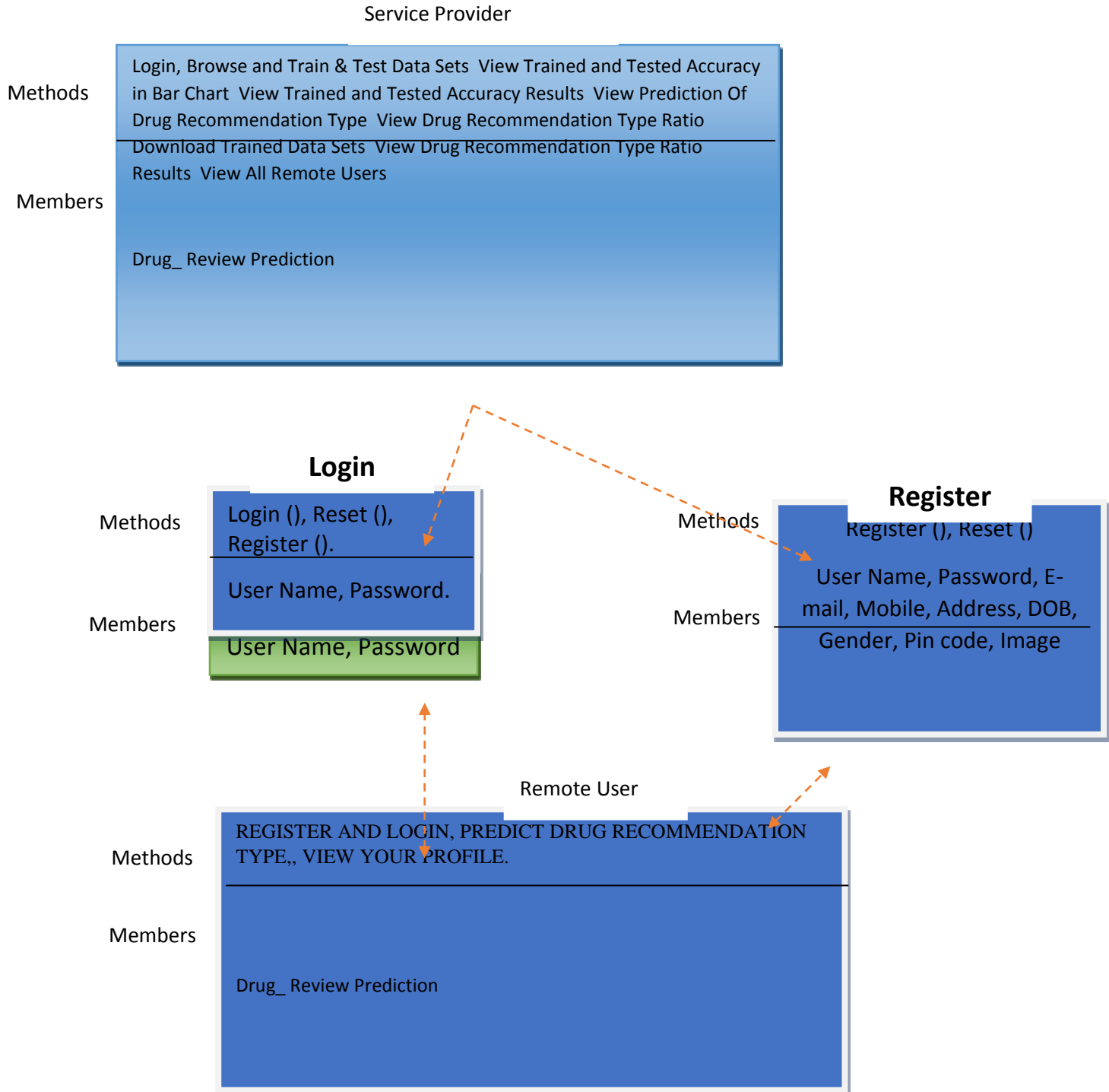
distinguishing and extracting emotional data, such as opinion and attitudes, from language [7]. On the other hand, Featuring engineering is the process of making more features from the existing ones; it improves the performance of models.

Advantages

The system is more effective since it presents the proposed algorithm used in natural language processing responsible for counting the number of times of all the tokens in review or document..

The system has exact sentiment analysis prediction techniques for Data Cleaning and Visualization.

Class Diagram :



4. CONCLUSIONS

Reviews are becoming an integral part of our daily lives; whether go for shopping, purchase something online or go to some restaurant, we first check the reviews to make the right decisions. Motivated by this, in this research sentiment analysis of drug reviews was studied



to build a recommender system using different types of machine learning classifiers, such as Logistic Regression, Perceptron, Multinomial Naive Bayes, Ridge classifier, Stochastic gradient descent, Linear SVC, applied on Bow, TF-IDF, and classifiers such as Decision Tree, Random Forest, Lgbm, and Cat boost were applied on Word2Vec and Manual features method. We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score, which reveal that the Linear SVC on TF-IDF outperforms all other models with 93% accuracy. On the other hand, the Decision tree classifier on Word2Vec showed the worst performance by achieving only 78% accuracy. We added best-predicted emotion values from each method, Perceptron on Bow (91%), Linear SVC on TF-IDF (93%), LGBM on Word2Vec (91%), Random Forest on manual features (88%), and multiply them by the normalized useful Count to get the overall score of the drug by condition to build a recommender system. Future work involves comparison of different oversampling techniques, using different values of n-grams, and optimization of algorithms to improve the performance of the recommender system.

5. REFERENCES

- [1] Telemedicine, <https://www.mohfw.gov.in/pdf/Telemedicine.pdf>
- [2] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. *Mayo Clin Proc.* 2014 Aug;89(8):1116-25.
- [3] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. *Practical Journal of Clinical Medicine*, 4.
- [4] Drug Review Dataset, <https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529#>
- [5] Fox, Susannah, and Maeve Duggan. "Health online 2013. 2013." URL: <http://pewinternet.org/Reports/2013/Health-online.aspx>
- [6] Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practice guidelines for the management of community-acquired pneumonia in adults. *Infectious Diseases Society of America. Clin Infect Dis.* 2000 Aug;31(2):347-82. doi: 10.1086/313954. Epub 2000 Sep 7. PMID: 10987697; PMCID: PMC7109923.
- [7] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.
- [8] T. N. Tekade and M. Emmanuel, "Probabilistic aspect mining approach for interpretation and evaluation of drug reviews," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), Paralakhemundi, 2016, pp. 1471-1476, doi: 10.1109/SCOPEs.2016.7955684.
- [9] Doulaverakis, C., Nikolaidis, G., Kleontas, A. et al. GalenOWL: Ontology-based drug recommendations discovery. *J Biomed Semant* 3, 14 (2012). <https://doi.org/10.1186/2041-1480-3-14>
- [10] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining



(KDD '16). Association for Computing Machinery, New York, NY, USA, 1865–1874.
DOI:<https://doi.org/10.1145/2939672.2939866>

[11] V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.

[12] Shimada K, Takada H, Mitsuyama S, et al. Drug-recommendation system for patients with infectious diseases. AMIA Annu Symp Proc. 2005;2005:1112.

DRUG RECOMMENDATION SYSTEM BASED ON SENTIMENT ANALYSIS OF DRUG REVIEWS USING MACHINE LEARNING

Lakamsani Haritha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G.Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Since coronavirus has shown up, inaccessibility of legitimate clinical resources is at its peak, like the shortage of specialists and healthcare workers, lack of proper equipment and medicines etc. The entire medical fraternity is in distress, which results in numerous individual's demise. Due to unavailability, individuals started taking medication independently without appropriate consultation, making the health condition worse than usual. As of late, machine learning has been valuable in numerous applications, and there is an increase in innovative work for automation. This paper intends to present a drug recommender system that can drastically reduce specialists heap. In this research, we build a medicine recommendation system that uses patient reviews to predict the sentiment using various vectorization processes like Bow, TF-IDF, Word2Vec, and Manual Feature Analysis, which can help recommend the top drug for a given disease by different classification algorithms. The predicted sentiments were evaluated by precision, recall, f1score, accuracy, and AUC score. The results show that classifier LinearSVC using TF-IDF vectorization outperforms all other models with 93% accuracy.

1. INTRODUCTION

With the number of corona virus cases growing exponentially, the nations are facing a shortage of doctors, particularly in rural areas where the quantity of specialists is less compared to urban areas. A doctor takes roughly 6 to 12 years to procure the necessary qualifications. Thus, the number of doctors can't be expanded quickly in a short time frame. A Telemedicine framework ought to be energized as far as possible in this difficult time [1].

Clinical blunders are very regular nowadays. Over 200 thousand individuals in China and 100 thousand in the USA are affected every year because of prescription mistakes. Over 40% medicine, specialists make mistakes while prescribing since specialists compose the solution as referenced by their knowledge, which is very restricted [2][3]. Choosing the toplevel medication is significant for patients who need specialists that know wide based information about microscopic organisms, antibacterial medications, and patients [6]. Every day a new study comes up with accompanying more drugs, tests, accessible for clinical staff every day. Accordingly, it turns out to be progressively challenging for doctors to choose which treatment or medications to give to a patient based on indications, past clinical history.

With the exponential development of the web and the web-based business industry, item reviews have become an imperative and integral factor for acquiring items worldwide. Individuals worldwide become adjusted to analyze reviews and websites first before settling on a choice to buy a thing. While most of past exploration zeroed in on rating



expectation and proposals on the E-Commerce field, the territory of medical care or clinical therapies has been infrequently taken care of. There has been an expansion in the number of individuals worried about their well-being and finding a diagnosis online. As demonstrated in a Pew American Research center survey directed in 2013 [5], roughly 60% of grown-ups searched online for health related subjects, and around 35% of users looked for diagnosing health conditions on the web. A medication recommender framework is truly vital with the goal that it can assist specialists and help patients to build their knowledge of drugs on specific health conditions.

A recommender framework is a customary system that proposes an item to the user, dependent on their advantage and necessity. These frameworks employ the customers' surveys to break down their sentiment and suggest a recommendation for their exact need. In the drug recommender system, medicine is offered on a specific condition dependent on patient reviews using sentiment analysis and feature engineering. Sentiment analysis is a progression of strategies, methods, and tools for distinguishing and extracting emotional data, such as opinion and attitudes, from language [7]. On the other hand, Featuring engineering is the process of making more features from the existing ones; it improves the performance of models.

2. EXISTING SYSTEM

The study [9] presents GalenOWL, a semantic-empowered online framework, to help specialists discover details on the medications. The paper depicts a framework that suggests drugs for a patient based on the patient's infection, sensitivities, and drug interactions. For empowering GalenOWL, clinical data and terminology first converted to ontological terms utilizing worldwide standards, such as ICD-10 and UNII, and then correctly combined with the clinical information. Leilei Sun [10] examined large scale treatment records to locate the best treatment prescription for patients. The idea was to use an efficient semantic clustering algorithm estimating the similarities between treatment records. Likewise, the author created a framework to assess the adequacy of the suggested treatment. This structure can prescribe the best treatment regimens to new patients as per their demographic locations and medical complications. An Electronic Medical Record (EMR) of patients gathered from numerous clinics for testing.

The result shows that this framework improves the cure rate. In this research [11], multilingual sentiment analysis was performed using Naive Bayes and Recurrent Neural Network (RNN). Google translator API was used to convert multilingual tweets into the English language. The results exhibit that RNN with 95.34% outperformed Naive Bayes, 77.21%.

The study [12] is based on the fact that the recommended drug should depend upon the patient's capacity. For example, if the patient's immunity is low, at that point, reliable medicines ought to be recommended. Proposed a risk level classification method to identify the patient's immunity. For example, in excess of 60 risk factors, hypertension, liquor addiction, and so forth have been adopted, which decide the patient's capacity to shield himself from infection. A web-based prototype system was also created, which uses a



decision support system that helps doctors select first-line drugs. Xiaohong Jiang et al. [13] examined three distinct algorithms,

decision tree algorithm, support vector machine (SVM), and backpropagation neural network on treatment data. SVM was picked for the medication proposal module as it performed truly well in each of the three unique boundaries - model exactness, model proficiency, model versatility. Additionally, proposed the mistake check system to ensure analysis, precision and administration quality. Mohammad Mehedi

Hassan et al. [14] developed a cloudassisted drug proposal (CADRE). As per patients' side effects, CADRE can suggest drugs with top-N related prescriptions.

This proposed framework was initially founded on collaborative filtering techniques in which the medications are initially bunched into clusters as indicated by the functional description data. However, after considering its weaknesses like computationally costly, cold start, and information sparsity, the model is shifted to a cloud-helped approach using tensor decomposition for advancing the quality of experience of medication suggestion.

Disadvantages

In the existing work, the system did not implement an exact sentiment analysis for large data sets.

This system is less performance due to lack Data Classification and Data Fragmentation technique.

3. PROPOSED SYSTEM

A recommender framework is a customary system that proposes an item to the user, dependent on their advantage and necessity. These frameworks employ the customers' surveys to break down their sentiment and suggest a recommendation for their exact need. In the drug recommender system, medicine is offered on a specific condition dependent on patient reviews using sentiment analysis and feature engineering. Sentiment analysis is a progression of strategies, methods, and tools for

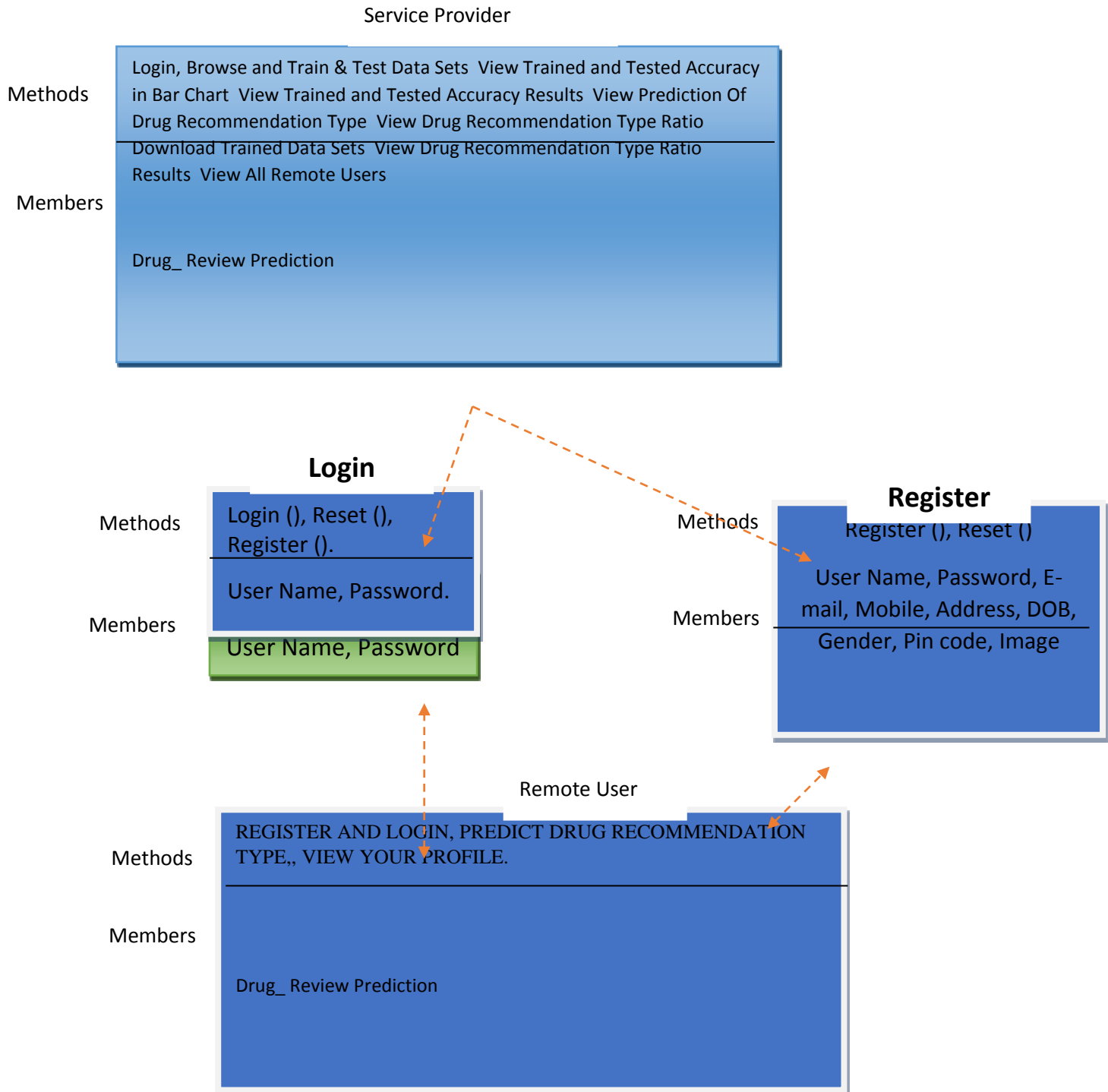
distinguishing and extracting emotional data, such as opinion and attitudes, from language [7]. On the other hand, Featuring engineering is the process of making more features from the existing ones; it improves the performance of models.

Advantages

The system is more effective since it presents the proposed algorithm used in natural language processing responsible for counting the number of times of all the tokens in review or document..

The system has exact sentiment analysis prediction techniques for Data Cleaning and Visualization.

Class Diagram :



4. CONCLUSIONS

Reviews are becoming an integral part of our daily lives; whether go for shopping, purchase something online or go to some restaurant, we first check the reviews to make the right decisions. Motivated by this, in this research sentiment analysis of drug reviews was studied



to build a recommender system using different types of machine learning classifiers, such as Logistic Regression, Perceptron, Multinomial Naive Bayes, Ridge classifier, Stochastic gradient descent, Linear SVC, applied on Bow, TF-IDF, and classifiers such as Decision Tree, Random Forest, Lgbm, and Cat boost were applied on Word2Vec and Manual features method. We evaluated them using five different metrics, precision, recall, f1score, accuracy, and AUC score, which reveal that the Linear SVC on TF-IDF outperforms all other models with 93% accuracy. On the other hand, the Decision tree classifier on Word2Vec showed the worst performance by achieving only 78% accuracy. We added best-predicted emotion values from each method, Perceptron on Bow (91%), Linear SVC on TF-IDF (93%), LGBM on Word2Vec (91%), Random Forest on manual features (88%), and multiply them by the normalized useful Count to get the overall score of the drug by condition to build a recommender system. Future work involves comparison of different oversampling techniques, using different values of n-grams, and optimization of algorithms to improve the performance of the recommender system.

5. REFERENCES

- [1] Telemedicine, <https://www.mohfw.gov.in/pdf/Telemedicine.pdf>
- [2] Wittich CM, Burkle CM, Lanier WL. Medication errors: an overview for clinicians. *Mayo Clin Proc.* 2014 Aug;89(8):1116-25.
- [3] CHEN, M. R., & WANG, H. F. (2013). The reason and prevention of hospital medication errors. *Practical Journal of Clinical Medicine*, 4.
- [4] Drug Review Dataset, <https://archive.ics.uci.edu/ml/datasets/Drug%2BReview%2BDataset%2B%2528Drugs.com%2529#>
- [5] Fox, Susannah, and Maeve Duggan. "Health online 2013. 2013." URL: <http://pewinternet.org/Reports/2013/Health-online.aspx>
- [6] Bartlett JG, Dowell SF, Mandell LA, File TM Jr, Musher DM, Fine MJ. Practice guidelines for the management of community-acquired pneumonia in adults. *Infectious Diseases Society of America. Clin Infect Dis.* 2000 Aug;31(2):347-82. doi: 10.1086/313954. Epub 2000 Sep 7. PMID: 10987697; PMCID: PMC7109923.
- [7] Fox, Susannah & Duggan, Maeve. (2012). Health Online 2013. Pew Research Internet Project Report.
- [8] T. N. Tekade and M. Emmanuel, "Probabilistic aspect mining approach for interpretation and evaluation of drug reviews," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), Paralakhemundi, 2016, pp. 1471-1476, doi: 10.1109/SCOPEs.2016.7955684.
- [9] Doulaverakis, C., Nikolaidis, G., Kleontas, A. et al. GalenOWL: Ontology-based drug recommendations discovery. *J Biomed Semant* 3, 14 (2012). <https://doi.org/10.1186/2041-1480-3-14>
- [10] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. 2016. Data-driven Automatic Treatment Regimen Development and Recommendation. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining



(KDD '16). Association for Computing Machinery, New York, NY, USA, 1865–1874.
DOI:<https://doi.org/10.1145/2939672.2939866>

[11] V. Goel, A. K. Gupta and N. Kumar, "Sentiment Analysis of Multilingual Twitter Data using Natural Language Processing," 2018 8th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2018, pp. 208-212, doi: 10.1109/CSNT.2018.8820254.

[12] Shimada K, Takada H, Mitsuyama S, et al. Drug-recommendation system for patients with infectious diseases. AMIA Annu Symp Proc. 2005;2005:1112.

AN EFFICIENT SPAM DETECTION FOR IOT DEVICES USING MACHINE LEARNING

Madhyannapu Pavani (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G.Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The Internet of Things (IoT) is a group of millions of devices having sensors and actuators linked over wired or wireless channel for data transmission. IoT has grown rapidly over the past decade with more than 25 billion devices are expected to be connected by 2020. The volume of data released from these devices will increase many-fold in the years to come. In addition to an increased volume, the IoT devices produces a large amount of data with a number of different modalities having varying data quality defined by its speed in terms of time and position dependency. In such an environment, machine learning algorithms can play an important role in ensuring security and authorization based on biotechnology, anomalous detection to improve the usability and security of IoT systems. On the other hand, attackers often view learning algorithms to exploit the vulnerabilities in smart IoT-based systems. Motivated from these, in this paper, we propose the security of the IoT devices by detecting spam using machine learning. To achieve this objective, Spam Detection in IoT using Machine Learning framework is proposed. In this framework, five machine learning models are evaluated using various metrics with a large collection of inputs features sets. Each model computes a spam score by considering the refined input features. This score depicts the trustworthiness of IoT device under various parameters. REFIT Smart Home dataset is used for the validation of proposed technique. The results obtained proves the effectiveness of the proposed scheme in comparison to the other existing schemes.

1. INTRODUCTION

Internet of Things (IoT) enables convergence and implementations between the real-world objects irrespective of their geographical locations. Implementation of such network management and control make privacy and protection strategies utmost important and

challenging in such an environment. IoT applications need to protect data privacy to fix security issues such as intrusions, spoofing attacks, DoS attacks, DoS attacks, jamming, eavesdropping, spam, and malware. The safety measures of IoT devices depends upon the size and type of organization in which it is imposed.

The behavior of users forces the security gateways to cooperate. In other words, we can say that the location, nature, application of IoT devices decides the security measures [1]. For instance, the smart IoT security cameras in the smart organization can capture the different parameters for analysis and intelligent decision making [2]. The maximum care to be taken is with web based devices as maximum number of IoT devices are web dependent. It is common at the workplace that the IoT devices installed in an organization can be used to implement security and privacy features efficiently.

For example, wearable devices collect and send user's health data to a connected smartphone should prevent leakage of information to ensure privacy. It has been found in the market that 25-30% of working employees connect their personal IoT devices with the organizational network. The expanding nature of IoT attracts both the audience, i.e., the users and the attackers. However, with the emergence of ML in various attacks scenarios, IoT devices choose a defensive strategy and decide the key parameters in the security protocols for trade-off between security, privacy and computation. This job is challenging as it is usually difficult for an IoT system with limited resources to estimate the current network and timely attack status.

2. EXISTING SYSTEM

_ Denial of service (DDoS) attacks: The attackers can flood the target database with unwanted requests to stop IoT devices from having access to various services. These malicious requests produced by a network of IoT devices are commonly known as bots [3]. DDoS can exhaust all the resources provided by the service provider. It can block authentic users and can make the network resource unavailable.

_ RFID attacks: These are the attacks imposed at the physical layer of IoT device. This attack leads to loose the integrity of the device. Attackers attempt to modify the data either at the node storage or while it is in the transmission within network. The common attacks possible at the sensor node are attacks on availability, attacks on authenticity, attacks on

confidentiality, Cryptography keys brute-forcing [4]. The countermeasures to ensure prevention of such attacks includes password protection, data encryption and restricted access control.

_ Internet attacks: The IoT device can stay connected with Internet to access various resources. The spammers who want to steal other systems information or want their target website to be visited continuously, use spamming techniques [5]. The common technique used for the same is Ad fraud. It generates the artificial clicks at a targeted website for monetary profit. Such practicing team is known as cyber criminals.

NFC attacks: These attacks are mainly concerned with electronic payment frauds. The possible attacks are unencrypted traffic, Eavesdropping, and Tag modification. The solution for this problem is the conditional privacy protection. So, the attacker fails to create the same profile with the help of user's public key [6]. This model is based on random public keys by trusted service manager.

Disadvantages

In the existing work, the system is less effective due to lack of Spam Detection in IoT using Machine Learning framework.

This system is less performance in which it is clear that Supervised machine learning techniques is absence.

3. PROPOSED SYSTEM

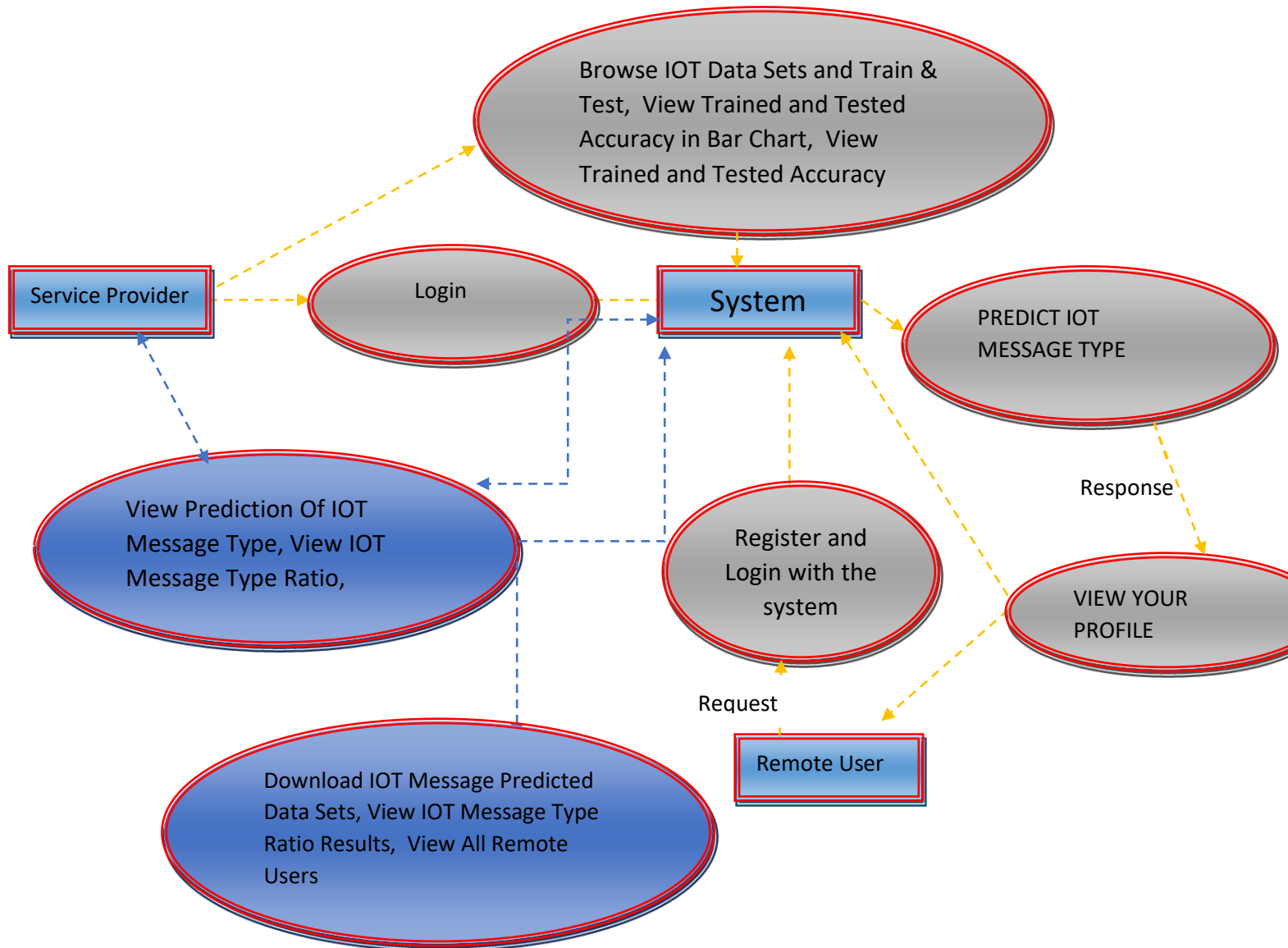
The digital world is completely dependent upon the smart devices. The information retrieved from these devices should be spam free. The information retrieval from various IoT devices is a big challenge because it is collected from various domains. As there are multiple devices involved in IoT, so a large volume of data is generated having heterogeneity and variety. We can call this data as IoT data. IoT data has various features such as real-time, multi-source, rich and sparse..

Advantages

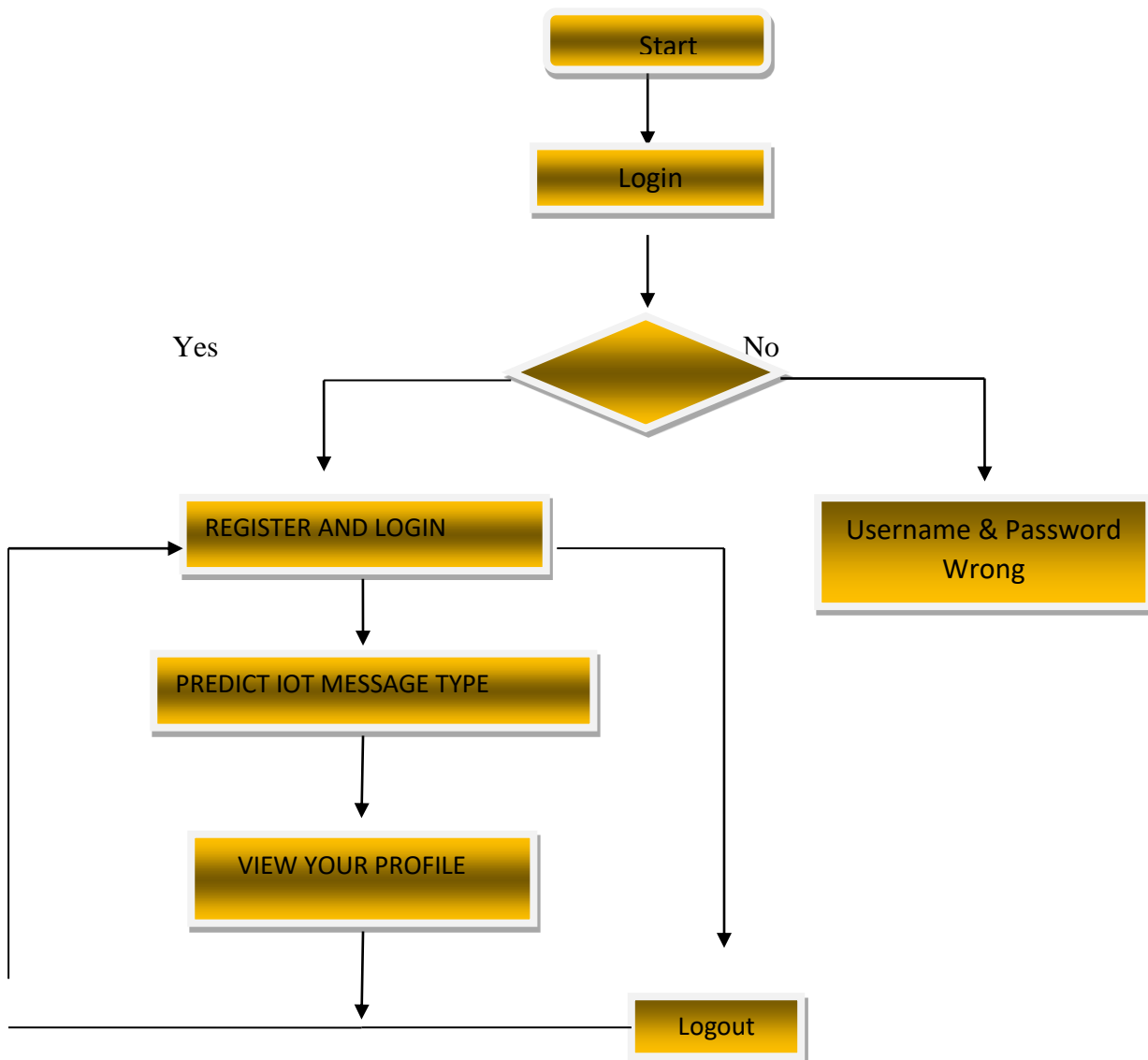
1) The proposed scheme of spam detection is validated using five different machine learning models.

- 2) An algorithm is proposed to compute the spamicity score of each model which is then used for detection and intelligent decision making.
- 3) Based upon the spamicity score computed in previous step, the reliability of IoT devices is analyzed using different evaluation metrics.

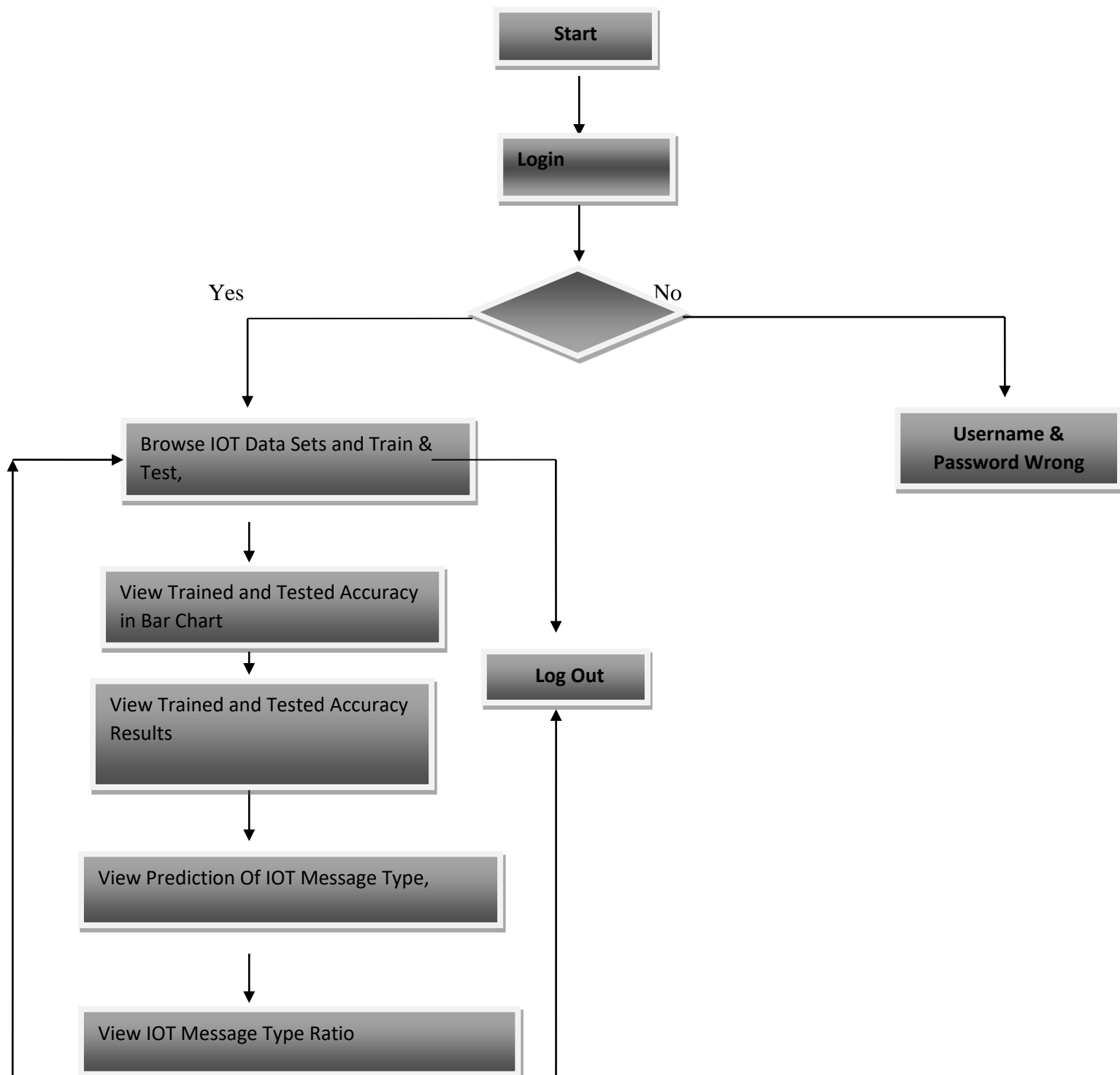
Data Flow Diagram :

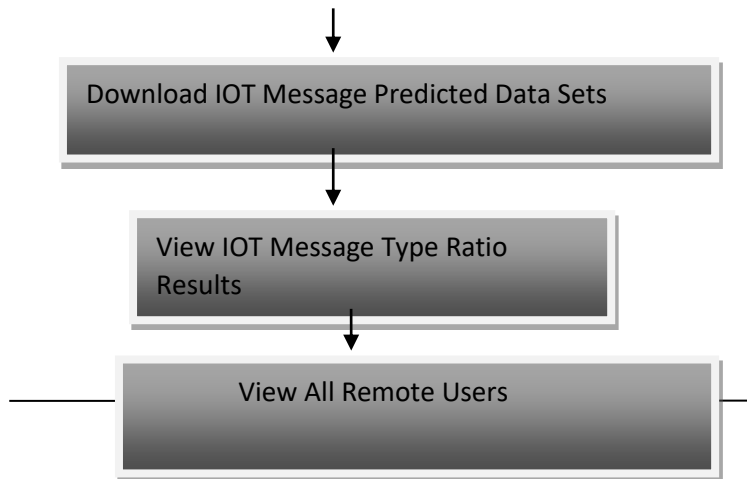


Flow Chart : Remote User



Flow Chart : Service Provider





Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

The following are the types of Integration Testing:

Top Down Integration

This method is an incremental approach to the construction of program structure. Modules are integrated by moving downward through the control hierarchy, beginning with the main program module. The module subordinates to the main program module are incorporated into the structure in either a depth first or breadth first manner.

In this method, the software is tested from main module and individual stubs are replaced when the test proceeds downwards.

4. CONCLUSIONS

The proposed framework, detects the spam parameters of IoT devices using machine learning models. The IoT dataset used for experiments, is pre processed by using feature engineering procedure. By experimenting the framework with machine learning models, each IoT appliance is awarded with a spam score. This refines the conditions to be taken for successful working of IoT devices in a smart home. In future, we are planning to consider the climatic and surrounding features of IoT device to make them more secure and trustworthy.

5. REFERENCES

- [1] Z.-K. Zhang, M. C. Y. Cho, C.-W. Wang, C.-W. Hsu, C.-K. Chen, and S. Shieh, "Iot security: ongoing challenges and research opportunities," in 2014 IEEE 7th international conference on service-oriented computing and applications. IEEE, 2014, pp. 230–234.
- [2] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot security and privacy: The case study of a smart home," in 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops). IEEE, 2017, pp. 618–623.

- [3] E. Bertino and N. Islam, "Botnets and internet of things security," *Computer*, no. 2, pp. 76–79, 2017.
- [4] C. Zhang and R. Green, "Communication security in internet of thing: preventive measure and avoid ddos attack over iot network," in *Proceedings of the 18th Symposium on Communications & Networking*. Society for Computer Simulation International, 2015, pp. 8–15.
- [5] W. Kim, O.-R. Jeong, C. Kim, and J. So, "The dark side of the internet: Attacks, costs and responses," *Information systems*, vol. 36, no. 3, pp. 675–705, 2011.
- [6] H. Eun, H. Lee, and H. Oh, "Conditional privacy preserving security protocol for nfc applications," *IEEE Transactions on Consumer Electronics*, vol. 59, no. 1, pp. 153–160, 2013.
- [7] R. V. Kulkarni and G. K. Venayagamoorthy, "Neural network based secure media access control protocol for wireless sensor networks," in *2009 International Joint Conference on Neural Networks*. IEEE, 2009, pp. 1680–1687.
- [8] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications,"

MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING, DEEP LEARNING AND STREAM-LIT

Mallula Venkatesh*¹

*¹B.V. Raju College, MCA Department, Adikavi Nannaya University, Bhimavaram, Andhra
Pradesh, India.

DOI : <https://www.doi.org/10.56726/IRJMETS42818>

ABSTRACT

Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit is a comprehensive project aimed at predicting various diseases including diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. This project leverages machine learning algorithms such as TensorFlow with Keras, Support Vector Machine (SVM), and Logistic Regression. The models are deployed using Streamlit Cloud and the Streamlit library, providing a user-friendly interface for disease prediction. The application interface comprises five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. Upon selecting a particular disease, the user is prompted to input the relevant parameters required for the prediction model. Once the parameters are entered, the application promptly generates the disease prediction result, indicating whether the individual is affected by the disease or not. This project addresses the need for accurate disease prediction using machine learning techniques, allowing for early detection and intervention. The userfriendly interface provided by Streamlit Cloud and the Streamlit library enhances accessibility and usability, enabling individuals to easily assess their risk for various diseases. The high accuracies achieved by the different models demonstrate the effectiveness of the employed machine learning algorithms in disease prediction.

Keywords: Machine Learning, Streamlit, TensorFlow, Keras, SVM, Logistic Regression, Diabetes, Heart Disease, Kidney Disease, Parkinson's Disease, Breast Cancer.

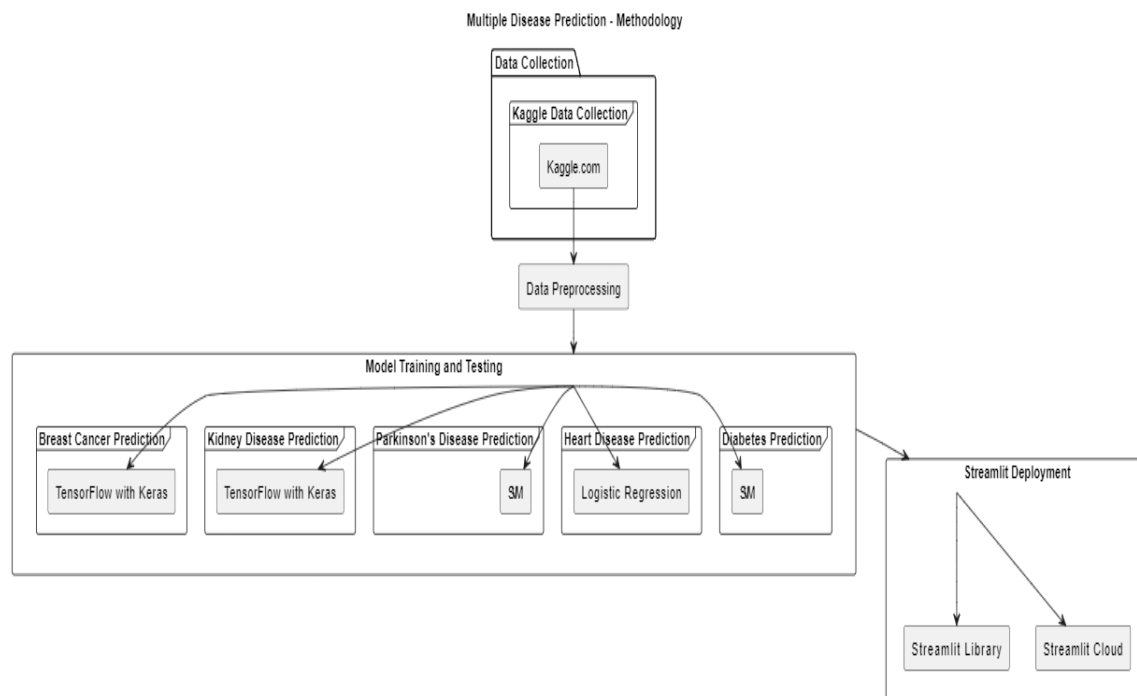
I. INTRODUCTION

The project "Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit" focuses on predicting five different diseases: diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The prediction models are built using machine learning algorithms, including Support Vector Machine (SVM) for diabetes and Parkinson's disease, Logistic Regression for heart disease, and TensorFlow with Keras for kidney disease and breast cancer. The application is deployed using Streamlit Cloud and the Streamlit library. The project begins by collecting relevant data from Kaggle.com, which is then preprocessed to prepare it for training and testing the prediction models. Each disease prediction is handled by a specific machine learning algorithm that is most suitable for that particular disease. SVM is employed for diabetes and Parkinson's disease, Logistic Regression for heart disease, and TensorFlow with Keras for kidney disease and breast cancer. The application interface offers five options, each corresponding to a specific disease. When a user selects a particular disease, the application prompts for the necessary parameters required by the corresponding model to predict the disease result. The user provides the required parameters, and the application displays the prediction result based on the input. To deploy the prediction models, Streamlit Cloud and the Streamlit library are utilized. Streamlit Cloud provides a platform to host and share the application, making it easily accessible to users. The Streamlit library simplifies the process of developing interactive and user-friendly web applications. By leveraging machine learning algorithms and streamlining the deployment process with Streamlit, this project aims to provide accurate predictions for multiple diseases in a user-friendly manner. The application's intuitive interface allows users to input disease-specific parameters and obtain prediction results, facilitating early detection and proactive healthcare management.

II. METHODOLOGY

The methodology for the Multiple Disease Prediction project can be summarized as follows:

1. **Data Collection:** Data is collected from Kaggle.com, a popular platform for accessing datasets. The data is obtained specifically for diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer.
2. **Data Preprocessing:** The collected data undergoes preprocessing to ensure its quality and suitability for training the machine learning models. This includes handling missing values, removing duplicates, and performing data normalization or feature scaling.
3. **Model Selection:** Different machine learning algorithms are chosen for each disease prediction task. Support Vector Machine (SVM), Logistic Regression, and TensorFlow with Keras are selected as the algorithms for various diseases based on their performance and suitability for the specific prediction tasks.
4. **Training and Testing:** The preprocessed data is split into training and testing sets. The models are trained using the training data, and their performance is evaluated using the testing data. Accuracy is used as the evaluation metric to measure the performance of each model.
5. **Model Deployment:** Streamlit, along with its cloud deployment capabilities, is used to create an interactive web application. The application offers a user-friendly interface with five options for disease prediction: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When a specific disease is selected, the application prompts the user to enter the required parameters for the prediction.



III. PROBLEM STATEMENT

Develop a machine learning-based application using TensorFlow with Keras, Support Vector Machine (SVM), and Logistic Regression to predict multiple diseases including diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The application should allow users to input relevant parameters for a specific disease and provide an accurate prediction of whether an individual is affected by the disease based on the trained models. The project aims to improve healthcare outcomes by enabling early detection and prediction of diseases using machine learning algorithms and streamlining the prediction process through an intuitive user interface.

IV. EXISTING SYSTEM

Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit The existing system is a project that focuses on predicting diabetes, heart disease, and Parkinson's disease using various machine learning algorithms. The algorithms employed in this project include Naive Bayes classifier, Decision Trees classifier, Random Forest classifier, Support Vector Machine (SVM), and Logistic Regression. To deploy the models, Streamlit Cloud and Streamlit library are utilized, providing a user-friendly interface for disease prediction. The

system collects data from various sources, preprocesses it, trains the models with the processed data, and tests their performance. One of the algorithms used in the system is SVM, which achieved a prediction accuracy of 76% for diabetes. This means that the SVM model correctly predicted diabetes in 76% of the cases it was tested on. The performance of the SVM algorithm indicates its effectiveness in distinguishing between diabetic and non-diabetic individuals. Similarly, for Parkinson's disease prediction, the SVM algorithm achieved a prediction accuracy of 71%. This means that the SVM model accurately predicted the presence or absence of Parkinson's disease in 71% of the cases. The performance of the SVM algorithm in Parkinson's disease prediction indicates its potential in assisting with early detection and intervention. The system incorporates other machine learning algorithms such as Naive Bayes, Decision Trees, and Random Forest, which may have varying performance metrics for different diseases. These algorithms are designed to leverage different characteristics of the data and make predictions based on distinct methodologies. Overall, the existing system demonstrates the effectiveness of machine learning algorithms in predicting diabetes, heart disease, and Parkinson's disease. The use of Streamlit Cloud and Streamlit library allows for easy deployment and provides a user-friendly interface for interacting with the prediction models. Further enhancements and optimizations can be made to improve the accuracy and performance of the models for better disease prediction and early intervention.

V. PROPOSED SYSTEM

The existing system the models are not implemented with TensorFlow and keras, but in the proposed system two new diseases are added to the existing system those are implemented by neural networks with the help of TensorFlow and keras. We use new techniques like data standardization to standardize the data, Label encoding technique to transform text data to numerical data and dimensionality reduction to reduce the features with loss of information from the data. We use the algorithms that are perfectly suitable for the dataset and we take simple models to increase the model performance.

The existing system uses flask api, in the proposed system we use stream lit library, stream lit cloud and GITHub. The proposed system is a comprehensive disease prediction project that utilizes machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, TensorFlow with Keras, to predict multiple diseases such as diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. The system aims to provide accurate disease predictions based on input parameters and a user-friendly interface developed using Streamlit and deployed on Streamlit Cloud. Data for the models is collected from the Kaggle platform, a popular data science community, and is preprocessed to ensure its quality and suitability for training the models. The preprocessed data is then used to train the respective machine learning algorithms specific to each disease. The trained models are tested to evaluate their accuracy in disease prediction.

The system employs the SVM algorithm to predict diabetes, achieving an accuracy of 78%. This indicates that the SVM model can accurately identify the presence or absence of diabetes in patients, aiding in early detection and effective management. For Parkinson's disease prediction, the system uses the SVM algorithm with an accuracy of 87%. This high accuracy demonstrates the capability of the SVM model to distinguish individuals with Parkinson's disease from healthy individuals.

Heart disease prediction is performed using the Logistic Regression algorithm, which achieves an accuracy of 85%. This model effectively identifies the likelihood of heart disease in patients, supporting timely intervention and appropriate treatment. For kidney disease prediction, the system utilizes TensorFlow with Keras, achieving an impressive accuracy of 97%. This high accuracy demonstrates the power of deep learning models in accurately predicting kidney diseases, enabling early detection and proactive care. Breast cancer prediction is also included in the system, utilizing TensorFlow with Keras and achieving an accuracy of 95%. The deep learning model developed using these technologies can effectively detect the presence of breast cancer, enabling early diagnosis and intervention.

The proposed system provides a user-friendly interface with a menu consisting of five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When a particular disease is selected, the system prompts the user to enter the required parameters for prediction. After providing the parameters, the system generates and displays the prediction result, facilitating informed decision-making and proactive healthcare management.

VI. INPUT AND OUTPUT DESIGN

Input Design: The Multiple Disease Prediction system requires user input in the form of parameters specific to each disease. When the user selects a particular disease from the options menu, the system prompts for the relevant parameters. The input design should ensure that the user can easily provide the required information. The application provides a user interface with a menu containing five disease options: heart disease, kidney disease, diabetes, Parkinson's disease, and breast cancer. When the user clicks on a specific disease, the application prompts for the required parameters for that particular disease prediction. The input design should ensure that the parameters requested are relevant and necessary for accurate disease prediction. The user should be able to enter the parameters in a user-friendly and intuitive manner.

Output Design:

The Multiple Disease Prediction system provides the predicted result of whether the person is affected by the selected disease or not. The output design should present the result in a clear and understandable format. The system should display the output after the user has entered the parameters. The output could be presented as:

- "Prediction: The person is affected by [Disease Name]." (If the prediction is positive)
- "Prediction: The person is not affected by [Disease Name]." (If the prediction is negative)

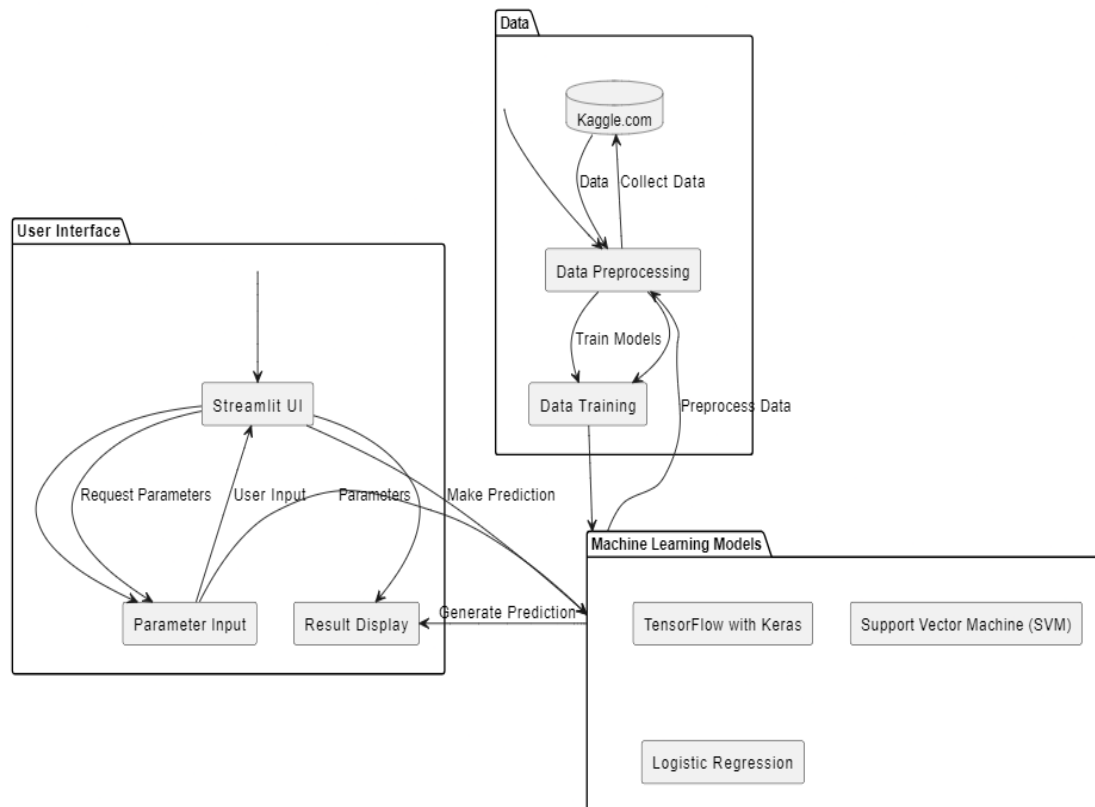
The output should be displayed on the user interface, allowing the user to easily interpret the prediction result. Overall, the input design ensures that the user can enter the necessary parameters for disease prediction, while the output design presents the prediction result clearly on the user interface.

VII. SYSTEM DESIGN

SYSTEM ARCHITECTURE:

The architecture diagram for the multiple disease prediction web application:

Multiple Disease Prediction - Architecture Diagram



VIII. RESULTS

The results for all the ML models and of final completed project are shown in the following figures and tables:

Table 1. Comparison of Accuracy of all 5 models

SN.	Disease Name	Algorithm Name	Existing system accuracy	Proposed system accuracy
1	Diabetes	SVM Classifier	76%	78%
2	Heart disease	Logistic Regression	80%	85%
3	Parkinson's disease	SVM Classifier	71%	87%
4	Kidney disease	TensorFlow and keras	-	97%
5	Breast cancer	TensorFlow andkeras	-	96%

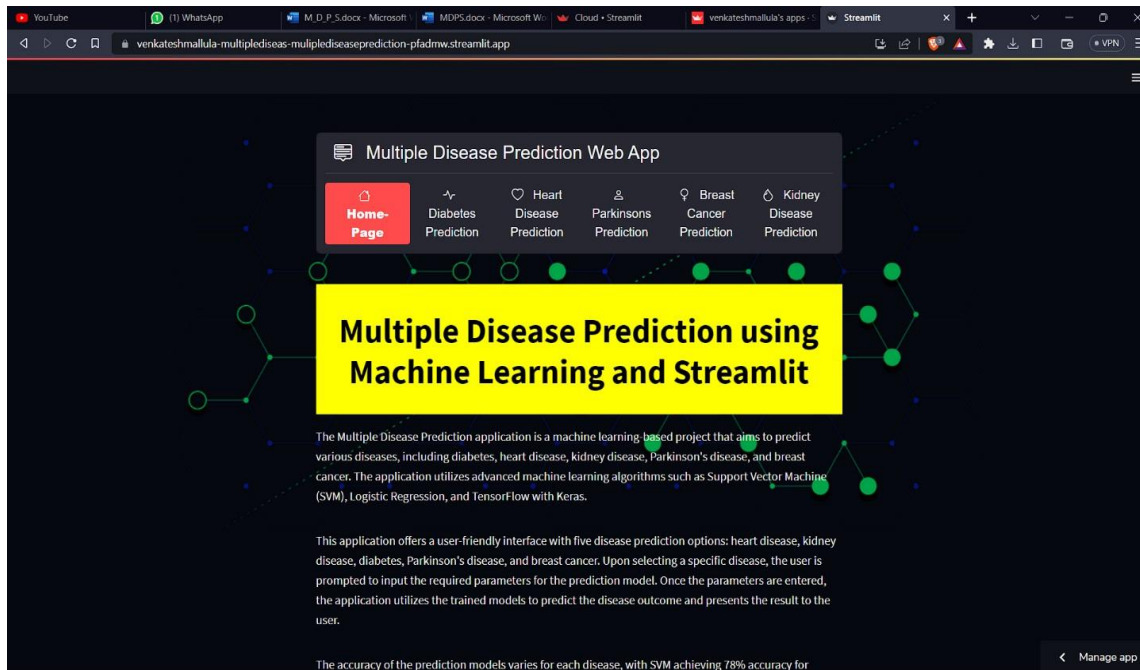
The existing system doesn't have kidney disease and breast cancer prediction system. that's why we leave "-" in the existing system accuracy for kidney disease and breast cancer. prediction system. that's why we leave "-" in the existing system accuracy for kidney disease and breast cancer.

After completion of project the application interfaces are look like following pictures:

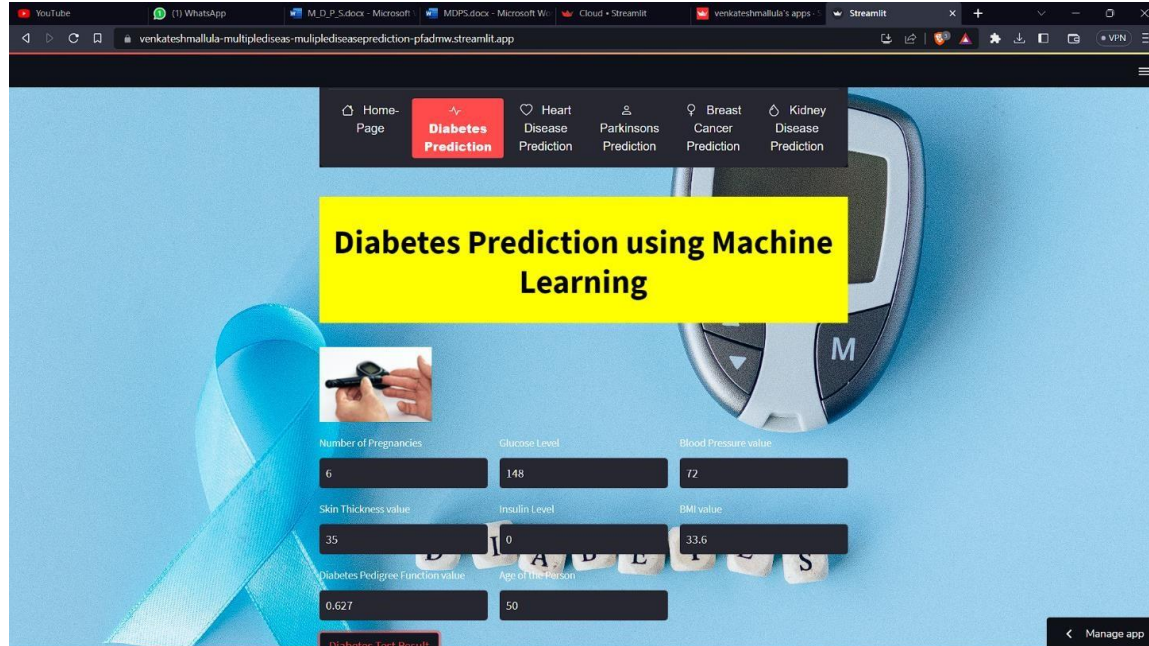
For diabetes and heart disease the features are less so we easily enter feature values in the respective feature input.

For Parkinson's disease, breast cancer and kidney disease the features are more so the application takes feature values in a single input field, the values must be separated by comma (",").

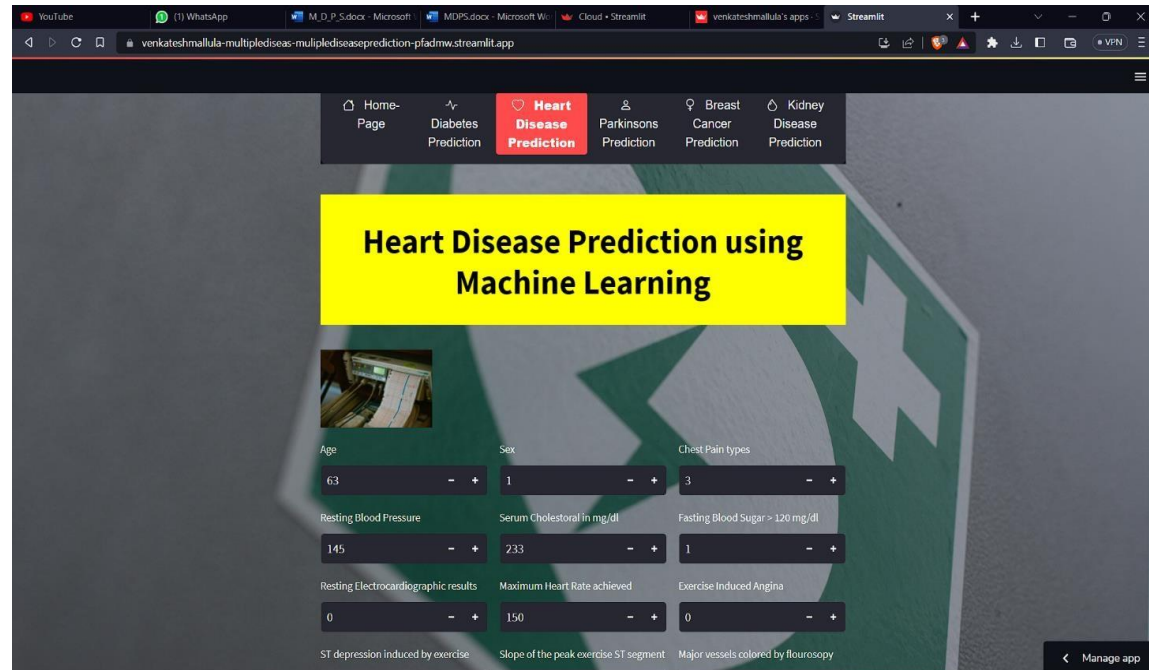
1. HOME PAGE:



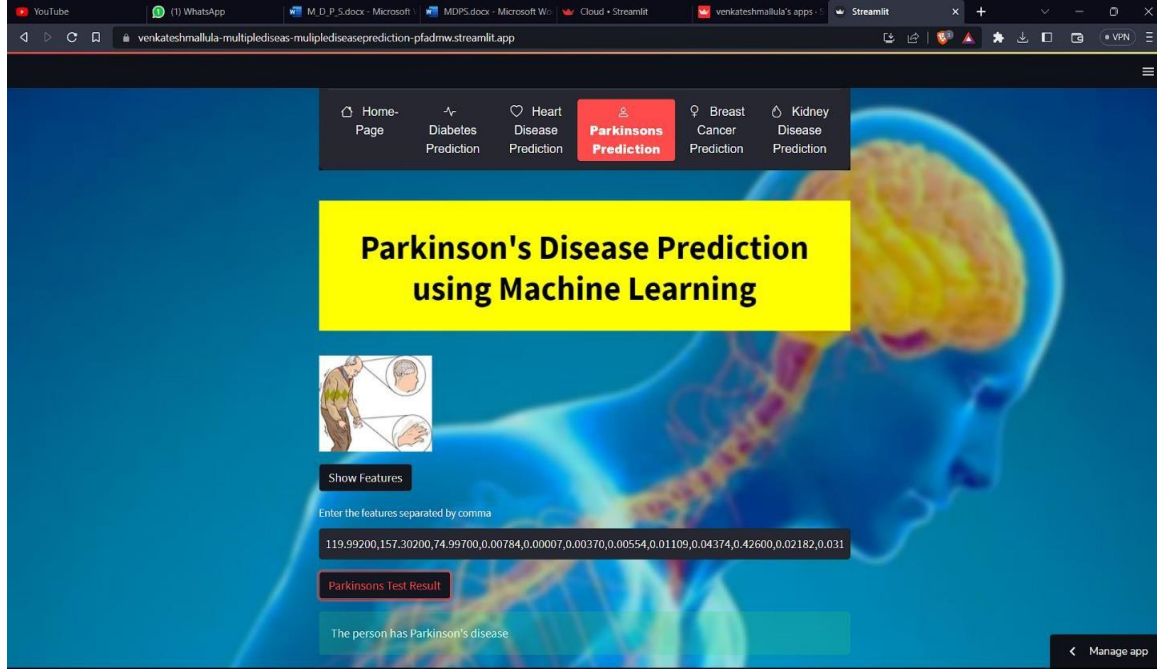
2. DIABETES PREDICTION:



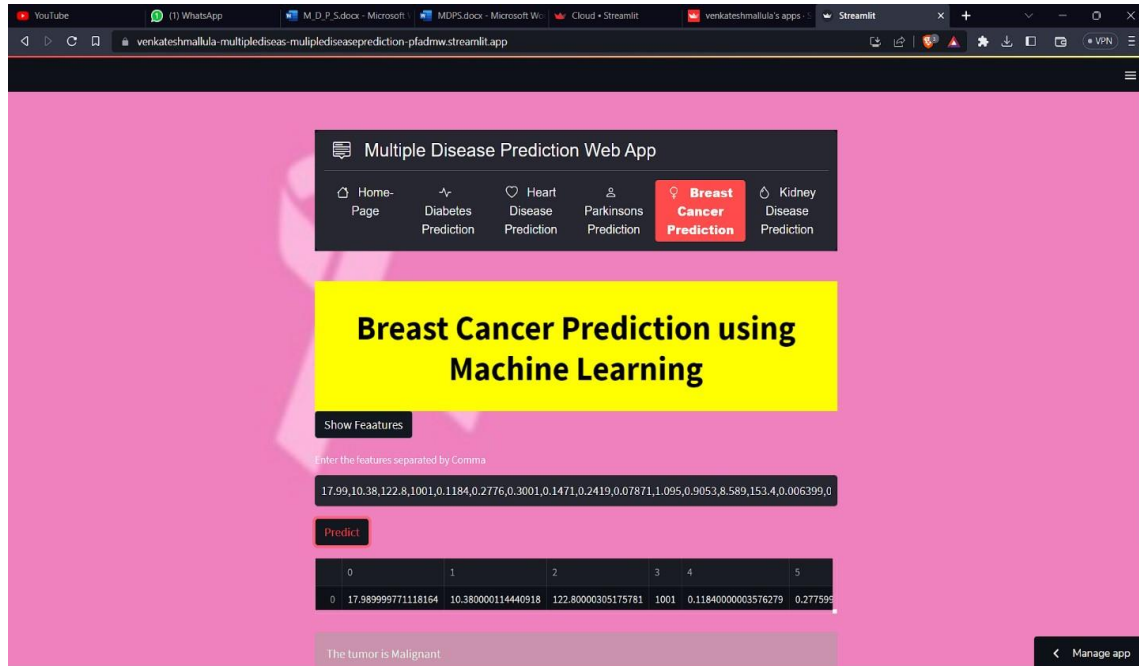
3. HEART DISEASE PREDICTION:



4. PARKINSON'S DISEASE PREDICTION:



5. BREAST CANCER PREDICTION:



6. KIDNEY DISEASE PREDICTION:



IX. CONCLUSION

In conclusion, our project utilized machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression, and TensorFlow with Keras, to develop a disease prediction system. The system focused on five diseases: diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. We collected data from Kaggle.com and performed preprocessing to ensure data quality. For diabetes prediction, we achieved an accuracy of 78% using the SVM algorithm. Similarly, for Parkinson's disease prediction, we achieved an accuracy of 89% with SVM. Logistic Regression was employed for heart disease prediction, resulting in an accuracy of 85%. For kidney disease and breast cancer prediction, we utilized TensorFlow with Keras, achieving accuracy rates of 97% and 95% respectively. The system is designed as a user-friendly application with a menu offering options for each disease. When a specific disease is selected, the user is prompted to enter the relevant parameters for the prediction model. Once the parameters are provided, the system displays the predicted disease result. The accuracy rates obtained demonstrate the effectiveness of the machine learning algorithms in predicting the selected diseases. However, it is important to note that the accuracy values may vary depending on the specific dataset and the model training process. Overall, this project demonstrates the potential of machine learning and streamlit library in developing disease prediction models. The application can be a valuable tool in assisting healthcare professionals and individuals in early detection and prevention of these diseases. Further enhancements and refinements can be made to improve the accuracy and usability of the system, making it an even more valuable resource in the field of disease prediction and prevention.

X. FUTURE SCOPE

The project "Multiple Disease Prediction using Machine Learning, Deep Learning and Streamlit" has shown promising results in predicting various diseases with respectable accuracies. Moving forward, there are several potential areas for future development and enhancement:

- **Expansion of Disease Prediction:** The current project focuses on diabetes, heart disease, kidney disease, Parkinson's disease, and breast cancer. In the future, additional diseases can be included to create a more comprehensive and diverse disease prediction system.
- **Integration of More Machine Learning Algorithms:** While the project already employs Support Vector Machines (SVM), Logistic Regression, and TensorFlow with Keras, there are many other machine learning

algorithms that can be explored. Incorporating algorithms such as Random Forest, Gradient Boosting, or Neural Networks may further improve the accuracy and performance of the disease prediction models.

- Integration of Advanced Feature Engineering Techniques: Feature engineering plays a crucial role in extracting meaningful information from the input data. Exploring advanced feature engineering techniques like dimensionality reduction, feature selection, and feature extraction can potentially enhance the prediction models and their interpretability.
- Real-time Monitoring and Feedback: Enhancing the application to provide real-time monitoring and feedback to users can be beneficial. Incorporating features like reminders for regular health check-ups, personalized recommendations for disease prevention, or alerts for abnormal health parameters can empower users to take proactive measures for their well-being.
- Integration of Explainable AI: Making the disease prediction models more interpretable and transparent is an important aspect for user trust and understanding. Exploring techniques for explainable AI, such as feature.

XI. ACKNOWLEDGEMENT

My sincere thanks to our B V RAJU COLLEGE, BHIMAVARAM for giving us a platform to prepare a project on the topic "Multiple Disease Prediction using machine Learning, deep learning and stream-lit". We are sincerely grateful for Dr. V. Bhaskara Murthy Sir(HOD) and Mr. G Ramesh Kumar sir as our guide for providing help during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestion

XII. REFERENCES

- [1] TensorFlow: Martín Abadi, Ashish Agarwal, et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. arXiv preprint arXiv:1603.04467.
- [2] Keras: François Chollet et al. (2015). Keras. GitHub repository.
- [3] Support Vector Machine (SVM): Corinna Cortes and Vladimir Vapnik (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- [4] Logistic Regression: Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). John Wiley & Sons.
- [5] Streamlit: Streamlit Documentation. <https://docs.streamlit.io/>
- [6] Kaggle: Kaggle website. <https://www.kaggle.com/>
- [7] Data sources: You can provide the specific datasets you used from Kaggle.com, mentioning the authors or contributors of the datasets.
- [8] Zhang, Y., & Ghorbani, A. (2019). A review on machine learning algorithms for diagnosis of heart disease. *IEEE Access*, 7, 112751-112760.
- [9] Arora, P., Chaudhary, S., & Rana, M. (2020). Prediction of diabetes using machine learning algorithms: A review. *Journal of Ambient Intelligence and Humanized Computing*, 11(6), 2575-2589.
- [10] Kaur, H., Batra, N., & Rani, R. (2020). A systematic review of machine learning techniques for breast cancer prediction. *Journal of Medical Systems*, 44(11), 1-15.
- [11] Gupta, D., & Rathore, S. (2021). A comprehensive review on machine learning algorithms for kidney disease diagnosis. *Journal of Medical Systems*, 45(1), 1-17.
- [12] Saeed, A., & Al-Jumaily, A. (2020). Machine learning techniques for Parkinson's disease diagnosis using handwriting: A review. *Computers in Biology and Medicine*, 122, 103804.



BULLYNET: UNMASKING CYBERBULLIES ON SOCIAL NETWORKS

Manan Narmada Bai (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

One of the most harmful consequences of social media is the rise of cyberbullying, which tends to be more sinister than traditional bullying given that online records typically live on the internet for quite a long time and are hard to control. In this paper, we present a three-phase algorithm, called BullyNet, for detecting cyberbullies on Twitter social network. We exploit bullying tendencies by proposing a robust method for constructing a cyberbullying signed network. We analyze tweets to determine their relation to cyberbullying, while considering the context in which the tweets exist in order to optimize their bullying score. We also propose a centrality measure to detect cyberbullies from a cyberbullying signed network, and we show that it outperforms other existing measures. We experiment on a dataset of 5.6 million tweets and our results shows that the proposed approach can detect cyberbullies with high accuracy, while being scalable with respect to the number of tweets.

1. INTRODUCTION

The Internet has created never before seen opportunities for human interaction and socialization. In the past decade, social media, in particular, has had a popularity explosion. From MySpace to Face book, Twitter, Flickr, and Instagram, people are connecting and interacting in a way that was previously impossible. The widespread usage of social media across people from all ages created a vast amount of data for several research topics, including recommender systems [1], link predictions [2], visualization, and analysis of social networks [3].

While the growth of social media has created an excellent platform for communications and information sharing, it has also created a new platform for malicious activities such as spamming [4], trolling [5], and cyber bullying [6]. According to the Cyber bullying Research Center (CRC) [7], cyber bullying occurs

when someone uses the technology to send messages to harass, mistreat or threaten a person or a group. Unlike traditional bullying where aggression is a short and temporary face to- face occurrence, cyber bullying contains hurtful messages which are present online for a long time. These messages can be accessed worldwide, and are often irrevocable. Laws about cyber bullying and how it is handled differ from one place to another. For example, in the United States, the majority of the states incorporate cyberbullying into their bullying laws, and cyber bullying is considered a criminal offense in most of them [8]. Popular social media platforms such as Face book and Twitter are very vulnerable to cyber bullying due to the popularity of these social media sites and the anonymity that the internet offers to the perpetrators. Although strict laws exist to punish cyber bullying, there are very



less tools available to effectively combat cyber bullying.

2. EXISTING SYSTEM

The first method of determining bullying messages was done using a combination of text-based analytics and a mix of text and user features. Zhao et al. [18] proposed a text based Embeddings-Enhanced Bag-of-Words (EBoW) model that utilizes a concatenation of bullying features, bag-of-words, and latent semantic features to obtain a final representation, which is then passed through a classifier to identify cyberbullies.

Xu et al. [21] used textual information to identify emotions in bullying traces, as opposed to determining whether or not a message was bullying. Singh et al. [19] proposed a probabilistic socio-textual information fusion for cyberbullying detection. This fusion uses social network features derived from a 1.5 ego network and textual features, such as density of bad words and part-of-speech-tags. Hosseinmardi et al. [20] used images and text to detect cyberbullying incidents. The text and image features were gathered from media sessions containing images and the corresponding comments, which was then fed into various classifiers. Chen [25] proposed a novel method in identifying cyberbullies within a multi-modal context. To understand cyberbullying Kao et al. [26] proposed a framework by studying social role detection. By using words and comments, temporal characteristics, and social information of a session as well as peer influence Cheng et al. [27], [28] proposed frameworks for detecting cyberbullies.

The second method was aimed at identifying the person behind the

cyberbullying incidents. Squicciarini et al. [22] used MySpace data to create a graph, which integrated user, textual, and network features. This graph was used to detect cyberbullies and predict the spreading of bullying behavior through node classification. Gal'an-Garc'ia et al. [23] used supervised machine learning to detect the real users behind troll profiles on Twitter, and demonstrated the technique in a

real case of cyberbullying. In a recent paper on aggression and bullying in Twitter, Chatzakou et al. [24] found cyberbullies and aggressors using user, text, and network-based features.

Disadvantages

The system is less effective due to lack of Constructing bullying signed network.

The system doesn't effective due to lack of training large scale datasets.

3. PROPOSED SYSTEM

In the proposed system, the system studies the problem of cyberbullying in social media in an attempt to answer the following research question: Can tweet contexts (conversations) help improve the detection of cyberbullying in Twitter?. Our intuition is that each tweet should be evaluated not only based on its contents, but also based on the context in which it exists. The system calls such a context a conversation, which is a set of tweets between two or more people exchanging information about a certain subject. Thus, our solution consists of three parts. First, for each conversation, a conversation graph is generated based on the sentiment and bullying words in the tweets. Second, we compute the bullying score for each pair of users in a conversation graph, and then combine all graphs to create an SSN

called bullying signed network (B). The inclusion of negative links can bring out information that would otherwise be missed with only positive links [16]. Finally, we propose a centrality measure called attitude & merit (A&M) to detect bullying users from the signed network B. Our main contributions are organized as follows:

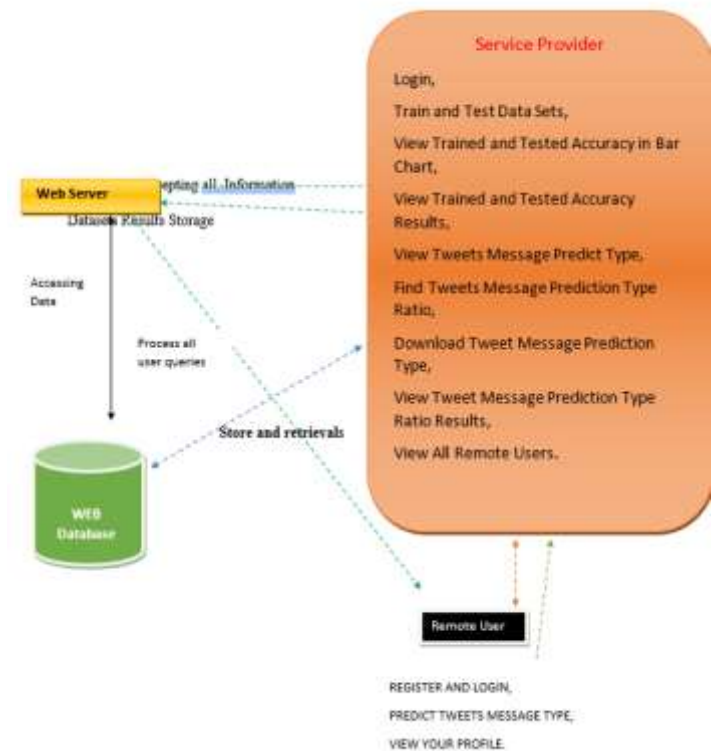
- 1) Collected, preprocessed and labelled the Twitter dataset.
- 2) Proposed a novel efficient algorithm for detecting cyberbullies on Twitter.
 - a) Built conversation.
 - b) Constructed Bullying Signed Network.
 - c) Proposed Attitude and Merit Centrality.
- 3) Experimented on 5.6 million tweets collected over 6 months. The results show that our approach can detect cyberbullies with high accuracy, while being scalable with respect to the number of tweets.

Advantages

The system is more effective due to presence of Conversation Graph Generation Algorithm, Bullying Signed Network Generation Algorithm, and Bully Finding Algorithm.

The system is more effective due to the techniques to analyze large number of datasets.

4. ARCHITECTURE DIAGRAM



5. INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

6. CONCLUSION

Although the digital revolution and the rise of social media enabled great advances in communication platforms and social interactions, a wider proliferation of harmful behavior known as bullying has also emerged. This paper presents a novel



framework of Bully Net to identify bully users from the Twitter social network. We performed extensive research on mining signed networks for better understanding of the relationships between users in social media, to build a signed network (SN) based on bullying tendencies. We observed that by constructing conversations based on the context as well as content, we could effectively identify the emotions and the behavior behind bullying. In our experimental study, the evaluation of our proposed centrality measures to detect bullies from signed network, we achieved around 80% accuracy with 81% precision in identifying bullies for various cases.

There are still several open questions deserving further investigation. First, our approach focuses on extracting emotions and behavior from texts and emojis in tweets. However, it would be interesting to investigate images and videos, given that many users use them to bully others. Second, it does not distinguish between bully and aggressive users. Devising new algorithms or techniques to distinguish bullies from aggressors would prove critical in better identification of cyber bullies. Another topic of interest would be to study the relationship between conversation graph dynamics and geographic location and how these dynamics are affected by the geographic dispersion of the users? Are the proximity increase the bullying behaviour?

7. REFERENCES

- [1] J. Tang, C. Aggarwal, and H. Liu, "Recommendations in signed social networks," in Proceedings of the International Conference on WWW, 2016, pp. 31–40.
- [2] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," Proceedings of the ASIS&T, vol. 58, no. 7, pp. 1019–1031, 2007.
- [3] U. Brandes and D. Wagner, "Analysis and visualization of social networks," in Graph drawing software, 2004, pp. 321–340.
- [4] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," In Proceedings of IEEE ICDM, pp. 180—189, 2014.
- [5] E. E. Buckels, P. D. Trapnell, and D. L. Paulhus, Trolls just want to have fun, 2014, pp. 67:97–102.
- [6] S. Kumar, F. Spezzano, and V. Subrahmanian, "Accurately detecting trolls in slashdot zoo via decluttering," in Proceedings of IEEE/ACM ASONAM, 2014, pp. 188–195.
- [7] J. W. Patchin and S. Hinduja, "2016 cyberbullying data," 2017.
- [8] C. R. Center, "https://cyberbullying.org/bullying-laws."
- [9] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." Psychological review, vol. 63, no. 5, p. 277, 1956.
- [10] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in Proceedings of the SIGCHI CHI, 2010, pp. 1361–1370.

TWITTER SENTIMENTAL ANALYSIS

Mandela Lakshmiajitha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

In recent years, research on Twitter sentiment analysis, which analyzes Twitter data (tweets) to extract user sentiments about a topic, has grown rapidly. Many researchers prefer the use of machine learning algorithms for such analysis. This study aims to perform a detailed sentiment analysis of tweets based on ordinal regression using machine learning techniques. The proposed approach consists of first pre-processing tweets and using a feature extraction method that creates an efficient feature. Then, under several classes, these features scoring and balancing. Multinomial logistic regression (SoftMax), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF) algorithms are used for sentiment analysis classification in the proposed framework. For the actual implementation of this system, a twitter dataset publicly made available by the NLTK corpora resources is used. Experimental findings reveal that the proposed approach can detect ordinal regression using machine learning methods with good accuracy. Moreover, results indicate that Decision Trees obtains the best results outperforming all the other algorithms.

1. INTRODUCTION

With the rapid development of social networks and microblogging websites. Microblogging websites have become one of the largest web destinations for people to express their thoughts, opinions, and attitudes about different topics [1], [2]. Twitter is a widely used microblogging platform and social networking service that generates a vast amount of information.

In recent years, researchers preferably made the use of social data for the sentiment analysis of people's opinions on a product, topic, or event. Sentiment analysis, also known as opinion mining, is an important natural language processing task. This process determines the sentiment orientation of a text as positive, negative, or neutral.

Twitter sentiment analysis is currently a popular topic for research. Such analysis is useful because it gathers and classifies public opinion by analyzing big social data.

However, Twitter data have certain characteristics that cause difficulty in conducting sentiment analysis in contrast to analyzing other types of data. Tweets are restricted to 140 characters, written in informal English, contain irregular expressions, and contain several abbreviations and slang words. To address these problems, researchers have conducted studies focusing on sentiment analysis of tweets [5]. Twitter sentiment analysis approaches can be generally categorized into two main approaches, the machine learning approach, and a lexicon-based approach.

In this study, we use machine learning techniques to tackle twitter sentiment analysis. Most classification algorithms are focused on predicting nominal class data labels. However, a rule for predicting categories or labels on an ordinal scale involves many pattern recognition issues. This type of problem, known as ordinal classification or ordinal regression.

Recently, ordinal regression has received considerable attention. Ordinal regression issues in many Fields of research are very common and have often been regarded as standard nominal problems that can lead to non-optimal solutions.

In fact, Ordinal regression problems with some similarities and differences can be said to be between classification and regression. Medical research, age estimation, brain-computer interface, face recognition, facial beauty evaluation, image classification, social sciences, text classification, and more are some of the Fields where ordinal regression is found.

Some studies suggest using machine learning techniques to solve regression problems to improve the sentiment analysis classification of Twitter data performance and predict new results. The main advantage of this method is the achievement of improved results.

The current study mainly focuses on the sentiment analysis of Twitter data (tweets) using different machine learning algorithms to deal with ordinal regression problems. In this paper, we propose an approach including pre-processing tweets, feature extraction methods, and constructing a scoring and balancing system, then using different techniques of machine learning to classify tweets under several classes.

2. EXISTING SYSTEM

In recent years, researchers preferably made the use of social data for the sentiment analysis of people's opinions on a product, topic, or event. Sentiment analysis, also known as opinion mining, is an important natural language processing task. This process determines the sentiment orientation of a text as positive, negative, or neutral.

Twitter sentiment analysis is currently a popular topic for research. Such analysis is useful because it gathers and classifies public opinion by analyzing big social data. However, Twitter data have certain characteristics that cause difficulty in conducting sentiment analysis in contrast to analyzing other types of data.

Tweets are restricted to 140 characters, written in informal English, contain irregular expressions, and contain several abbreviations and slang words. To address these problems, researchers have conducted studies focusing on sentiment analysis of tweets

Most classification algorithms are focused on predicting nominal class data labels. However, a rule for predicting categories or labels on an ordinal scale involves many pattern recognition issues. This type of problem, known as ordinal classification or ordinal regression. Recently, ordinal regression has received considerable attention.

3. PROPOSED SYSTEM

Substantial work has also been performed by Go et al. [7] who proposed a solution for sentiment analysis based on tweets using distant supervision. In their method, they used training data containing tweets with emoticons, which served as noisy labels. They built models using naive Bayes classifiers, maximum entropy (MaxEnt), and support vector machine. Their features comprised unigrams, bigrams and POS. They concluded that SVM outperformed other models and that unigrams were more effective as features.

There has been a growing interest in Sentiment Analysis based on Twitter data research as well as ordinal regression over the past decade. Ordinal regression problem is one of the main study areas in machine learning and data mining, with the aim of classifying patterns using a categorical scale showing a natural order between labels [12][14]. However, less attention was paid to the problems of ordinal regression (also known as ordinal classification). Recently, the field of ordinal regression has developed, many algorithms have been proposed from a machine learning approach for ordinal regression such as support vector ordinal regression and the perceptron ranking (PRank) algorithm.

Li and Lin proposed a reduction framework based on expanded examples from ordinal regression to binary classification. The framework can perform with any reasonable cost matrix and any binary classifier. The framework consists of three steps: removing expanded examples from the original examples, learning a binary classifier with any binary classification algorithm on the expanded examples, and building a binary classifier ranking rule. Their framework enables not only good ordinal regression algorithms based on well-tuned binary classification methods, but also new generalization boundaries for ordinal regression to be derived from recognized binary classification boundaries. Their framework also unifies many current ordinal regression algorithms.

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

- Valid Input : identified classes of valid input must be accepted.
- Invalid Input : identified classes of invalid input must be rejected.
- Functions : identified functions must be exercised.
- Output : identified classes of application outputs must be exercised.
- Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

ECONOMICAL FEASIBILITY**TECHNICAL FEASIBILITY****SOCIAL FEASIBILITY****ECONOMICAL FEASIBILITY**

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

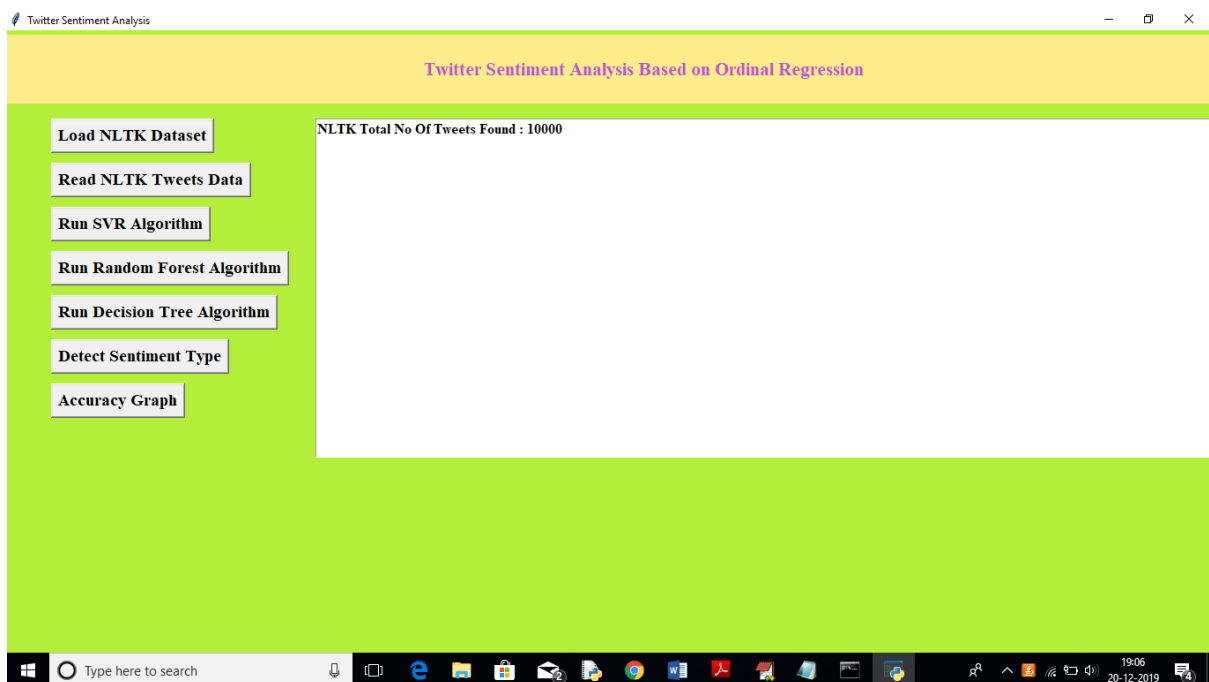
This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

4. OUTPUT SCREENS

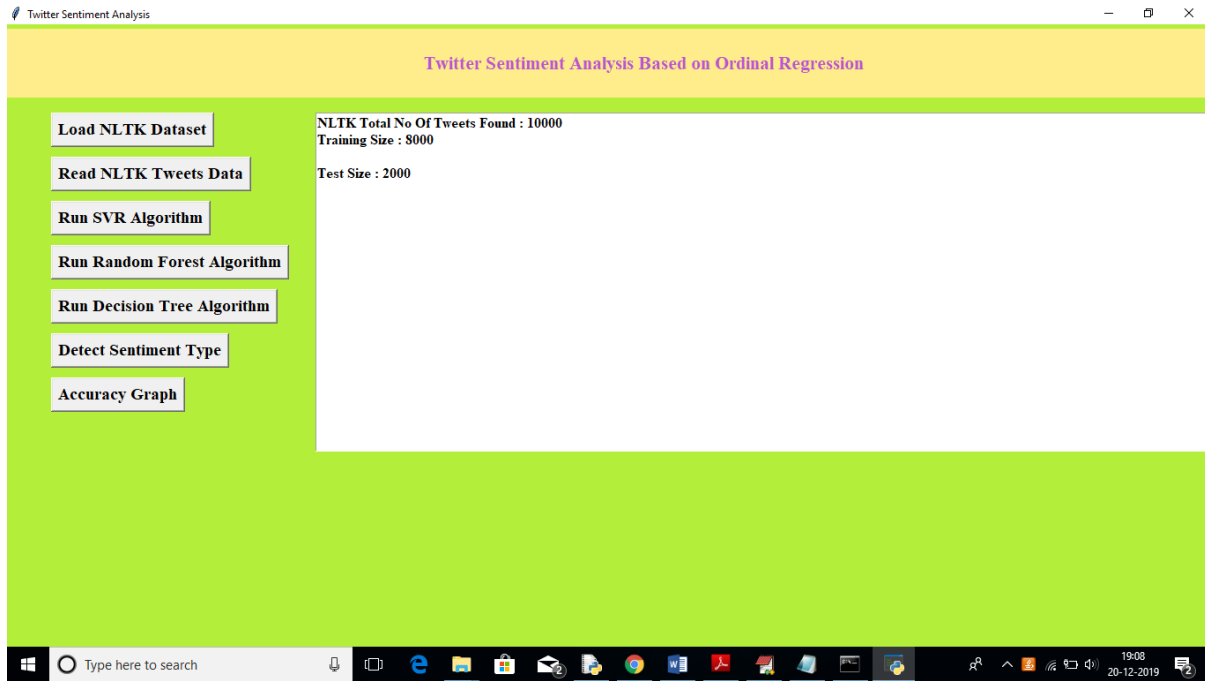
To run this project double click on 'run.bat' file to get below screen



In above screen click on 'Load NLTK Dataset' to load tweets dataset from NLTK library



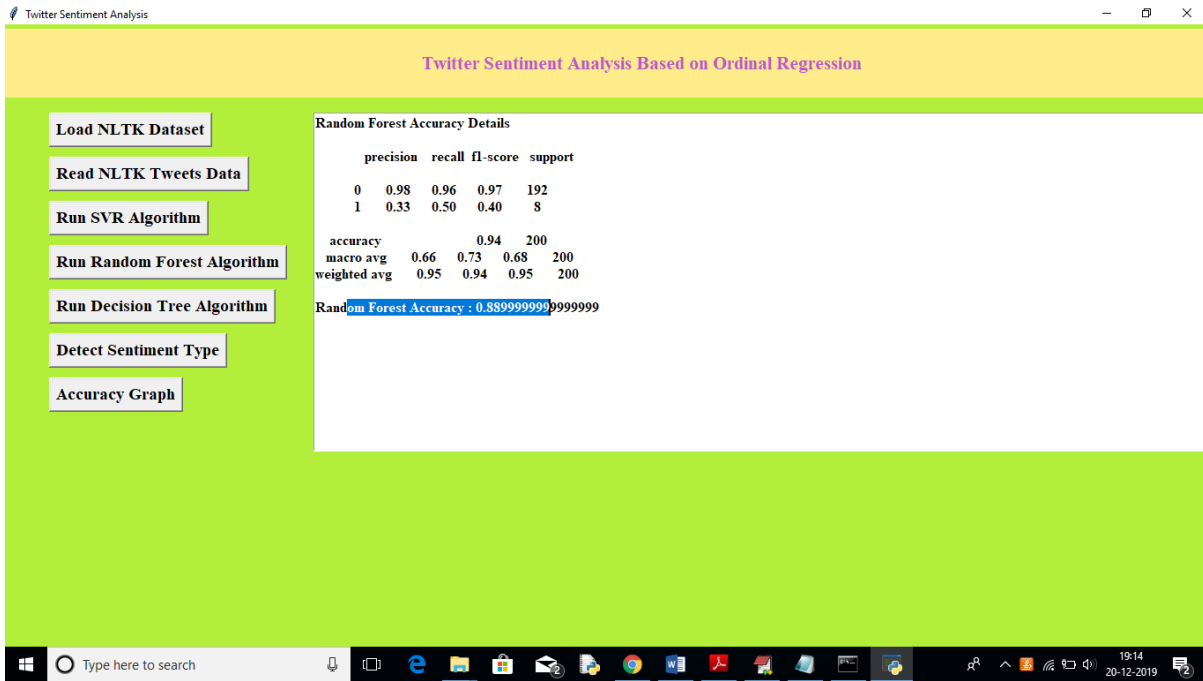
In above screen we can see total 10000 tweets are there in NLTK library, now click on 'Read NLTK Tweets Data' button to read all tweets and to build TFIDF vector. Upon each button click you need to wait for some seconds to get output. See below screen



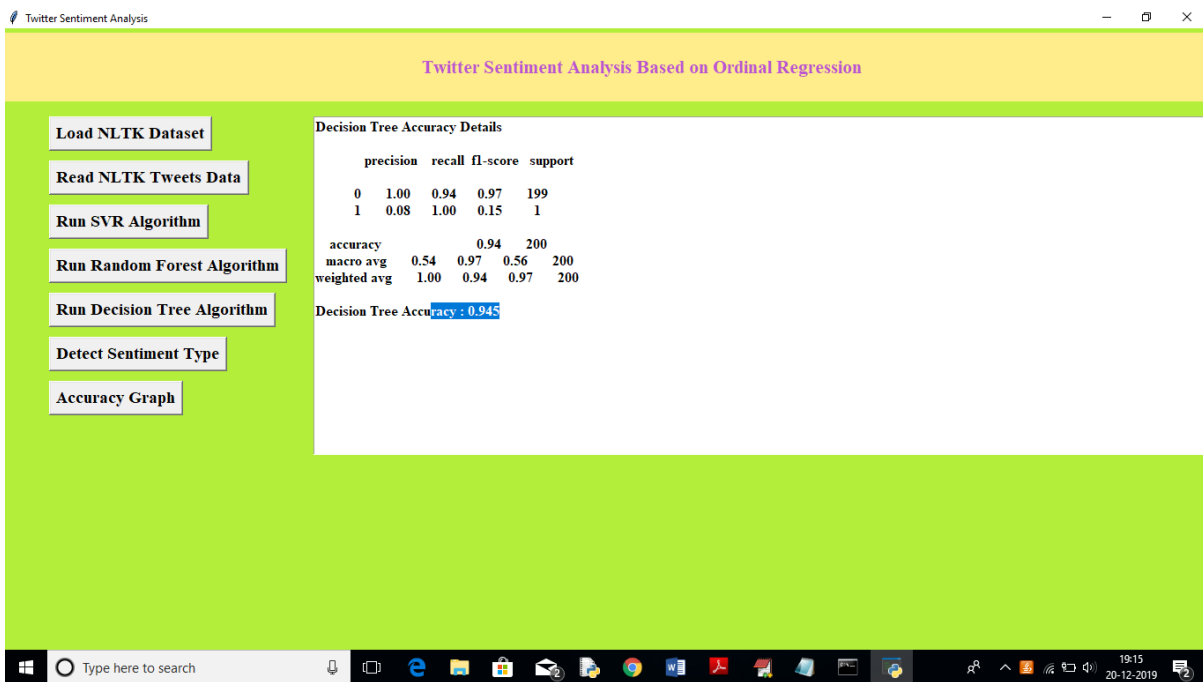
In above screen we can see total 8000 tweets vector used for training purpose and 2000 tweets used for testing purpose. Now click on ‘Run SVR Algorithm’ to build train model on that dataset and to calculate accuracy



In above screen we can see SVR generate 0.71% prediction accuracy, now click on ‘Run Random Forest Algorithm’ button to calculate its accuracy



In above screen Random Forest got 0.88% accuracy, now click on ‘Run Decision Tree Algorithm’ button to calculate its accuracy



In above screen Decision Tree got 0.94% accuracy, Now click on ‘Detect Sentiment Type’ button and upload test tweets to predict it sentiment. In test folder inside test.txt you can see there is no sentiment label and application will detect it.

5. CONCLUSION

This study aims to explain sentiment analysis of twitter data regarding ordinal regression using several machine learning techniques. In the context of this work, we present an approach that aims to extract Twitter sentiment analysis by building a balancing and scoring model, afterward, classifying tweets into several ordinal classes using machine learning classifiers. Classifiers, such as Multinomial logistic regression, Support vector regression, Decision Trees, and Random Forest, are used in this study. This approach is optimized using Twitter data set that is publicly available in the NLTK corpora resources.

Experimental results indicate that Support Vector Regression and Random Forest have an almost similar accuracy, which is better than that of the Multinomial logistic regression classifier. However, the Decision Tree gives the highest accuracy at 91.81%. Experimental results concluded that the proposed model can detect ordinal regression in Twitter using machine learning methods with a good accuracy result. The performance of the model is measured using accuracy, Mean Absolute Error, and Mean Squared Error.

In the future, we plan to improve our approach by attempting to use bigrams and trigrams. Furthermore, we intend to investigate different machine learning techniques and deep learning techniques, such as Deep Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks.

6. REFERENCES

- [1] B. O'Connor, R. Balasubramanian, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series.," in *Proc. ICWSM*, 2010, vol. 11, nos. 122_129, pp. 1_2.
- [2] M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons," in *Proc. 5th Int. Conf. Inf., Intell., Syst. Appl. (IISA)*, Jul. 2014, pp. 94_97.
- [3] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proc. 20th Int. Conf. Comput. Linguistics*, Aug. 2004, p. 1367.

- [4] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, Oct./Nov. 2005, pp. 625_631.
- [5] H. Saif, M. Fernández, Y. He, and H. Alani, "Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold," in *Proc. 1st International Workshop Emotion Sentiment Social Expressive Media, Approaches Perspect. AI (ESSEM)*, Turin, Italy, Dec. 2013.
- [6] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul. 2016, pp. 628_632.
- [7] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, vol. 150, no. 12, pp. 1_6, 2009.
- [8] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multi-class sentiment analysis in Twitter," *IEEE Access*, vol. 5, pp. 20617_20639, 2017.
- [9] R. Sara, R. Alan, N. Preslav, and S. Veselin, "SemEval-2016 task 4: Sentiment analysis in Twitter," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 1_18.

ANALYZING AND DETECTING MONEY-LAUNDERING ACCOUNTS IN ONLINE SOCIAL NETWORKS.

Maruboina Karthik (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:

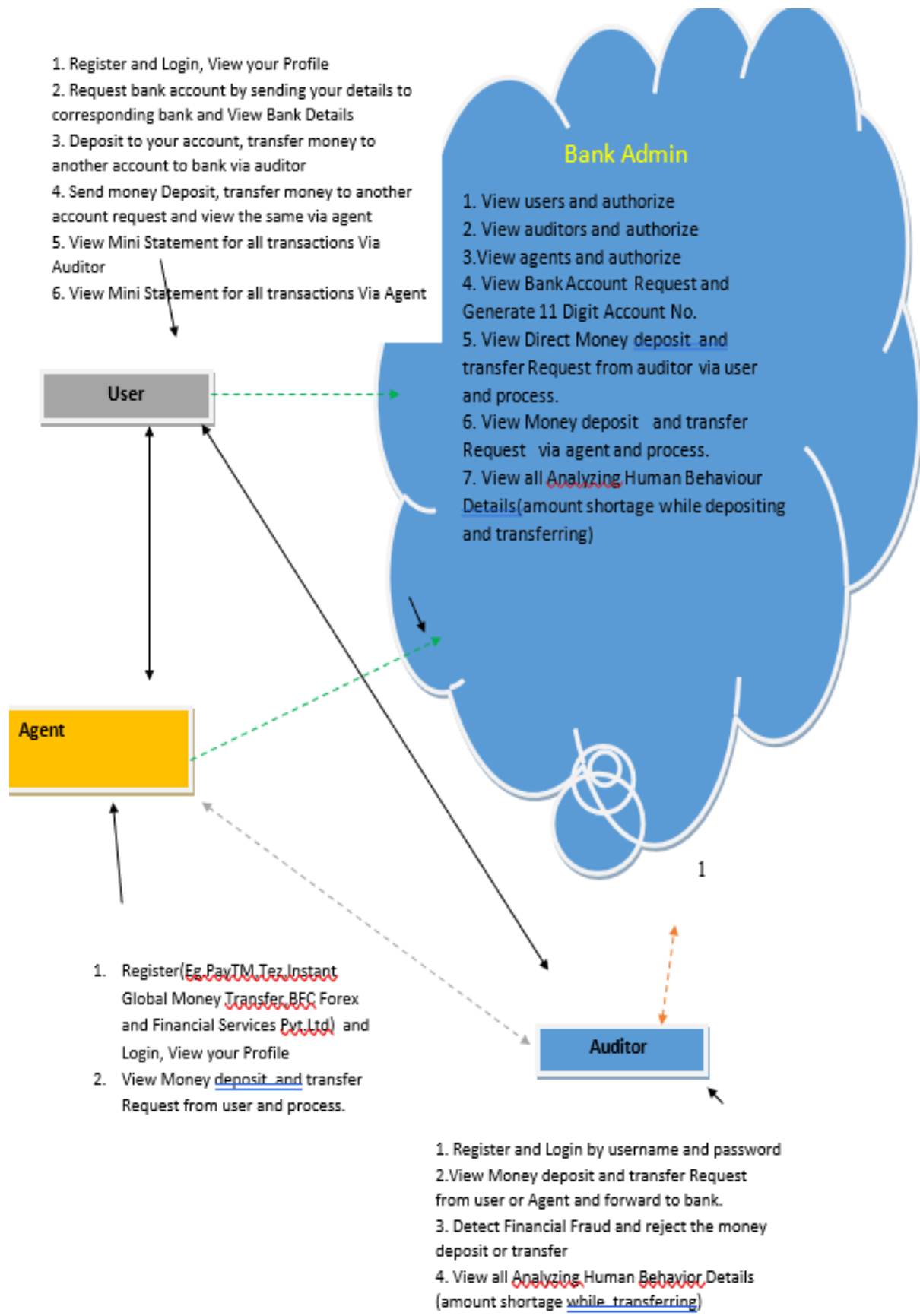
Virtual currency in OSNs plays an increasingly important role in supporting various financial activities such as currency exchange, online shopping, and paid games. Users usually purchase virtual currency using real currency. This fact motivates attackers to instrument an army of accounts to collect virtual currency unethically or illegally with no or very low cost and then launder the collected virtual money for massive profit. Such attacks not only introduce significant financial loss of victim users, but also harm the viability of the ecosystem. It is therefore of central importance to detect malicious OSN accounts that engage in laundering virtual currency. To this end, we extensively study the behavior of both malicious and benign accounts based on operation data collected from Tencent QQ, one of the largest OSNs in the world. Then, we devise multi-faceted features that characterize accounts from three aspects: account viability, transaction sequences, and spatial correlation among accounts. Finally, we propose a detection method by integrating these features using a statistical classifier, which can achieve a high detection rate of 94.2 percent

1. INTRODUCTION

Fraud is a worldwide phenomenon that affects public and private organizations, covering a wide variety of illegal practices and acts that involve intentional deception or misrepresentation. According to the Association of Certified Fraud Examiners (ACFE) [1] fraud includes any intentional or deliberate act of depriving another of property or money by cunning, deception or other unfair acts. The 2016 PwC Global Economic Crime Survey report describes that more than a third of organizations worldwide have been victims of some kind of economic crime such as asset misappropriation, bribery, cybercrime, fraud and

money laundering. Approximately 22% of respondents experienced losses of between one hundred thousand and one million, 14% suffered losses of more than one million and 1% of those surveyed suffered losses of one hundred million dollars. These high loss rates represent a rising trend in costs caused by fraud. In organizations, 56% of cases are related to internal fraud and 40% to external, this difference is since any individual related to accounting and financial activities is considered a potential risk factor for fraud [2]. When observing the behavior of people in the scope of business processes, it can be concluded that the human factor is closely linked and related to the fraud triangle theory of the Donald R. Cressey [3], where three basic concepts: pressure, opportunity and rationalization; are needed. Nowadays, there are different solutions in the commercial field [4], [5] as well as the the academic field, where some works in progress had been identified [6], [7] aimed at detecting financial fraud. In both cases, these solutions are focused on the use of different tools that perform statistical and parametric analysis, as well as behavioral analysis, based on data mining techniques and Big Data; but none of them solve the problem of detection financial fraud in real time. FraudFind, unlike other proposals, detects, reports and stores fraudulent activities in real time through the periodic analysis of the information generated by users for further analysis and treatment. This paper presents FraudFind, a conceptual framework that allows detecting and identifying potential criminals who work in the banking field, in real time, based on the theory of the fraud triangle. For the design of the FraudFind framework, some software components related to the processing of informtion were analyzed, among them, RabbitMQ, Logstash and ElasticSearch. In addition, the computerization of the triangle of fraud and the use of semantic techniques will allow finding possible bank delinquents with a lower false positive rate. The rest of the document is structured as follows. Section 2 presents the theoretical framework on the definition of Fraud and the concept of the fraud triangle. Section 3 presents the related works found in the literature. Section 4 details the architecture of the model and the prototype to be implemented as future work. Section 5 continues with the discussion and section 6 concludes with the conclusions and future work.

2. ARCHITECTURE DIAGRAM



ATTACKER

--- Create an attacker to deposit money by un authorized way

--- Create an attacker to transfer money by un authorized way

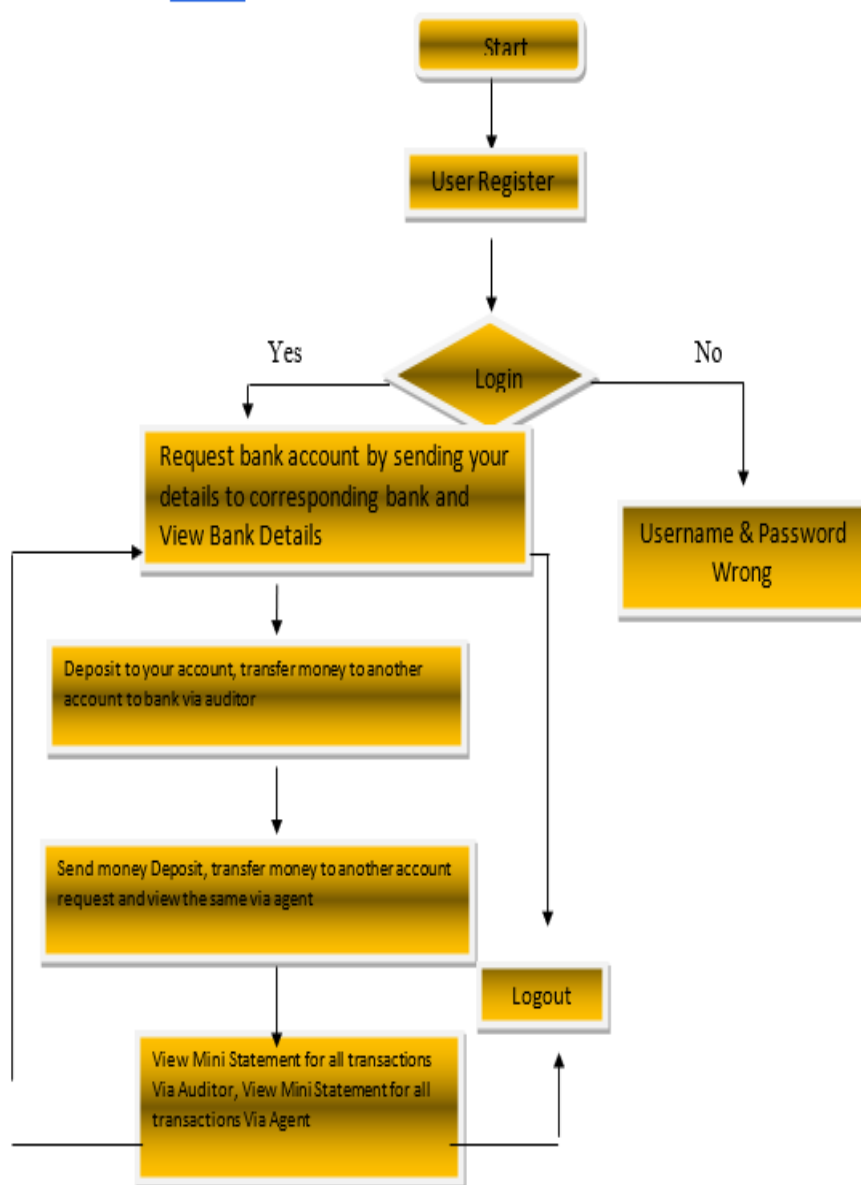
Enter Agent Name or user name ---

Enter user account no --

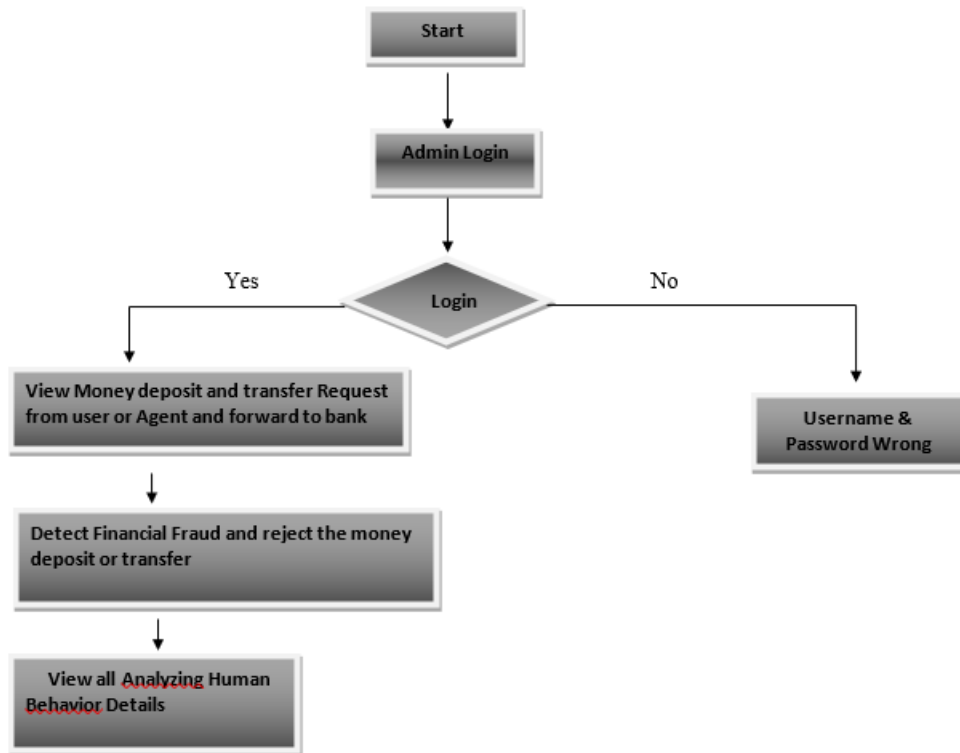
Enter money --

Give transfer or deposit button

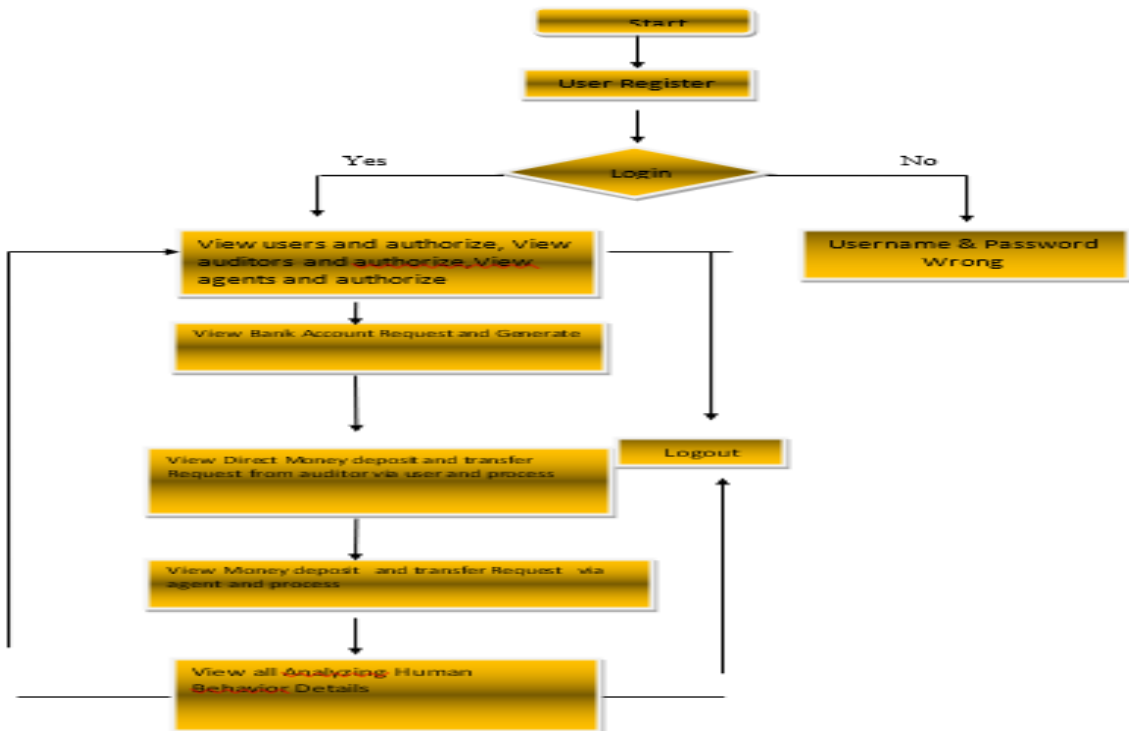
➤ **Flow Chart :** User

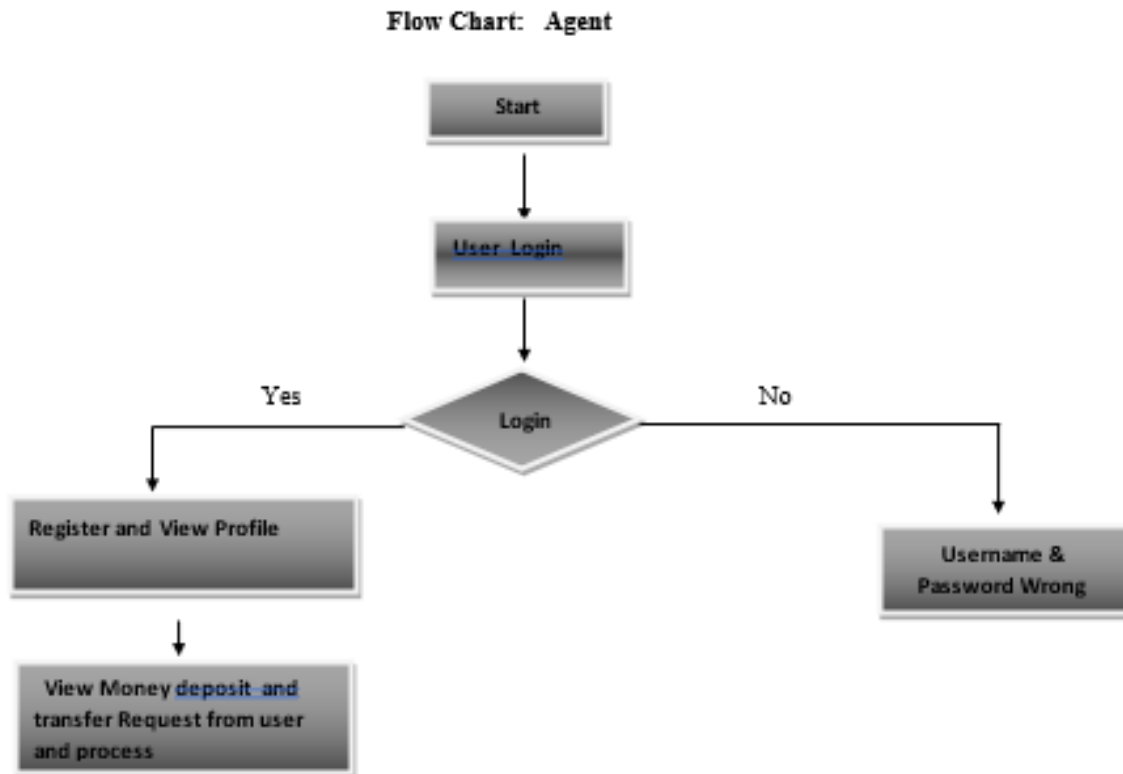


Flow Chart: Auditor



Flow Chart: Bank Admin





4. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

5. CONCLUSIONS

The present work proposes Fraud Find, a conceptual framework to detect financial fraud supported by the fraud triangle factors which, compared to the classic audit analysis, makes a significant contribution to the early detection of fraud within an organization. Taking into account human behavior factors, it is possible to detect unusual transactions that would have not been considered using traditional audit methods. These patterns of behavior can be found

in the information that users generate when using the different applications on a workstation. The collected data is examined during data mining techniques to obtain patterns of suspicious behavior evidencing possible fraudulent behavior. Nevertheless, the legal framework and the different regulations that are applied in public and private institutions of a particular region represent a high risk for the non-implementation of this architecture as an alternative solution. Future work will have as its main objective the implementation and evaluation of the framework as a tool for continuous auditing within an organization.

6. REFERENCES

- [1] "ACFE Asociaci3n de Examinadores de Fraudes Certificados," (Date last accessed 15-July-2014). [Online]. Available: <http://www.acfe.com/uploadedfiles/acfewebiste/content/documents/rtn-2010.pdf>
- [2] "PwC," (Date last accessed 15-July-2014). [Online]. Available: <https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf>
- [3] N. B. Omar and H. F. M. Din, "Fraud diamond risk indicator: An assessment of its importance and usage," in 2010 International Conference on Science and Social Research (CSSR 2010). IEEE, dec 2010.
- [4] "Lynx," (Date last accessed 15-July-2014). [Online]. Available: <http://www.iic.uam.es/soluciones/banca/lynx/>
- [5] "Ibm," (Date last accessed 15-July-2014). [Online]. Available: <https://www.ibm.com/developerworks/ssa/local/analytics/prevencion-de-fraude/index.html>
- [6] C. Holton, "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem," *Decision Support Systems*, vol. 46, no. 4, pp. 853–864, mar 2009.
- [7] S. Hoyer, H. Zakhariya, T. Sandner, and M. H. Breitner, "Fraud prediction and the human factor: An approach to include human behavior in an automated fraud audit," in 2012 45th Hawaii International Conference on System Sciences. IEEE, jan 2012.



MACHINE LEARNING MODEL FOR AVERAGE FUEL CONSUMPTION IN HEAVY VEHICLE

Meka Subrahmanya Vara Prasad (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract

we used vehicle travel distance rather than the traditional time period when developing individualized machine learning models for fuel consumption. This approach is used in conjunction with seven predictors derived from vehicle speed and road grade to produce a highly predictive neural network model for average fuel consumption in heavy vehicles. The proposed model can easily be developed and deployed for each individual vehicle in a fleet in order to optimize fuel consumption over the entire fleet. The predictors of the model are aggregated over fixed window sizes of distance travelled. Different window sizes are evaluated and the results show that a 1 km window is able to predict fuel consumption with a 0.91 coefficient of determination and mean absolute peak-to-peak percent error less than 4% for routes that include both city and highway duty cycle segments.

1. INTRODUCTION

Fuel consumption models for vehicles are of interest to manufacturers, regulators, and consumers. They are needed across all the phases of the vehicle life-cycle. we focus on modeling average fuel consumption for heavy vehicles during the operation and maintenance phase. In general, techniques used to develop models for fuel consumption fall under three main categories:

Physics-based models:-which are derived from an in-depth understanding of the physical system. These models describe the dynamics of the components of the vehicle

Machine learning models:- which are data-driven and represent an abstract mapping from an input space consisting of a selected set of predictors to an output space that represents the target output.

Statistical models:- which are also data-driven and establish a mapping between the probability distribution of a selected set of predictors and the target outcome.

a model that can be easily developed for individual heavy vehicles in a large fleet is proposed.

Relying on accurate models of all of the vehicles in a fleet, a fleet manager can optimize the route planning for all of the vehicles based on each unique vehicle predicted fuel consumption thereby ensuring the route assignments are aligned to minimize overall fleet fuel consumption. These types of fleets exist in various sectors including, road transportation of goods , public transportation , construction trucks and refuse trucks .

For each fleet, the methodology must apply and adapt to many different vehicle technologies (including future ones) and configurations without detailed knowledge of the vehicles specific physical characteristics and measurements.

2. EXISTING SYSTEM

model that can be easily developed for individual heavy vehicles in a large fleet is proposed. Relying on accurate models of all of the vehicles in a fleet, a fleet



manager can optimize the route planning for all of the vehicles based on each unique vehicle predicted fuel consumption thereby ensuring the route assignments are aligned to minimize overall fleet fuel consumption. This approach is used in conjunction with seven predictors derived from vehicle speed and road grade to produce a highly predictive neural network model for average fuel consumption in heavy vehicles.

Different window sizes are evaluated and the results show that a 1 km window is able to predict fuel consumption with a 0.91 coefficient of determination and mean absolute peak-to-peak percent error less than 4% for routes that include both city and highway duty cycle segments.

Disadvantages of existing system:

Physics-based models, which are derived from an in-depth understanding of the physical system. These models describe the dynamics of the components of the vehicle each time step using detailed mathematical equations.

Statistical models, which are also data-driven and establish a mapping between the probability distribution of a selected set of predictors and the target outcome.

3. PROPOSED SYSTEM

As mentioned above Artificial Neural Networks (ANN) are often used to develop digital models for complex systems. The models proposed in [15] highlight some of the difficulties faced by machine learning models when the input and output have different domains. In this study, the input is aggregated in the time domain over 10 minutes intervals and the output is fuel consumption over the distance traveled during the same time

period. The complex system is represented by a transfer function $F(p) = o$, where $F(\cdot)$ represents the system, p refers to the input predictors and o is the response of the system or the output. The ANNs used in this paper are Feed Forward Neural Networks (FNN).

Training is an iterative process and can be performed using multiple approaches including particle swarm optimization [20] and back propagation. Other approaches will be considered in future work in order to evaluate their ability to improve the model's predictive accuracy. Each iteration in the training selects a pair of (input, output) features from F_{tr} at random and updates the weights in the network. This is done by calculating the error between the actual output value and the value predicted by the model

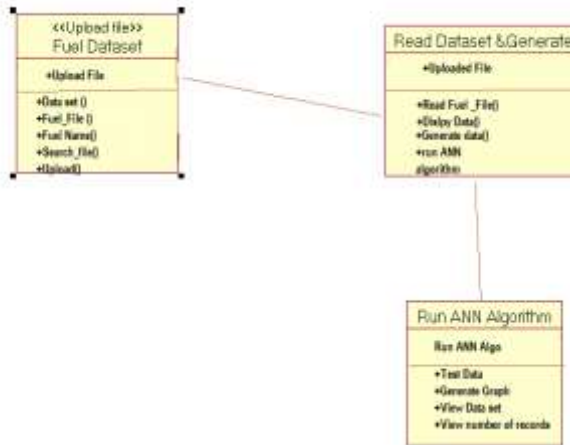
ADVANTAGES OF PROPOSED SYSTEM:

Data is collected at a rate that is proportional to its impact on the outcome. When the input space is sampled with respect to time, the amount of data collected from a vehicle at a stop is the same as the amount of data collected when the vehicle is moving.

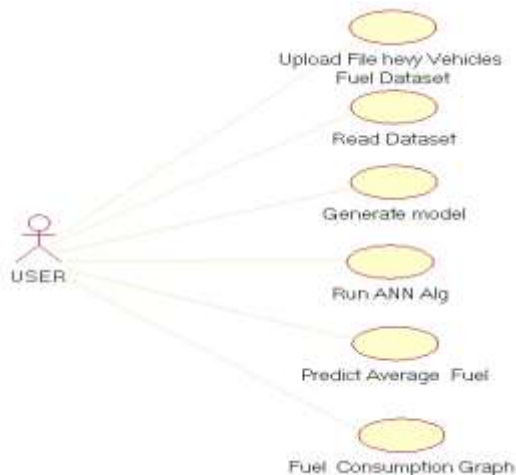
The predictors in the model are able to capture the impact of both the duty cycle and the environment on the average fuel consumption of the vehicle (e.g., the number of stops in an urban traffic over a given distance).

Data from raw sensors can be aggregated on-board into few predictors with lower storage and transmission bandwidth requirements. Given the increase in computational capabilities of new vehicles, data summarization is best performed on-board near the source of the data.

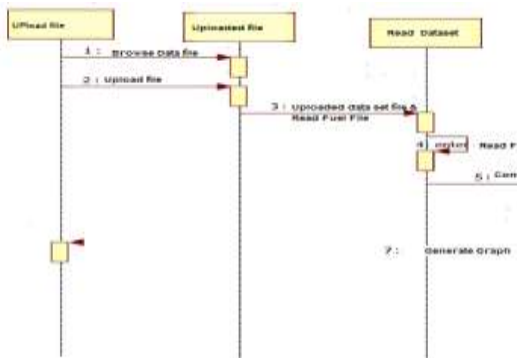
CLASS DIAGRAMS



USE CASE DIAGRAMS



SEQUENCE DIAGRAMS



4. CONCLUSIONS

This paper presented a machine learning model that can be conveniently developed

for each heavy vehicle in a fleet. The model relies on seven predictors: number of stops, stop time, average moving speed, characteristic acceleration, aerodynamic speed squared, change in kinetic energy and change in potential energy. The last two predictors are introduced in this paper to help capture the average dynamic behaviour of the vehicle. All of the predictors of the model are derived from vehicle speed and road grade. These variables are readily available from telematics devices that are becoming an integral part of connected vehicles. Moreover, the predictors can be easily computed on-board from these two variables.

The model predictors are aggregated over a fixed distance traveled (i.e., window) instead of a fixed time interval. This mapping of the input space to the distance domain aligns with the domain of the target output, and produced a machine learning model for fuel consumption with an RMSE < 0.015 l/100 km.

Different model configurations with 1, 2, and 5 km window sizes were evaluated. The results show that the 1 km window has the highest accuracy. This model is able to predict the actual fuel consumption on a per 1 km-basis with a CD of 0.91. This performance is closer to that of physics-based models and the proposed model improves upon previous machine learning models that show comparable results only for entire long-distance trips.

Selecting an adequate window size should take into consideration the cost of the model in terms of data collection and onboard computation. Moreover, the window size is likely to be application-dependent. For fleets with short trips (e.g., construction vehicles within a site) or



urban traffic routes, a 1 km window size is recommended. For long-haul fleets, a 5 km window size may be sufficient. In this study, the duty cycles consisted of both highway and city traffic and therefore, the 1 km window was more adequate than the 5 km window. Future work includes understanding these differentiating factors and the selection of the appropriate window size. Expanding the model to other vehicles with different characteristics such as varying masses and aging vehicles is being studied. Predictors for these characteristics will be added in order to allow for the same model to capture the impact on fuel consumption due to changes in vehicle mass and wear.

5. REFERENCES

- [1] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2016, pp. 452-455.
- [2] L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in *IEEE Access*, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.
- [3] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 700-704, doi: 10.1109/ICCS45141.2019.9065812.
- [4] Z. Singla, S. Randhawa and S. Jain, "Statistical and sentiment analysis of consumer product reviews," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203960.
- [5] C. Nanda, M. Dua and G. Nanda, "Sentiment Analysis of Movie Reviews in Hindi Language Using Machine Learning," 2018 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2018, pp. 1069-1072, doi: 10.1109/ICCSP.2018.8524223.
- [6] B. Seetharamulu, B. N. K. Reddy and K. B. Naidu, "Deep Learning for Sentiment Analysis Based on Customer Reviews," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-5, doi: 10.1109/ICCCNT49239.2020.9225665.
- [7] Rahul, V. Raj and Monika, "Sentiment Analysis on Product Reviews," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2019, pp. 5-9, doi: 10.1109/ICCCIS48478.2019.8974527.
- [8] Y. Saito and V. Klyuev, "Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019, pp. 681-685, doi: 10.23919/ICACT.2019.8702039.

A DEEP LEARNING-BASED APPROACH FOR INAPPROPRIATE CONTENT DETECTION AND CLASSIFICATION OF YOUTUBE VIDEOS

Mindyala Sirisha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

G. Ramesh Kumar, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The exponential growth of videos on YouTube has attracted billions of viewers among which the majority belongs to a young demographic. Malicious uploaders also find this platform as an opportunity to spread upsetting visual content, such as using animated cartoon videos to share inappropriate content with children. Therefore, an automatic real-time video content filtering mechanism is highly suggested to be integrated into social media platforms. In this study, a novel deep learning-based architecture is proposed for the detection and classification of inappropriate content in videos. For this, the proposed framework employs an ImageNet pre-trained convolutional neural network (CNN) model known as EfficientNet-B7 to extract video descriptors, which are then fed to bidirectional long short-term memory (BiLSTM) network to learn effective video representations and perform multiclass video classification. An attention mechanism is also integrated after BiLSTM to apply attention probability distribution in the network. These models are evaluated on a manually annotated dataset of 111,156 cartoon clips collected from YouTube videos. Experimental results demonstrated that EfficientNet-BiLSTM (accuracy 95.66%) performs better than attention mechanism based EfficientNet-BiLSTM (accuracy 95.30%) framework. Secondly, the traditional machine learning classifiers perform relatively poor than deep learning classifiers. Overall, the architecture of EfficientNet and BiLSTM with 128 hidden units yielded state-of-the-art performance (f1 score 0.9267). Furthermore, the performance comparison against existing state-of-the-art approaches verified that BiLSTM on top of CNN captures better contextual information of video descriptors in network architecture, and hence achieved better results in child inappropriate video content detection and classification.

1. INTRODUCTION

The creation and consumption of videos on social media platforms have grown drastically over the past few years. Among the social media sites, YouTube predominates as a video sharing platform with plethora of videos from diverse categories. According to YouTube statistics [1], the global user base of YouTube is over 2 billion registered users and more than 500 hours of video content is uploaded every minute. Consequently, billions of hours of videos are available where users of all age groups can explore generic as well as personalized content [2]. Considering such a large-scale crowd sourced database, it is extremely challenging to monitor and regulate the uploaded content as per platform guidelines. This creates opportunities for malicious users to indulge in spamming activities by misleading the audiences with falsely advertised content (i.e., video, audio or text). The most disruptive behavior by malicious users is to expose the young audiences to disturbing content, particularly when it is fabricated as safe for them. Children today spend most of their time on the Internet and the YouTube platform for them has distinctly established itself as an alternative to traditional screen media (e.g., television) [3], [4]. The YouTube press release [5] also confirmed the high popularity of this social media site among younger audiences compared to other age groups, and the reason for this high level of approval is due to fewer restrictions [6].

Unlike television, children can be presented with any type of content on the Internet due to lack of regulations. Exposing children to disturbing content is considered as one among other internet safety threats (like cyber bullying, cyber predators, hate etc.) [7]. Bushman and Huesmann [8] confirmed that frequent exposure to disturbing video content may have a short-term or long-term impact on children's behavior, emotions and cognition. Many reports [9]_[12] identified the trend of distributing inappropriate content in children's videos. This trend got people's attention when mainstream media reported about the Elsatage controversy [13], [14], where such video material was found on YouTube featuring famous childhood cartoon characters (i.e., Disney characters, superheroes, etc.) portrayed in disturbing scenes; for instance, performing mild violence, stealing, drinking alcohol and involving in nudity or sexual activities.

In an attempt to provide a safe online platform, laws like the children's online privacy protection act (COPPA) imposes certain requirements on websites to adopt safety mechanisms for children under the age of 13. YouTube has also included a "safety mode" option to filter out unsafe content. Apart from that, YouTube developed the YouTube Kids

application to allow parental control over videos that are approved as safe for a certain age group of children [15]. Regardless of YouTube's efforts in controlling the unsafe content phenomena, disturbing videos still appear [16]_[19] even in YouTube Kids [20] due to difficulty in identifying such content. An explanation for this may be that the rate at which videos are uploaded every minute makes YouTube vulnerable to unwanted content. Besides, the decision-making algorithms of YouTube rely heavily on the metadata of video (i.e., video title, video description, view count, rating, tags, comments, and community _ags). Hence, altering videos based on the metadata and community _agging is not sufficient to assure the safety of children [21]. Many cases exist on YouTube where safe video titles and thumbnails are used for disturbing content to trick children and their parents. The sparse inclusion of child inappropriate content in videos is another common technique followed by malicious up loaders.

2. EXISTING SYSTEM

Rea *et al.* [37] proposed a periodicity-based audio feature extraction method which was later combined with visual features for illicit content detection in videos.

The machine learning algorithms are usually employed as classifiers Liu *et al.* [38] classified the periodicity-based audio and visual segmentation features through support vector machine (SVM) algorithm with Gaussian radial basis function (RBF) kernel. Later on, they extended the framework [39] by applying the energy envelope (EE) and bag-of-words (BoW)-based audio representations and visual features.

Ulges *et al.* [23] used MPEG motion vectors and Mel-frequency cepstral coefficient (MFCC) audio features with skin color and visual words. Each feature representation

is processed through an individual SVM classifier and combined in a weighted sum of late fusion. Ochoa *et al.* [40] performed binary video genre classification for adult content detection by processing the spatiotemporal features with two types of SVM algorithms: sequential minimal optimization (SMO) and LibSVM.

Jung *et al.* [41] worked with the one dimensional signal of spatiotemporal motion trajectory and skin color. Tang *et al.* [42] proposed a pornography detection system_PornProbe, based on a hierarchical latent Dirichlet allocation (LDA) and SVM algorithm. This system combined an unsupervised clustering in LDA and supervised learning in SVM, and achieved high efficiency than a single SVM classifier. Lee *et al.* [43] presented a multilevel

hierarchical framework by taking the multiple features of different temporal domains. Lopes *et al.* [44] worked with the bag-of-visual features (BoVF) for obscenity detection.

Kaushal *et al.* [21] performed supervised learning to identify the child unsafe content and content uploaders by feeding the machine learning classifiers (i.e., random forest, K-nearest neighbor, and decision tree) with video-level, user-level and comment-level metadata of YouTube Reddy *et al.* [45] handled the explicit content problem of videos through text classification of YouTube comments. They applied bigram collocation and fed the features to the naïve Bayes classifier for final classification.

Disadvantages

An existing system doesn't ANALYSIS OF PRE-TRAINED CNN MODEL VARIANTS.

An existing system doesn't ANALYSIS OF EFFICIENT-NET FEATURES WITH DIFFERENT CLASSIFIER VARIANTS.

4. PROPOSED SYSTEM

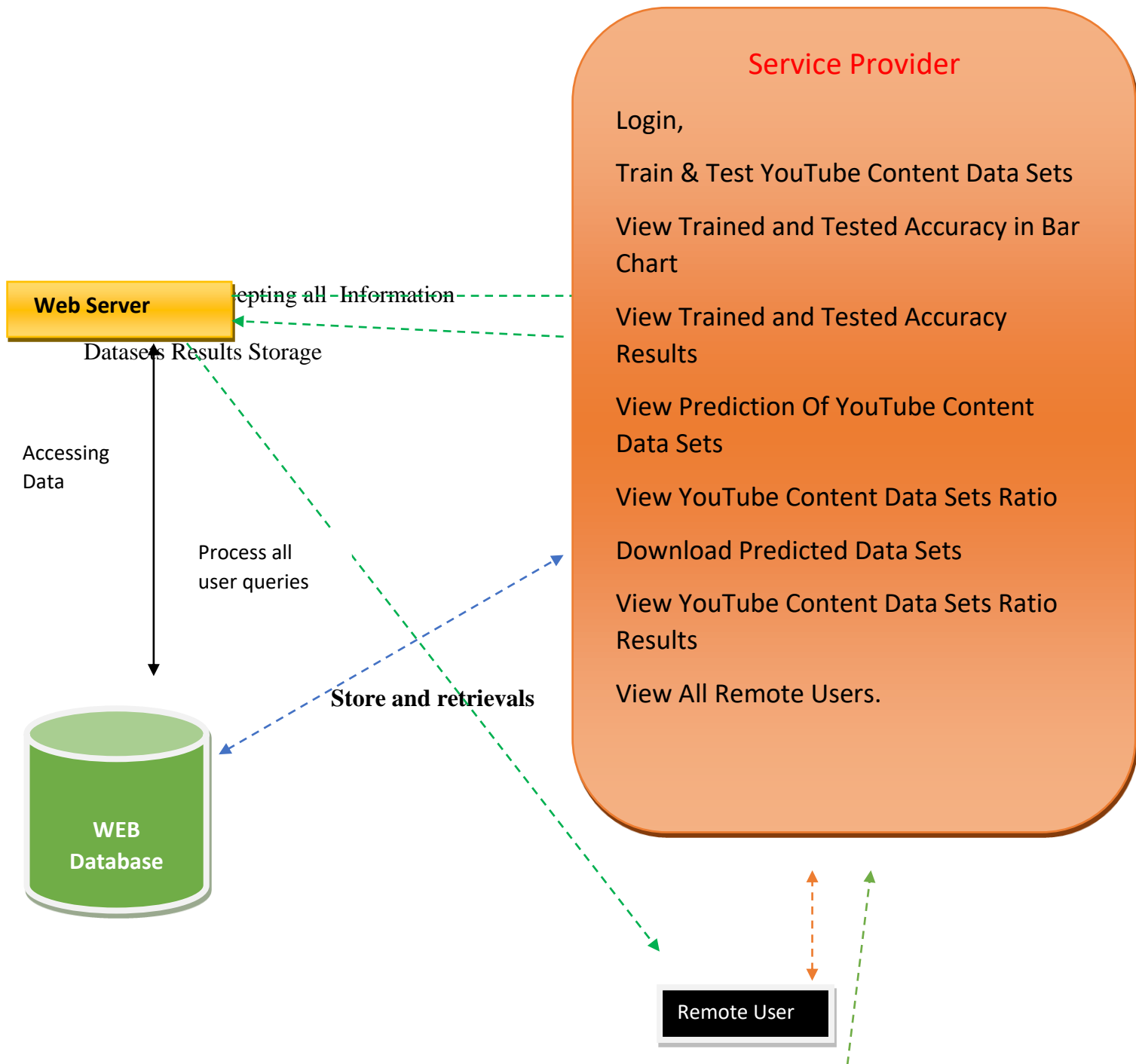
1. The system proposes a novel CNN (EfficientNet-B7) and BiLSTM-based deep learning framework for inappropriate video content detection and classification.
2. The system presents a manually annotated ground truth video dataset of 1860 minutes (111,561 seconds) of cartoon videos for young children (under the age of 13). All videos are collected from YouTube using famous cartoon names as search keywords. Each video clip is annotated for either safe or unsafe class. For the unsafe category, fantasy violence and sexual-nudity explicit content are monitored in videos. We also intend to make this dataset publicly available for the research community.
3. The system evaluates the performance of our proposed CNN-BiLSTM framework. Our multiclass video classifier achieved the validation accuracy of 95.66%. Several other state-of-the-art machine learning and deep learning architectures are also evaluated and compared for the task of inappropriate video content detection.

Advantages

The most frequent applications of image/video classification employed the convolutional neural networks.

The EfficientNet model is a convolutional neural network model and scaling method that uniformly scales network depth, width and resolution through compound co efficient.

Architecture Diagram

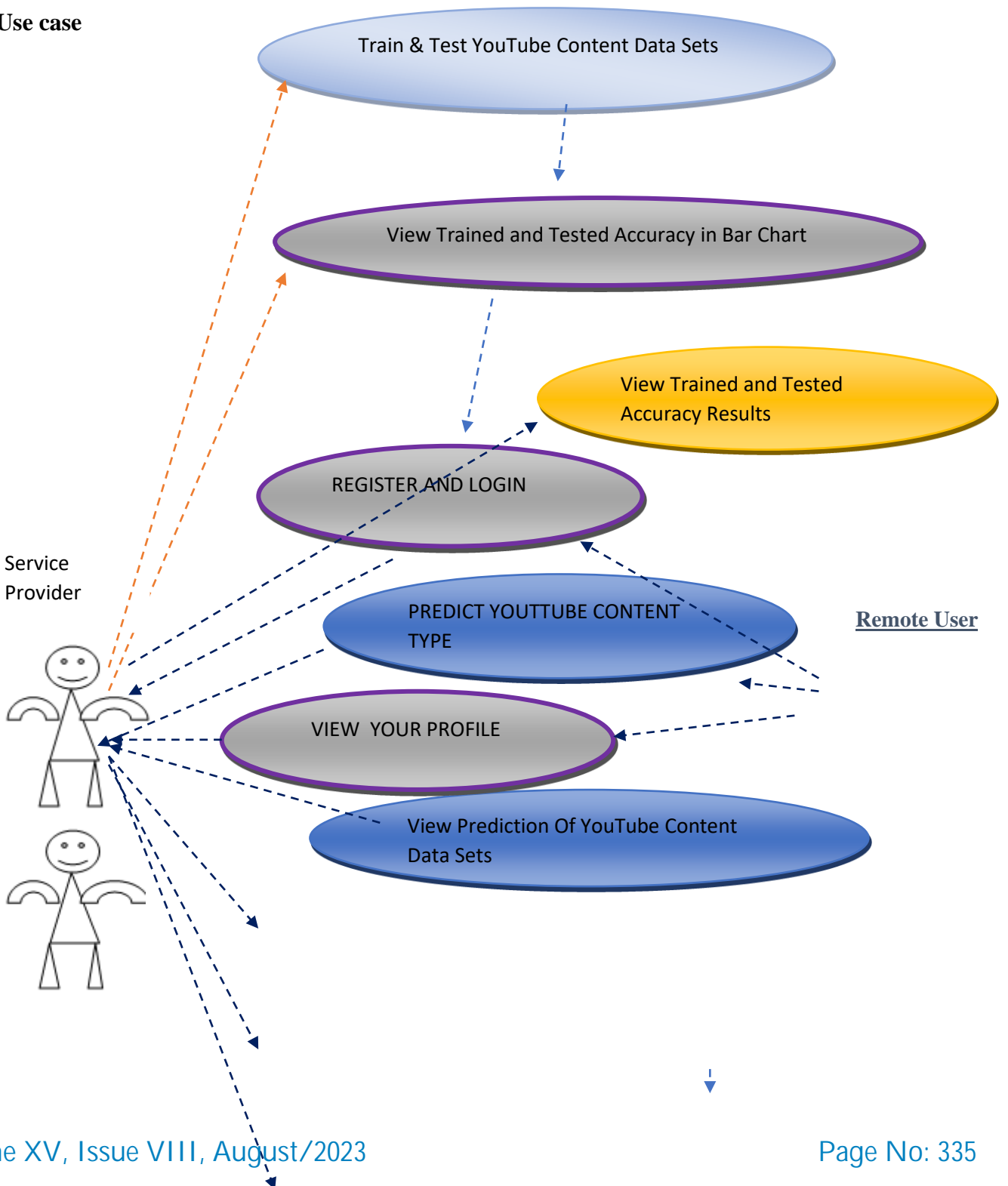


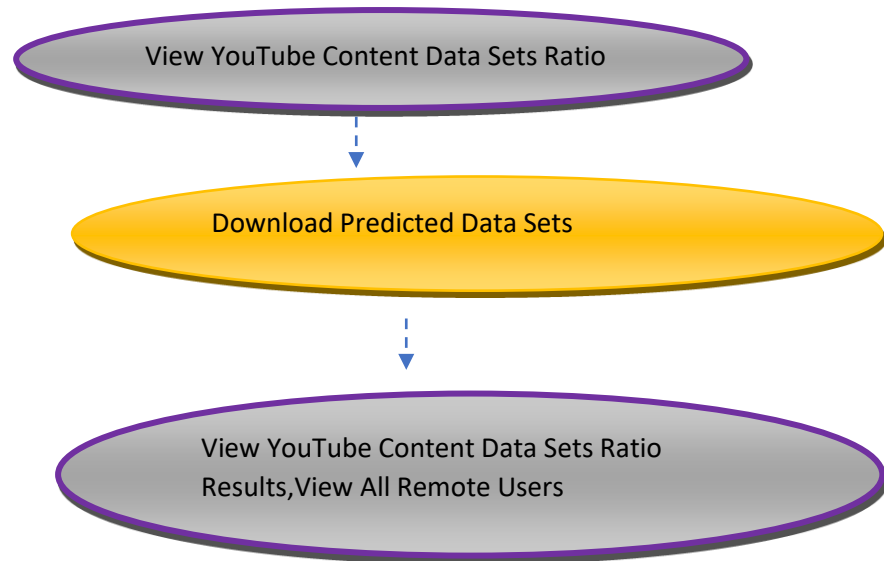
REGISTER AND LOGIN,

PREDICT YOUTTUBE CONTENT TYPE

VIEW YOUR PROFILE.

Use case





Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.

During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

4. CONCLUSION

In this paper, a novel deep learning-based framework is proposed for child inappropriate video content detection and classification. Transfer learning using EfficientNet-B7 architecture is employed to extract the features of videos. The extracted video features are

processed through the BiLSTM network, where the model learns the effective video representations and performs multiclass video classification. All evaluation experiments are performed by using a manually annotated cartoon video dataset of 111,156 video clips collected from YouTube. The evaluation results indicated that proposed framework of Efficient Net-BiLSTM (with hidden units $D = 128$) exhibits higher performance (accuracy 95.66%) than other experimented models including Efficient Net-FC, Efficient Net-SVM, Efficient Net-KNN, Efficient Net-Random Forest, and Efficient Net-BiLSTM with attention mechanism-based models (with hidden units $D = 64, 128, 256, \text{ and } 512$). Moreover, the performance comparison with existing state-of-the-art models also demonstrated that our BiLSTM-based framework surpassed other existing models and methods by achieving the highest recall score of 92.22%. The advantages of the proposed deep learning-based children inappropriate video content detection system are as follows:

- 1) It works by considering the real-time conditions by processing the video with a speed of 22 fps using EfficientNet-B7 and BiLSTM-based deep learning framework, which helps in filtering the live-captured videos.
- 2) It can assist any video sharing platform to either remove the video containing unsafe clips or blur/hide any portion with unsettling frames.
- 3) It may also help in the development of parental control solutions on the Internet through plugins or browser extensions where child unsafe content can be filtered automatically.

Furthermore, our methodology to detect inappropriate children content from YouTube is independent of YouTube video metadata which can easily be altered by malicious up loaders to deceive the audiences. In the future, we intend to combine the temporal stream using optical flow frames with the spatial stream of the RGB frames to further improve the model performance by better understanding the global representations of videos. We also aim to increase the classification labels to target the different types of inappropriate children content of YouTube videos.

5. REFERENCES

- [1] L. Ceci. *YouTube Usage Penetration in the United States 2020, by Age Group*. Accessed: Nov. 1, 2021. [Online]. Available: <https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/>

- [2] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. 10th ACM Conf. Recommender Syst.*, Sep. 2016, pp. 191_198, doi: 10.1145/2959100.2959190.
- [3] M. M. Neumann and C. Herodotou, "Evaluating YouTube videos for young children," *Educ. Inf. Technol.*, vol. 25, no. 5, pp. 4459_4475, Sep. 2020, doi: 10.1007/s10639-020-10183-7.
- [4] J. Marsh, L. Law, J. Lahmar, D. Yamada-Rice, B. Parry, and F. Scott, *Social Media, Television and Children*. Shef_eld, U.K.: Univ. Shef_eld, 2019. [Online]. Available: https://www.stac-study.org/downloads/STAC_Full_Report.pdf
- [5] L. Ceci. *YouTube Statistics & Facts*. Accessed: Sep. 01, 2021. [Online]. Available: <https://www.statista.com/topics/2019/youtube/>
- [6] M. M. Neumann and C. Herodotou, "Young children and YouTube: A global phenomenon," *Childhood Educ.*, vol. 96, no. 4, pp. 72_77, Jul. 2020, doi: 10.1080/00094056.2020.1796459.

DEFENSIVE MODELING OF FAKE NEWS THROUGH ONLINE SOCIAL NETWORKS

Mudunuri Lokesh Varma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract:-

Online social networks (OSNs) have become an integral mode of communication among people and even nonhuman scenarios can also be integrated into OSNs. The ever growing rise in the popularity of OSNs can be attributed to the rapid growth of Internet technology. OSN becomes the easiest way to broadcast media (news/content) over the Internet. In the wake of emerging technologies, there is dire need to develop methodologies, which can minimize the spread of fake messages or rumors that can harm society in any manner. In this article, a model is proposed to investigate the propagation of such messages currently coined as fake news. The proposed model describes how misinformation gets disseminated among groups with the influence of different misinformation refuting measures. With the onset of the novel coronavirus-19 pandemic, dubbed COVID-19, the propagation of fake news related to the pandemic is higher than ever. In this article, we aim to develop a model that will be able to detect and eliminate fake news from OSNs and help ease some OSN users stress regarding the pandemic. A system of differential equations is used to formulate the model. Its stability and equilibrium are also thoroughly analyzed. The basic reproduction number (R_0) is obtained which is a significant parameter for the analysis of message spreading in the OSNs. If the value of R_0 is less than one ($R_0 < 1$), then fake message spreading in the online network will not be prominent, otherwise if $R_0 > 1$ the rumor will persist in the OSN. Realworld trends of misinformation spreading in OSNs are discussed. In addition, the model discusses the controlling mechanism for untrusted message propagation. The proposed model has also been validated through extensive simulation and experimentation.

1. INTRODUCTION

IN THE 20th century, the Internet has become the most powerful tool for communication. It facilitates efficient and effective transfer of media from one location to another. With the development of Internet technology, social networks such as Facebook, WhatsApp, Twitter, Instagram, and Google plus have become a vital platform for information exchange [1]. Nowadays, people are connected through online social networks (OSNs) and exchange information in a cost efficient manner through data transfer. However, information exchanged on OSN platforms may comprise rumors that may affect the social lives of people [2]. Take COVID-19 as an example, where the proliferation of fake news related to the virus has left many people skeptical of any information they read information related to the virus on social media [3]. Some recent fake news related to a cure for COVID 19 has spread through Facebook [4].

Due to this type of misinformation, people from different corners of the world died. The impact of fake news on people related to a well-known Zika virus case study was presented by Sommariva et al. [5]. The authors found that the speed of fake news spread on OSNs is tremendous and tends to cover large audiences. One major challenge that is associated with OSNs is verification of messages exchanged as well as the authenticity of users.

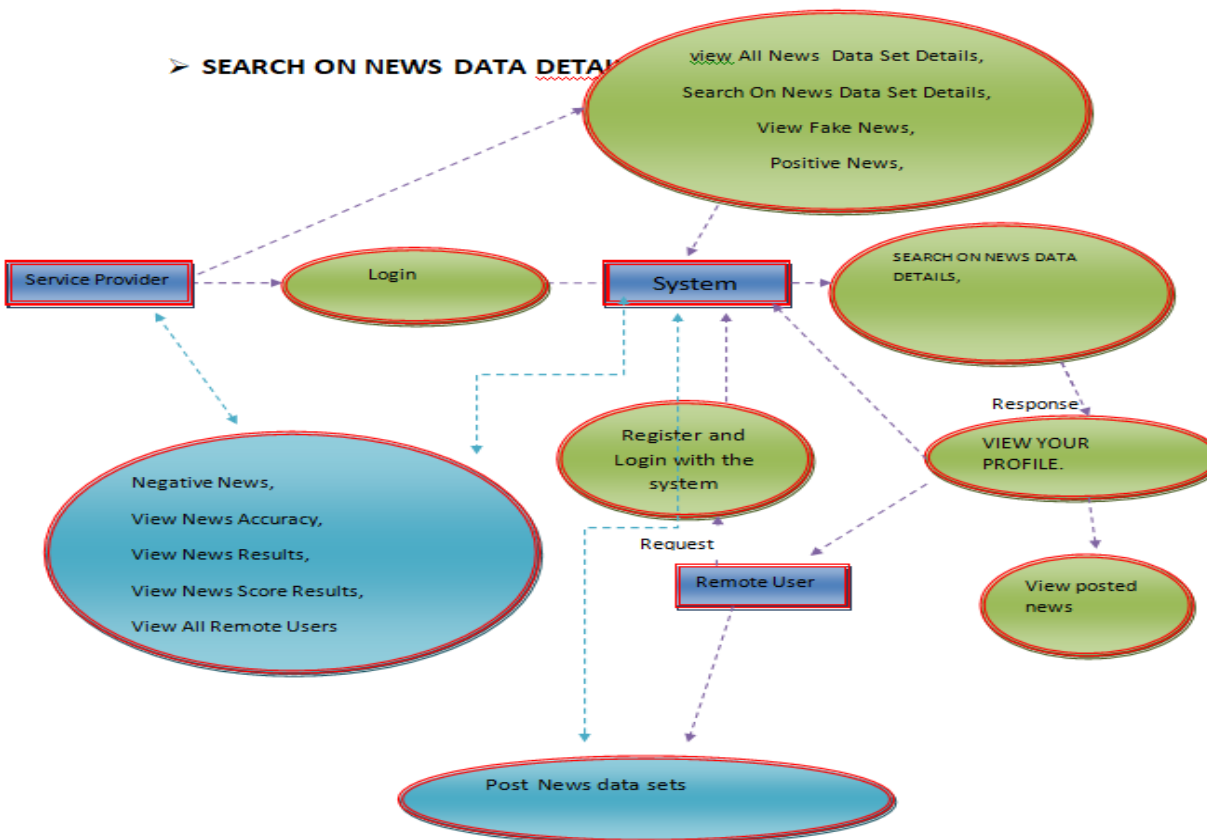
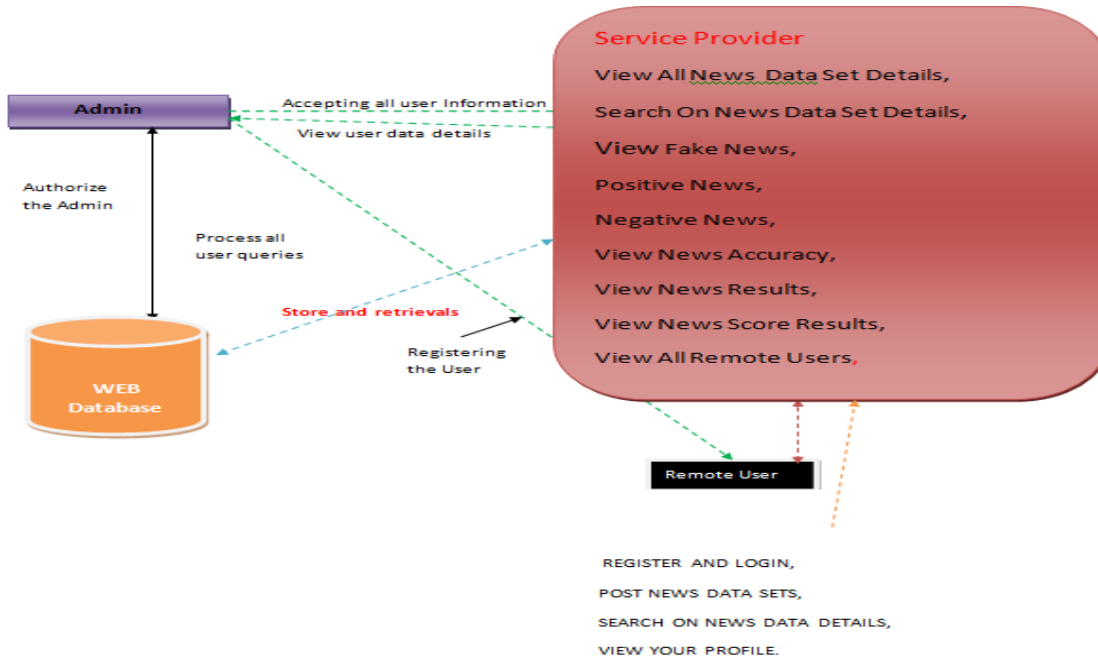
EXISTING SYSTEM

- ❖ The improved SIR model has been discussed by Zhang *et al.* [29] who considered the variable rate of infection and the resultant function for infected individuals and nonlinear Ordinary Differential Equation (ODE) is developed. This model also discusses the crowding effect on OSN and also derives an expression for the basic reproduction number. This model has been used for the analysis of rumor spreading dynamics in social network and predicts the spreading behavior of rumor. They discussed the control strategies of rumor spread in social networks.
- ❖ Zhu *et al.* [41] proposed an epidemic SIRS model, in which they described joining and leaving of users in OSNs. This article considers the dynamics of demography and the model is validated by simulation. More epidemic models are discussed related to rumors. Some of the researchers examined the temporal dynamics using the ODE [47]. Singh and Singh [48] discussed the spatial and temporal dynamics of rumor propagation and developed a strategy for counter measures using. They used partial differential equation for the study of rumor propagation dynamics in the social network. Huang and Su [44] proposed an epidemic model for the study of news propagation on OSN and also

suggested a method for controlling the rumor. They explained the effects of rumor spreading on OSN. For the study of rumor spreading in OSN, they evaluated the value of basic reproduction number and observed that if its value is less than one then the OSN will be free from unauthenticated news, otherwise unauthenticated news will be present in the OSN forever. The result of the proposed model has been verified by numerical calculation as well as simulation results.

- ❖ Dong *et al.* [49] analyzed the rumor spreading dynamics on OSN by SEIR epidemic model. They considered the varying user's number on OSN with time. The joining and deactivation rate of user in this model is discussed. They also found the basic reproduction number and exact equilibrium points of the model. The effect of user variation on rumor spreading in OSN is explained. They found that the new incoming users influence the rumor spreading rate in OSN. The proposed model is verified by simulation results.
- ❖ Furthermore, Zhu *et al.* [50] using the same model as in [49] obtained a local and global equilibrium as well as calculated the basic reproduction number using the next generation matrix concept. The authors explained the effect of time delay on rumor propagation and developed an effective control mechanism. A hesitating mechanism-based SEIR model is proposed by Liu *et al.* [51] for the study of rumor spreading in OSN. They used mean field theory for analysis of rumor spreading in OSN. They discussed the rumor-free equilibrium condition and global stability of the OSN and also obtained the value of basic reproduction number. They also analyzed the effects of feedback method on rumor spreading. They established the analysis feedback mechanism to reduce the rate of rumor spreading but were not able to reduce the value of basic reproduction number.

Architecture Diagram



3. SYSTEM STUDY

3.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement..

5. CONCLUSIONS

The research work presented in this article proposes a mathematical model to study the dynamic spreading and controlling activities of message transmission in OSNs. The proposed model employs differential equations for investigating the effect of verification and blocking of users and the spread of messages on OSNs. The expression for basic reproduction R_0 is obtained, which is used to analyze the status of rumor in the social network. Results obtained indicates that if R_0 is less than 1, then rumors and fake news will be eliminated and OSNs gets stabilized locally. The local stability of rumor free equilibrium is established by the Jacobian matrix. It is found that if the eigen values of the matrix are less than zero then the network will be

asymptotically stabilize locally in nature and free from the rumors. The Lyapunov function used to establish the global asymptotic stable status of the social network. Mathematical analysis has been performed to depict the accuracy of the rumor-free equilibrium. The activities of different classes of users have also been examined in the social network. In future, the method of latent and isolation can be used for the prevention of social network from rumor spread and fake news propagation. The issues examined in this article are of direct current concern, and the pandemic COVID-19 is creating a global crisis in rumors and fake news propagating freely on OSNs which may continue until it is cured/handled. Real world data clearly show that fake news propagation can be harmful for people, businesses, and many other facets of society. The results in this article therefore, may help solve some of the current global issues related to fake news spread.

6. REFERENCES

- [1] S. Wen, W. Zhou, J. Zhang, Y. Xiang, W. Zhou, and W. Jia, "Modeling propagation dynamics of social network worms," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 8, pp. 1633–1643, Aug. 2013.
- [2] E. Lebensztayn, F. P. Machado, and P. M. Rodríguez, "On the behaviour of a rumour process with random stifling," *Environ. Model. Softw.*, vol. 26, no. 4, pp. 517–522, Apr. 2011.
- [3] L. Li et al., "Characterizing the propagation of situational information in social media during COVID-19 epidemic: A case study on weibo," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 556–562, Apr. 2020.
- [4] A. Legon and A. Alsalman. How Facebook Can Flatten the Curve of the Coronavirus Infodemic. Accessed: Apr. 20, 2020. [Online]. Available: https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/
- [5] S. Sommariva, C. Vamos, A. Mantzarlis, L. U.-L. Dào, and D. Martinez Tyson, "Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study," *Amer. J. Health Edu.*, vol. 49, no. 4, pp. 246–255, Jul. 2018, doi: 10.1080/19325037.2018.1473178.
- [6] G. Whitehouse, "Pete/Repeat Tweet/Retweet Blog/reblog: A hoax reveals media mimicking," *J. Mass Media Ethics*, vol. 27, no. 1, pp. 57–59, Jan. 2012.
- [7] M. Kosfeld, "Rumours and markets," *J. Math. Econ.*, vol. 41, no. 6, pp. 646–664, Sep. 2005.

- [8] Y. Xiao, D. Chen, S. Wei, Q. Li, H. Wang, and M. Xu, “Rumor propagation dynamic model based on evolutionary game and antirumor,” *Nonlinear Dyn.*, vol. 95, no. 1, pp. 523–539, Jan. 2019.
- [9] A. V. Banerjee, “The economics of rumours,” *Rev. Econ. Stud.*, vol. 60, no. 2, pp. 309–327, Apr. 1993.
- [10] K. Dietz, “Epidemics and rumours: A survey,” *J. Roy. Stat. Soc., A (Gen.)*, vol. 130, no. 4, pp. 505–528, 1967.
- [11] D. J. Daley and D. G. Kendall, “Stochastic rumours,” *IMA J. Appl. Math.*, vol. 1, no. 1, pp. 42–55, 1965.
- [12] S. Dubey et al., “Psychosocial impact of covid-19,” *Diabetes Metabolic Syndrome*, vol. 14, no. 5, pp. 779–788, May 2020.
- [13] F. Ren, S.-P. Li, and C. Liu, “Information spreading on mobile communication networks: A new model that incorporates human behaviors,” *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 334–341, Mar. 2017.
- [14] T. Wang, J. He, and X. Wang, “An information spreading model based on online social networks,” *Phys. A, Stat. Mech. Appl.*, vol. 490, pp. 488–496, Jan. 2018.
- [15] S. Dagher, *Assad or We Burn Country: How One Family’s Lust for Power Destroyed Syria*. Boston, MA, USA: London Back Pay Books, 2019.



PREDICTING STOCK MARKET TRENDS USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS VIA CONTINUOUS AND BINARY DATA; A COMPARATIVE ANALYSIS

Muggalla Nandini (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

The nature of stock market movement has always been ambiguous for investors because of various influential factors. This study aims to significantly reduce the risk of trend prediction with machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals from Tehran stock exchange, are chosen for experimental evaluations. This study compares nine machine learning models (Decision Tree, Random Forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression and Artificial Neural Network (ANN)) and two powerful deep learning methods (Recurrent Neural Network (RNN) and Long short-term memory (LSTM)). Ten technical indicators from ten years of historical data are our input values, and two ways are supposed for employing them. Firstly, calculating the indicators by stock trading values as continuous data, and secondly converting indicators to binary data before using. Each prediction model is evaluated by three metrics based on the input ways. The evaluation results indicate that for the continuous data, RNN and LSTM outperform other prediction models with a considerable difference. Also, results show that in the binary data evaluation, those deep learning methods are the best; however, the difference becomes less because of the noticeable improvement of models' performance in the second way.

1. INTRODUCTION

The task of stock prediction has always been a challenging problem for statistics experts and finance. The main reason behind this prediction is buying stocks that are likely to increase in price and then selling stocks that are probably to fall. Generally, there are two ways for stock market prediction. Fundamental analysis is one of them and relies on a company's technique and fundamental information like mar-ket

position, expenses and annual growth rates. The second one is the technical analysis method, which concentrates on previous stock prices and values. This analysis uses historical charts and patterns to predict future prices [1], [2]. Stock markets were normally predicted by financial experts in the past time. However, data scientists have started solving prediction problems with the progress of learning techniques.



Also, computer scientists have begun using machine learning methods to improve the performance of prediction models and enhance the accuracy of predictions. Employing deep learning was the next phase in improving prediction models with better performance [3], [4]. Stock market prediction is full of challenges, and data scientists usually confront some problems when they try to develop a predictive model. Complexity and nonlinearity are two main challenges caused by the instability of stock market and the correlation between investment psychology and market behavior [5]. It is clear that there are always unpredictable factors such as the public image of companies or political situation of countries, which affect stock markets trend.

Therefore, if the data gained from stock values are efficiently preprocessed and suitable algorithms are employed, the trend of stock values and index can be predicted. In stock market prediction systems, machine learning and deep learning approaches can help investors and traders through their decisions. These methods intend to automatically recognize and learn patterns among big amounts of information. The algorithms can be effectively self-learning, and can tackle the predicting task of price punctuations in order to improve trading strategies [6].

2. EXISTING SYSTEM

- ❖ Stock market trends can be affected by external factors such as public sentiment and political events. The goal of this research

is to find whether or not public sentiment and political situation on a given day can affect stock market trends of individual companies or the overall market. For this purpose, the sentiment and situation features are used in a machine learning model to find the effect of public sentiment and political situation on the prediction accuracy of algorithms for 7 days in future. Besides, interdependencies among companies and stock markets are also studied. For the sake of experimentation, stock market historical data are downloaded from Yahoo! Finance and public sentiments are obtained from Twitter. Important political events data of Pakistan are crawled from Wikipedia.

- ❖ The raw text data are then pre-processed, and the sentiment and situation features are generated to create the final data sets. Ten machine learning algorithms are applied to the final data sets to predict the stock market future trend. The experimental results show that the sentiment feature improves the prediction accuracy of machine learning algorithms by 0–3%, and political situation feature improves the prediction accuracy of algorithms by about 20%. Furthermore, the sentiment attribute is most effective on day 7, while the political situation

attribute is most effective on day 5. SMO algorithm is found to show the best performance, while ASC and Bagging show poor performance. The interdependency results indicate that stock markets in the same industry show a medium positive correlation with each other.

PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- Request Clarification
- Feasibility Study
- Request Approval

REQUEST CLARIFICATION

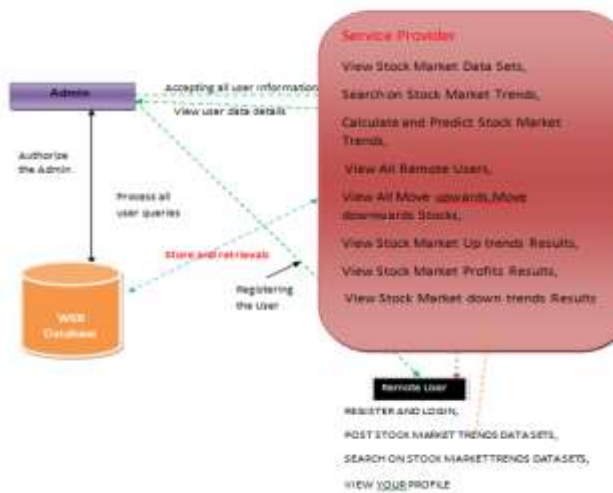
After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

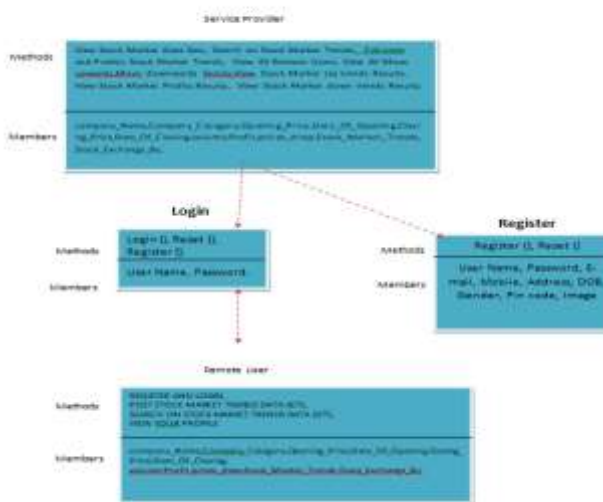
FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the

Architecture Diagram



Class Diagram





system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**
- **Operational Feasibility**

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser

connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations. This system has input screens in almost all the modules. Error messages are developed to alert the



user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4. CONCLUSIONS

The purpose of this study was the prediction task of stock market movement by machine learning and deep learning algorithms. Four stock market groups, namely diversified financials, petroleum, non-metallic minerals and basic metals, from Tehran stock exchange were chosen, and the dataset was based on ten years of historical records with ten technical features. Also, nine machine learning models (Decision Tree, Random Forest, Ada boost, XG Boost, SVC, Naïve Bayes, KNN, Logistic Regression and ANN) and two deep learning methods (RNN and LSTM) were employed as predictors. We supposed two approaches for input values to models, continuous data and binary data, and we employed three classification metrics for evaluations. Our experimental works showed that there was a significant improvement in the performance of models when they use binary data instead of



continuous one. Indeed, deep learning algorithms (RNN and LSTM) were our superior models in both approaches..

5. REFERENCES

[J. Murphy, Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. Penguin, 1999.

[2] T. Turner, A Beginner's Guide To Day Trading Online, 2nd ed. New York, NY, USA: Simon and Schuster, 2007.

[3] H. Maqsood, I. Mehmood, M. Maqsood, M. Yasir, S. Afzal, F. Aadil, M. M. Selim, and K. Muhammad, "A local and global event sentiment based efficient stock exchange forecasting using deep learning," *Int. J. Inf. Manage.*, vol. 50, pp. 432451, Feb. 2020.

[4] W. Long, Z. Lu, and L. Cui, "Deep learning-based feature engineering for stock price movement prediction," *Knowl.-Based Syst.*, vol. 164, pp. 163173, Jan. 2019.

[5] J. B. Duarte Duarte, L. H. Talero Sarmiento, and K. J. Sierra Juárez, "Evaluation of the effect of investor psychology on an artificial stock market through its degree of efficiency," *Contaduría y Administración*, vol. 62, no. 4, pp. 13611376, Oct. 2017.

[6] Lu, Ning, A Machine Learning Approach to Automated Trading. Boston, MA, USA: Boston College Computer Science Senior, 2016.

[7] M. R. Hassan, B. Nath, and M. Kirley, "A fusion model of HMM, ANN and GA for stock market forecasting," *Expert Syst. Appl.*, vol. 33, no. 1, pp. 171180, Jul. 2007.

[8] W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market

movement direction with support vector machine," *Comput. Oper. Res.*, vol. 32, no. 10, pp. 25132522, Oct. 2005.

[9] J. Sun and H. Li, "Financial distress prediction using support vector machines: Ensemble vs. Individual," *Appl. Soft Comput.*, vol. 12, no. 8, pp. 22542265, Aug. 2012.

[10] P. Ou and H. Wang, "Prediction of stock market index movement by ten data mining techniques," *Modern Appl. Sci.*, vol. 3, no. 12, pp. 2842, Nov. 2009.

[11] F. Liu and J. Wang, "Fluctuation prediction of stock market index by legendre neural network with random time strength function," *Neurocomputing*, vol. 83, pp. 1221, Apr. 2012.

[12] C.-F. Tsai, Y.-C. Lin, D. C. Yen, and Y.-M. Chen, "Predicting stock returns by classifier ensembles," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 24522459, Mar. 2011.

HATE CLASSIFY A SERVICE FRAMEWORK FOR HATE SPEECH IDENTIFICATION ON SOCIAL MEDIA

Mura Sushma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari
District, Andhra Pradesh, India, 534202.

V.Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

It is indeed a challenge for the existing machine learning approaches to segregate the hateful content from the one that is merely offensive. One prevalent reason for low accuracy of hate detection with the current methodologies is that these techniques treat hate classification as a multiclass problem. In this article, we present the hate identification on the social media as a multilabel problem. To this end, we propose a CNN-based service framework called “HateClassify” for labeling the social media contents as the hate speech, offensive, or nonoffensive. Results demonstrate that the multiclass classification accuracy for the CNN-based approaches particularly sequential CNN (SCNN) is competitive and even higher than certain state-of-the-art classifiers. Moreover, in the multilabel classification problem, sufficiently high performance is exhibited by the SCNN among other CNN-based techniques. The results have shown that using multilabel classification instead of multiclass classification, hate speech detection is increased up to 20%..

1. INTRODUCTION

Social media has emerged as a great platform to share feelings and emotions. However, the widespread acceptance of social media has also resulted in dissemination of hate content in the name of freedom of expression. The hate content on the social media has increased around 900% from year 2014 till year 2016 (<https://www.usatoday.com/story/news/2017/02/23/hate-groups-explode-social-media/98284662/>). According to a report, 73% of Internet users have seen online harassment and 40% personally experienced the online harassment (<https://www.pewresearch.org/internet/2014/10/22/onlineharassment/>). The term “hate speech” is defined by Council of Europe’s Protocol to the Convention on Cybercrime as the speech to “spread, incite, promote, or justify racial hatred, xenophobia, anti semitism or other forms of

hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants, and people of immigrant origin". However, under the free speech provisions of the first amendment, hate speech is protected in the United States. Online social media sites, such as Google, Facebook, and Twitter have their own policies for deciding "what is hate speech?" in their online social media. There exists a disagreement among the social media sites about dealing with the hate and offensive speech. Among, Google, Facebook, and Twitter, Twitter is the only one that does not ban hate speech at all. Twitter differentiates between the hate speech and direct specific threats. The twitter only considers hateful behavior of accounts whose primary purpose is to target others and their reported behavior is "one-sided". Although, Twitter claims that nobody is above their rules, it still faces criticism due to the vague nature of the company rules. As on May 31, 2016, Facebook, Twitter, Google's YouTube, and Microsoft have agreed to voluntary code of conduct to remove hate speech as defined by European Union. Most recently, the issue of hate speech on social media gained significant attention when the Facebook CEO was questioned about the company's policy about the flagging and identifying the hate speech or hateful content. The remarks of the company's CEO depict that the current approach being used by the Facebook for flagging the hateful content is not effective to deeply identify the emotions at varying levels of intensities. The reason is that there is difference in defining the hate speech content by different individuals. Several previous works, for example the work by Del Vigna et al.¹ considered the offensive and hate speech as one problem. However, Davidson et al.² differentiated hate speech from the offensive speech. The authors of the study argued that people often use highly offensive terms in their normal routines. Therefore, the problem of hate speech classification was presented as multiclass classification problem among the hate, offensive, and non offensive speech. We agree with the categorization of speeches provided by Davidson et al.² However, we consider the hate speech problem as multi label problem instead of multiclass problem. There is a very minute difference between offensive and hate speech and drawing a distinction between offensive and hate speech has confused human experts as well. Therefore, strictly labeling only one class can never resolve the conflicts between two arguing parties. Our results demonstrate that presenting the problem as multi label problem increases the accuracy in detecting offensive and hate speech. The proposed service framework called Hate Classify is a combination of a crowd-

source and machine learning techniques to detect the offensive and hate speech in online social media platforms. The main contributions of this article are as follows.

› We present a framework for detection of hate and offensive speech as a service for social media companies.

› Contrary to the social media platforms where the policies regarding hate speech are regulated by the specific organizations, the proposed framework employs a crowd-sourced approach for hate speech identification.

› The problem of hate speech detection is presented as multi label classification problem and sufficiently high classification accuracy is achieved.

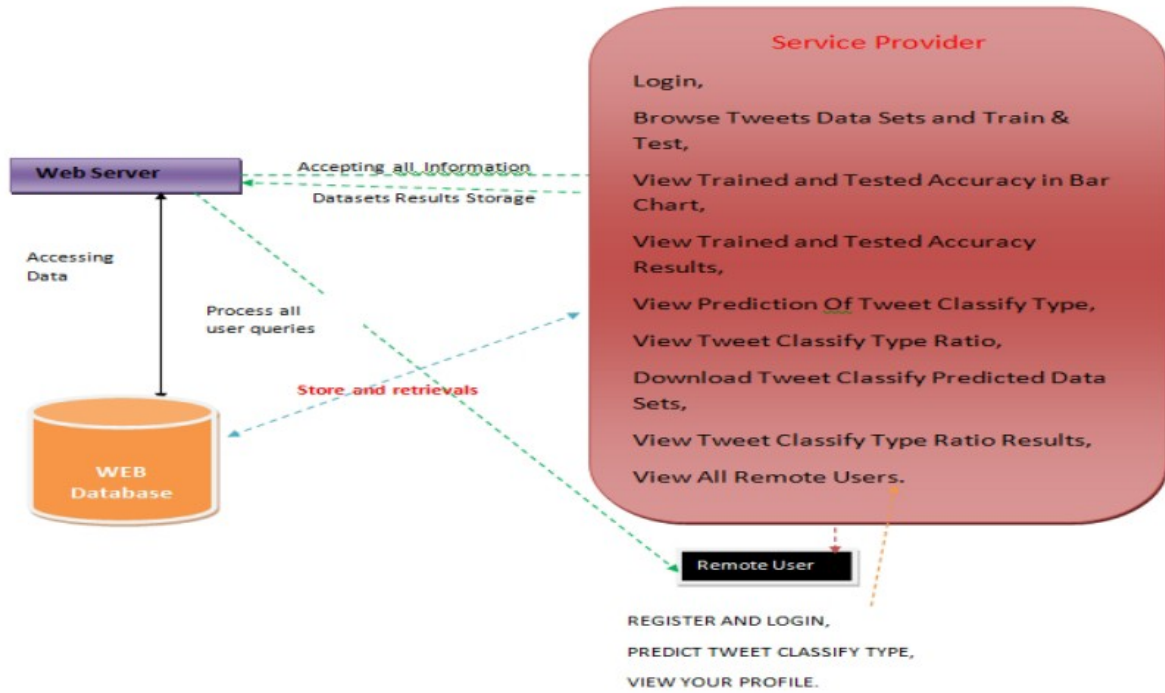
› The multi label classification used in Hate Classify framework yields 20% improvement in detection of hate speech on social media. The rest of this article is organized as follows. The “Related Work” section discusses the related work. The service framework is presented in the “Framework for Hate Speech Detection” section. The results of multiclass and multi label classification and comparisons with state-of-the-art techniques are presented in the “Experimental Results” section, whereas the “Conclusions” section finally concludes this article.

2. EXISTING SYSTEM

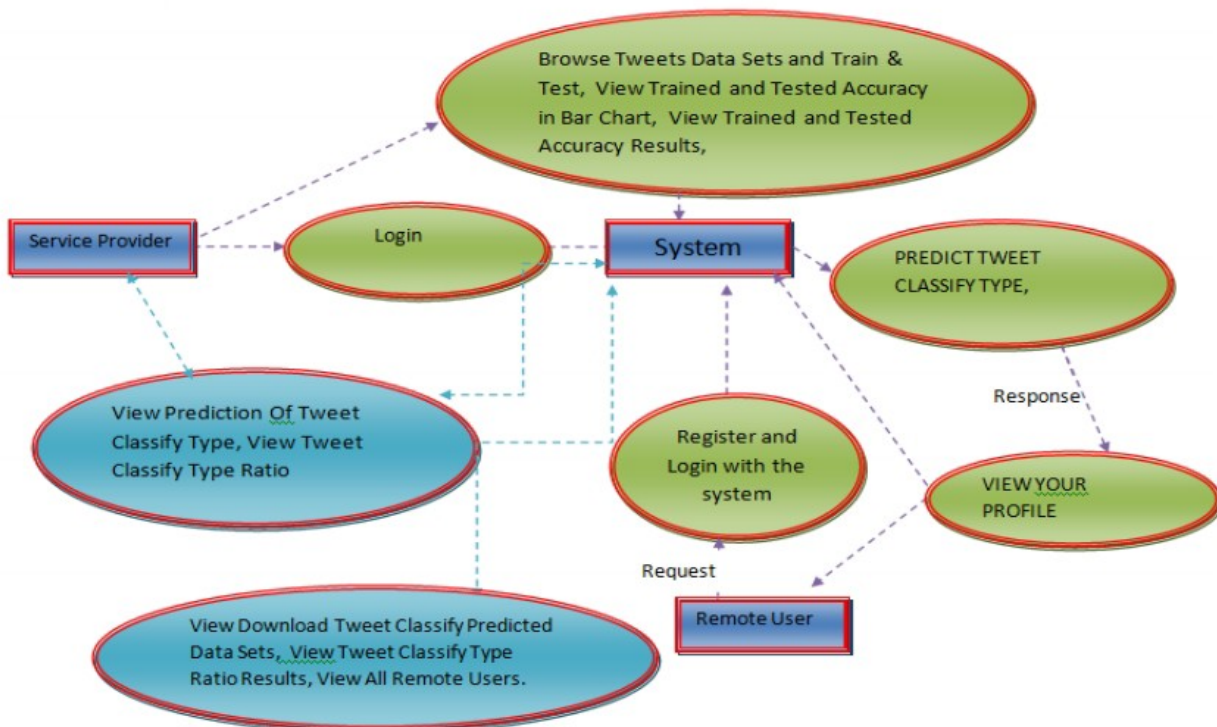
The work on the hate speech detection mostly revolves around finding the best features that can be used in text classification algorithms. The basic features that are used by most of the authors in their studies are n-grams and Bag-of-Words (BoW). Warner et al.³ argued that hatred against different groups can be categorized with the usage of small set of high frequency words. Chen et al.⁴ used n-grams with syntactic rules, such as user’s writing style. Hosseinmardi et al.⁵ used n-grams along with the number of comments for the images. Length of a tweet, geographical location, and gender information of the tweeting person were used along with the n-grams for hate speech detection by Waseem and Hovy.⁶ Finding the grammatical usage of hate content has also gained popularity among the researchers. Van Hee et al.⁷ used the sentiment features along with the n-grams and the BoW for studying and detecting hate speech. Xu et al.⁸ used n-grams with the Part-Of-Speech tagging (POS tagging) to study bullying traces on the social media. Davidson et al.² used TF-IDF weighted unigram, bigrams, trigrams, sentiment score of the tweet.

number of hashtags, retweets, URLs, characters, words, and syllables in each tweet as the feature set. To overcome the problem of sparsity due to short length of texts in tweets or online comments during hate detection, numerous researchers have utilized the concept of word generalization. Warner and Hirschberg³ used Brown Clustering technique for word generalization. Unlike Brown Clustering that assigns word to exactly one cluster, latent Dirichlet allocation (LDA) predict the probabilities of word in different clusters. Xiang et al.⁹ used the LDA for word generalization. Recently, several distributed word representations, termed as the word embedding have been developed for word generalizations. The word embedding takes the large text as the input and develops a vector space of words. The word vectors are placed in such a manner that words with similar context are placed closer to each other. Zhong et al.¹⁰ used word2vec (a word embedding technique) along with the BoW and hate effectiveness score to detect the hate speech. Paragraph2vec another word embedding technique was studied for hate speech detection against the BoW approach by Djuric et al. For classification, state vector machine^{12,3-5,7-9} and logistic regression (LR)^{2;6;9} have outperformed the other techniques for the hate speech detection studies. Nobata et al.¹³ preferred Vowpal Wabbit's regression model over other models. Mehdad and Tetreault¹⁴ have used recurrent neural network (RNN) models for hate speech detection.

Architecture Diagram



➤ **RData Flow Diagram :**



PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

3. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4. CONCLUSIONS

In this article, we presented a service framework called Hate Classify for hate speech detection on social media. The Hate Classify framework employs a crowd-sourced approach that permits the social media users to vote about any textual speech or content that is deemed inappropriate. To evaluate the performance in terms of classification, the CNNs were employed and experimental results demonstrate that the classification accuracy achieved through the CNN models, particularly the SCNN is significantly competitive and even better than several state-of-the-art approaches. An important contribution of this article is that it presents the problem of hate speech classification as the multi label classification problem. The experimental results attained by employing the CNN approaches both for the multiclass classification and multi label classification are sufficiently encouraging and signify the feasibility of these approaches for hate speech classification on social media.

5. REFERENCES

1. F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on Facebook," in Proc. 1st Italian Conf. Cybersecurity, 2017, pp. 86–95.
2. T. Davidson, D. Warmley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in Proc. 11th Int. AAAI Conf. Web Social Media, 2017, pp. 512–515.
3. W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in Proc. 2nd Workshop Lang. Social Media, 2012, pp. 19–26.
4. Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in Proc. IEEE Int. Conf. Privacy, Secur., Risk Trust Int. Conf. Soc. Comput., 2012, pp. 71–80.
5. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Analyzing labeled cyberbullying incidents on the instagram social network," Social Inform., T. Y. Liu, C. N. Scollon, and W. Zhu, Eds., 2015, pp. 49–66.

6. Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in Proc. NAACL Student Res. Workshop, 2016, pp. 88–93.
7. C. VanHee et al., "Detection and fine-grained classification of cyberbullying events," in Proc. Int. Conf. Recent Adv. Natural Lang. Process., 2015, pp. 672–680.
8. J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol., 2012, pp. 656–666.
9. G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale Twitter corpus," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1980–1984.
10. H. Zhong et al., "Content-driven detection of cyberbullying on the Instagram social network," in Proc. Int. Joint Conf. Artif. Intell., 2016, pp. 3952–3958.

STUDENTS PERFORMANCE PREDICTION IN ONLINE COURSES USING MACHINE LEARNING ALGORITHMS

Nandhyala Phanindhra Varma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Advances in Information and Communications Technology (ICT) have increased the growth of Massive open online courses (MOOCs) applied in distance learning environments. Various tools have been utilized to deliver interactive content including pictures, figures, and videos that can motivate the learners to build new cognitive skills. High ranking universities have adopted MOOCs as an efficient dashboard platform where learners from around the world can participate in such courses. The students learning progress is evaluated by using set computer marked assessments. In particular, the computer gives immediate feedback to the student once he or she completes the online assessments. The researchers claim that student success rate in an online course can be related to their performance at the previous session in addition to the level of engagement. Insufficient attention has been paid by literature to evaluate whether student performance and engagement in the prior assessments could affect student achievement in the next assessments. In this paper, two predictive models have been designed namely students' assessments grades and final students' performance. The models can be used to detect the factors that influence students' learning achievement in MOOCs. The result shows that both models gain feasible and accurate results. The lowest RSME gain by RF acquire a value of 8.131 for students assessments grades model while GBM yields the highest accuracy in final students' performance, an average value of 0.086 was achieved.

1. INTRODUCTION

Massive Open Online Courses (MOOCs) is one of the most widespread e-learning platforms. The MOOCs present the course using digital tool materials in various forms such as visual, audio, video and plain text. Most students prefer using video lectures to understand the contents of lessons over thoroughly reading plain text documents. The interactive video in the MOOCs could reduce students' stress, help them to feel relaxed and learn quickly [1] [2].MOOCs can be

classified into two distinct types mainly, connectivist Massive Open Online Courses (cMOOCs) and eXtended Massive Open Online Courses (xMOOCs). The xMOOCs are learning paradigm based on the principles of cognitivist behaviorist theory [4]. The structure of the courses is similar to the traditional course where the syllabus consists of a set of video lectures and a set of multiple choice quizzes in addition to the final exam. The video lectures featuring the course instructor reviewing the content of the previous online lesson are released weekly. The participants can watch and pause the video at their own pace. Moreover, the students can socially interact with other participants and the instructor through posting in discussion forums. The instructors usually post questions, provide task solutions and reply to student questions via these discussion forums; as a consequence the discussion forums play a vital role in enhancing the course quality and make online sessions collaborative and engaging [3] [5]. The cMOOCs are a new learning model based on connectivist learning theory [3][4]. With the connectivism approach, the instructor would not provide the actual learning material; the students get the course syllabus by asking the questions and sharing this information with other participants. References [3][4][5] posit the learning strategy of cMOOCs focused on a collaborative approach in which learning material combined remix, repurposable and provided, forwarded to other students. With cMOOCs, it is impossible to involve expertise to assess the students' knowledge whereas in xMOOCs, university lecturers can evaluate the students' knowledge through the use of computer-marked assessment feedback. In particular, the computer gives immediate feedback to the student when he completes the online assessment. The learner, upon successful completion, will be awarded their certification in xMOOCs. The cMOOCs do not include a formal assessment. Hence, universities are not considered cMOOCs as an official course [5][6]. With rapid advancements in technology, artificial intelligence has recently become an effective approach in the evaluation and testing of student performance in online courses. Many researchers applied machine learning to predict student performance in [7], however few works have been done to examine the trajectories performance [8]. As a result, educators could not monitor the real-time students learning curve. Two sets of experiments are conducted in this work. In the first set of experiments, regression analysis is implemented for estimation of students' assessment scores. The student past and current activities in addition to past performance are employed to predict student outcome. In the second set of experiments, supervised machine learning method has been utilized to predict long-term student performance.

Three types of candidate predictors have been considered firstly behavioral features, followed by temporal and demographic features. The proposed models offer new insight into determining the most critical learning activity and assist the educators in keeping tracking of timely student performance. To the best of our knowledge, student performance has been evaluated in online course using only two targets: “success” and “fail”. Our model predicts the performance with three-class labels “success”, “fail” and “withdrew”.

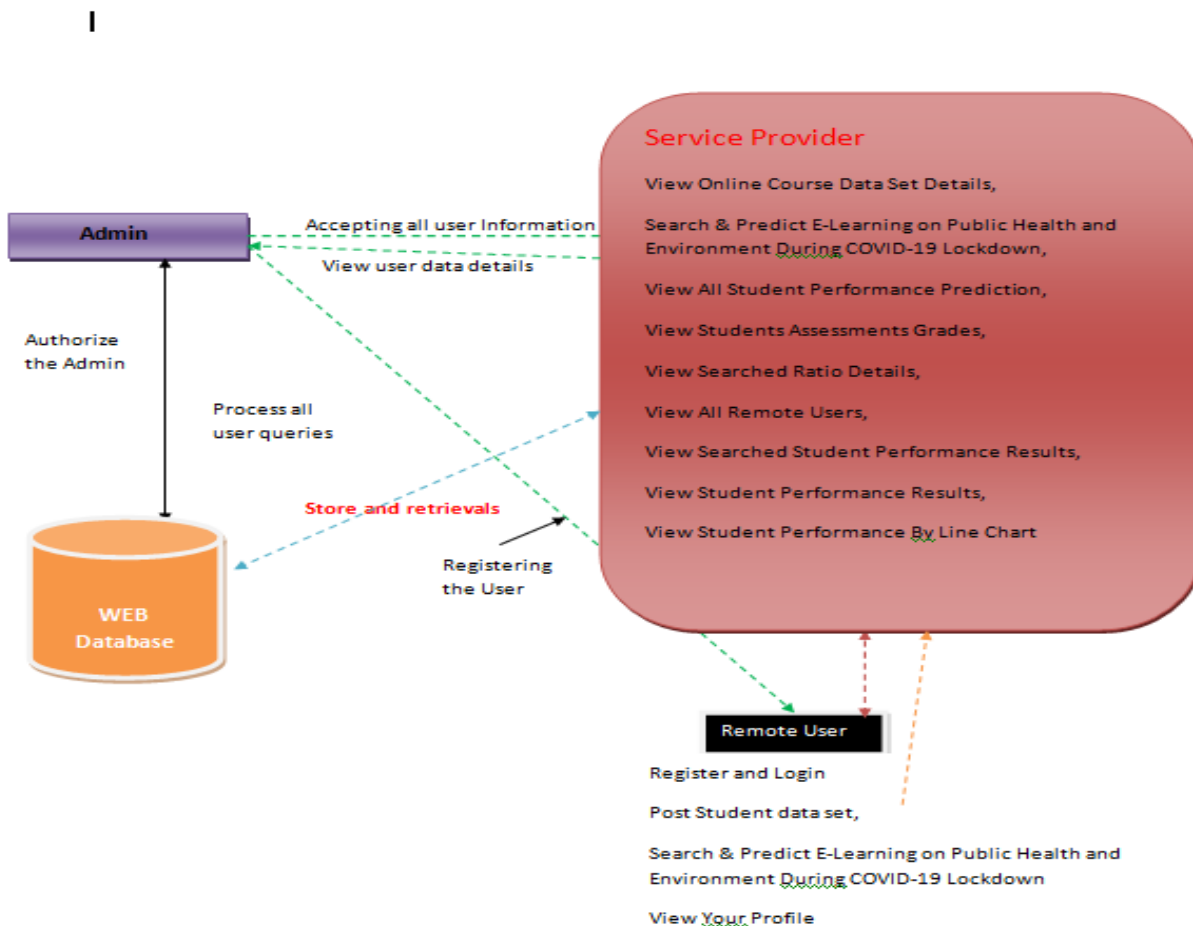
EXISTING SYSTEM

- ❖ The Factor Analysis Model (FAM) was proposed to predict the student's performance in Intelligent Tutoring System (ITS) taking into consideration the difficulty level of assessments based on Item Response Theory concept [9] [10]. The difficulty level of tasks can infer measurement of the correlation between the student's performances and assessment questions. To compute the probability of a student solving a task correctly, a set of predictor variables are defined in the FAM including the number of opportunities presented to the student at each task, the duration spent on each step and the difficulty level of each question or latent variable. The results reveal that incorporating the latent variables into the estimates of student performance can significantly enhance the model [10].
- ❖ To measure how the activities of learners could impact their learning achievement in MOOCs, the researchers found that Learning Analytics (LAs) in conjunction with machine learning, are effective tools that offer the potential to trace student knowledge. The researchers demonstrated that machine learning could help the educator in providing cohort information about the learning process, furnishing researchers with the ability to both visualise and analyse the information obtained from each tier of the learner. Thus, an accurate predictive model can be acquired in such courses[11] [12][13]. Students' marks in the first assessment and quiz scores in conjunction with social factors are used to predict students' final performance in online course [14].
- ❖ Two predictive models were introduced. In the first model, logistic regression was used to predict whether students gained a normal or distinction certificate. In the second predictive model, logistic regression was also used to predict if students achieved certification or not. The results indicated that the number of peer assessment is the most effective feature for acquiring a distinction. The average quiz scores were considered the

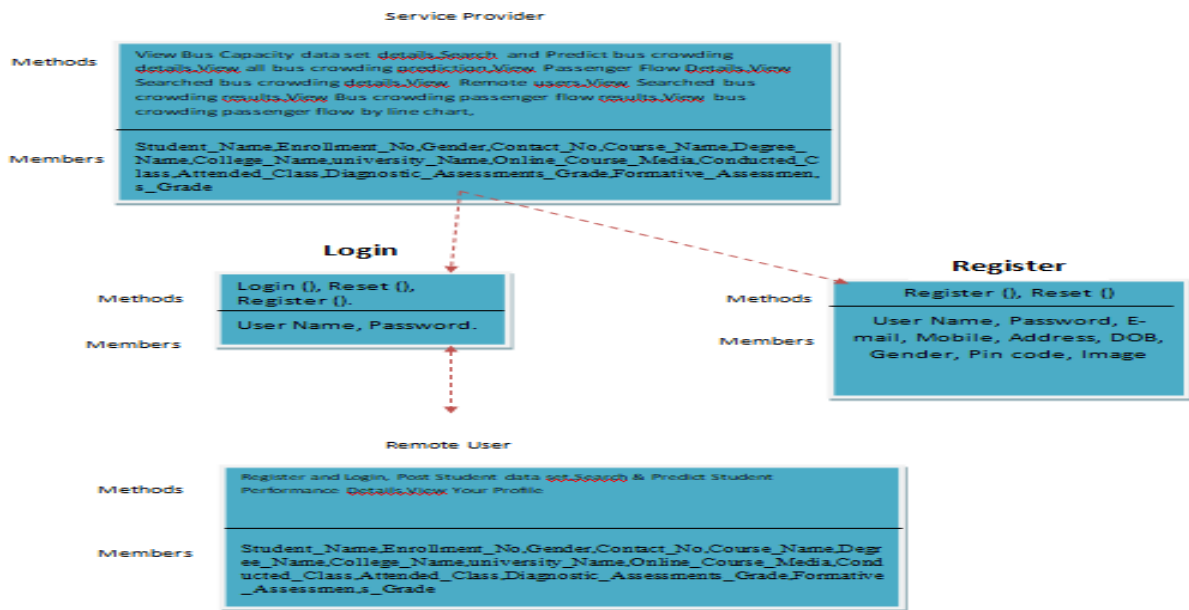
most reliable predictor for earning a certificate. The accuracy of distinction and normal models were reported with the percentage of 92.6% for the first model and 79.6 % for the second model, respectively[14].

- ❖ The association between the Virtual Learning Environment (VLE) data and student performance has been investigated at the University of Maryland, Baltimore County (UMBC) [12]. LA used through the implementation of the Check My Activity (CMA) tool. CMA can be defined as an LA tool, which compares students VLE activities with other activities and provides lecturers frequent feedback of students’ emotional states. The results showed the students who engage with the course frequently are more likely to earn mark C or higher than those who did not regularly engage [12].

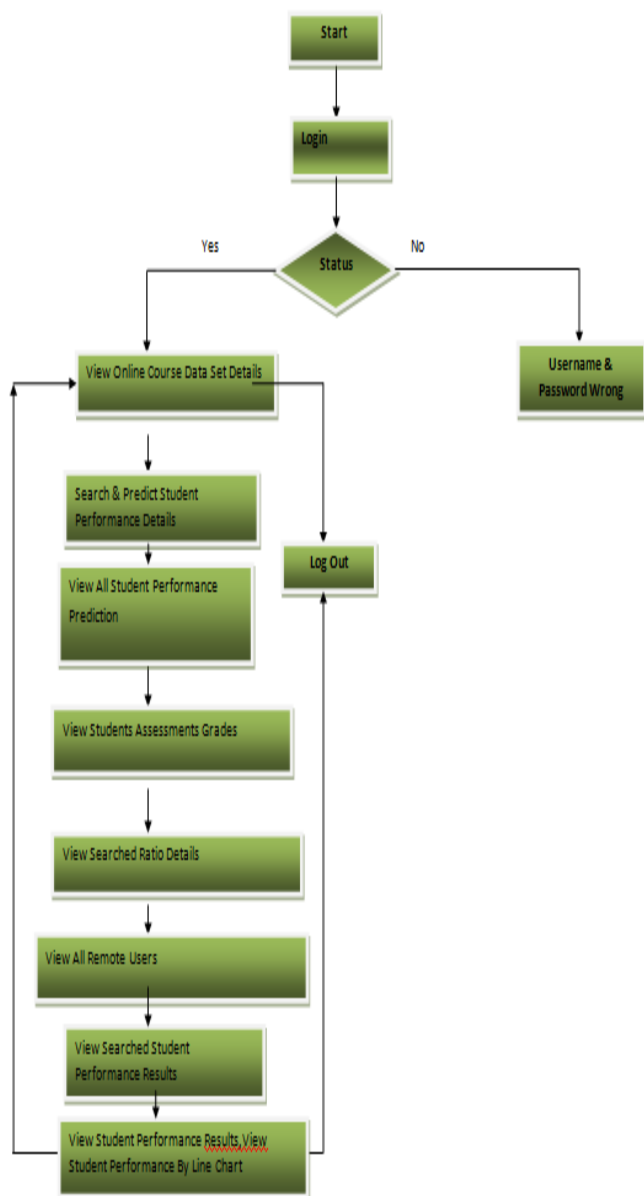
Architecture Diagram



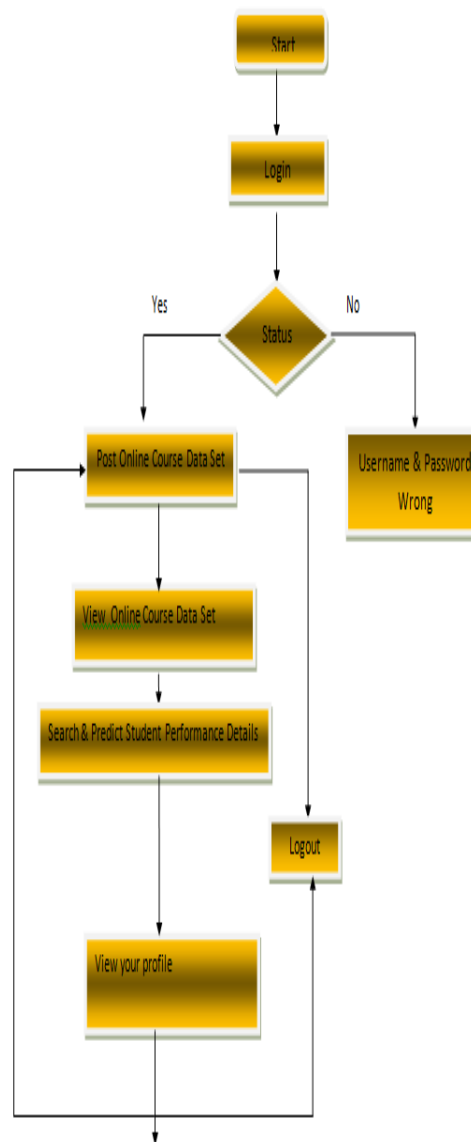
➤ **Class Diagram :**



➤ Flow Chart : Service Provider



➤ Flow Chart : Remote User



2. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins.

The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of

employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, it cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

3. CONCLUSIONS

Two sets of exterminates have been carried out in this study using regression and classification analysis. The results of predicting students' assessments grades model show that the students' performance in a particular assignment relies on students' mark in the previous assignment within single Courses. The researchers conclude that students' prior grade point average (GPA) with a low mark is considered as a significant factor of withdrawal from the next course in the traditional classroom setting, Both conventional classroom setting and virtual class share similar characteristic in term of the effective of pervious performance into student learning achievement in the future.

The final student performance predictive model revealed that student engagement with digital material has a significant impact on their success in the entire course. The findings' results also demonstrate that long-term students' performance achieves better accuracy than students' assessments grades prediction model, due to the exclusion of temporal features in regression analysis. The date of student deregistration from the course is a valuable predictor that is

significantly correlated with student performance. With the regression analysis, the data does not provide the last date of students' activity prior to undertaken assessments. The findings' results have been recommended to take into account the temporal features on predicting of subsequent assessments grades.

Future research direction involves the use of temporal features for predicting students' assessments grades model. With temporal feature time series analysis will be undertaken, might be more advanced machine learning will be utilized.

4. REFERENCES

- [1] K. F. Hew and W. S. Cheung, "Students' and instructors' use of massive open online courses (MOOCs): Motivations and challenges," *Educ. Res. Rev.*, vol. 12, pp. 45–58, 2014.
- [2] H. B. Shapiro, C. H. Lee, N. E. Wyman Roth, K. Li, M. Çetinkaya-Rundel, and D. A. Canelas, "Understanding the massive open online course (MOOC) student experience: An examination of attitudes, motivations, and barriers," *Comput. Educ.*, vol. 110, pp. 35–50, 2017.
- [3] J. Renz, F. Schwerer, and C. Meinel, "openSAP: Evaluating xMOOC Usage and Challenges for Scalable and Open Enterprise Education.," *Int. J. Adv. Corp. Learn.*, vol. 9, no. 2, pp. 34–39, 2016.
- [4] S. Li, Q. Tang, and Y. Zhang, "A Case Study on Learning Difficulties and Corresponding Supports for Learning in cMOOCs| Une étude de cas sur les difficultés d'apprentissage et le soutien correspondant pour l'apprentissage dans les cMOOC," *Can. J. Learn. Technol. Rev. Can. l'apprentissage la Technol.*, vol. 42, no. 2, 2016.
- [5] S. Zutshi, S. O'Hare, and A. Rodafinos, "Experiences in MOOCs: The Perspective of Students," *Am. J. Distance Educ.*, vol. 27, no. 4, pp. 218–227, 2013.
- [6] Z. Wang, T. Anderson, L. Chen, and E. Barbera, "Interaction pattern analysis in cMOOCs based on the connectivist interaction and engagement framework," *Br. J. Educ. Technol.*, vol. 48, no. 2, pp. 683–69 2017.
- [7] W. Xing and D. Du, "Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention," *J. Educ. Comput. Res.* p.0735633118757015., 2018.
- [8] M. J. Gallego Arrufat, V. Gamiz Sanchez, and E. Gutierrez Santiuste, "Trends in Assessment in Massive Open Online Courses," *Educ. Xx1*, vol. 18, no. 2, pp. 77–96, 2015.

- [9] B. J. Hao Cen, Kenneth Koedinger and Carnegie, Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement,” in In International Conference on Intelligent Tutoring Systems, 2006, vol. 8, pp. 164–175.
- [10] K. R. Koedinger, E. A. McLaughlin, and J. Stamper, “Automated Student Model Improvement,” in Educational Data Mining, proceedings of the 5th International Conference on, 2012, pp. 17–24.
- [11] J. Sinclair and S. Kalvala, “Student engagement in massive open online courses,” *Int. J. Learn. Technol.*, vol. 11, no. 3, pp. 218–237, 2016.
- [12] J. Mullan, “Learning Analytics in Higher Education,” London, 2016.
- [13] P. and K. Al-Shabandar, R., Hussain, A.J., Liatsis, “Detecting At-Risk Students With Early Interventions Using Machine Learning Techniques,” *IEEE Access*, vol. 7, pp. 149464–149478, 2019.
- [14] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, “Predicting MOOC Performance with Week 1 Behavior,” in Proceedings of the 7th International Conference on Educational Data Mining (EDM), 2014, pp. 273–275.
- [15] L. Analytics and C. Exchange, “OU Analyse : Analysing at - risk students at The Open University,” in in Conference, 5th International Learning Analytics and Knowledge (LAK) (ed.), 2015, no. October 2014.
- [16] R. Alshabandar, A. Hussain, R. Keight, A. Laws, and T. Baker, “The Application of Gaussian Mixture Models for the Identification of At-Risk Learners in Massive Open Online Courses,” in 2018 IEEE Congress on Evolutionary Computation, CEC 2018 - Proceedings, 2018.
- [17] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, “Identifying At-Risk Students for Early Interventions—A Time-Series Clustering Approach,” *IEEE Trans. Emerg. Top. Comput.*, vol. 5, no. 1, pp. 45–55, 2017.
- [18] C. Yun, D. Shin, H. Jo, J. Yang, and S. Kim, “An Experimental Study on Feature Subset Selection Methods,” 7th IEEE Int. Conf. Comput. Inf. Technol. (CIT 2007), pp. 77–82, 2007.



HOSPITAL MANAGEMENT SYSTEM WITH CHATBOT

Nitta Sree Himaja (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Healthcare is very important to lead a good life. However, it is very difficult to obtain the consultation with the doctor for every health problem. The idea is to create a medical chatbot using Artificial Intelligence that can diagnose the disease and provide basic details about the disease before consulting a doctor. This will help to reduce healthcare costs and improve accessibility to medical knowledge through medical chatbot. The chatbots are computer programs that use natural language to interact with users. The chatbot stores the data in the database to identify the sentence keywords and to make a query decision and answer the question. Ranking and sentence similarity calculation is performed using n-gram, TFIDF and cosine similarity. The score will be obtained for each sentence from the given input sentence and more similar sentences will be obtained for the query given. The third party, the expert program, handles the question presented to the bot that is not understood or is not present in the database.

Keywords:-Chatbot, Healthcare, Artificial Intelligence, Virtual Assistance, TFID, N-gram

1. INTRODUCTION

An AI chatbot is a computer program that simulates human communication. It is a piece of software that interacts with a human through written language. It is often embedded in web pages or other digital applications to answer customer inquiries without the need for human agents, thus providing affordable effortless customer service. Chatbots based on Machine Learning make an AI chatbot that is very capable of having an organic conversation with the user and answering their queries. Chatbots make use of the data given to them and using different training algorithms they can answer the queries in the best way possible. In our proposed system we create a conversational chatbot that is integrated into a hospital website. It is trained using

Machine Learning algorithms and acts as a very efficient interface between the user and the application. There is no predefined format for the users to ask their queries in, the chatbot manages to answer the query in the best possible way. Users have the flexibility to raise a query both in text and speech format. With this chatbot, users have access to hospital information, doctor availability, diagnostics, and other related data. They are navigated to different pages according to their requests which makes it easier and faster for them to explore. They can book appointments, and identify the problem by specifying symptoms to try to know about it beforehand, doing this they can take any required precautionary measures and book an appointment with the doctor as soon as possible.



2. LITERATURE REVIEW

A literature gives overview of previous related works in the current domain. With the Literature review, it can bring focus on area of research and broaden your knowledge of the domain. In the paper by Mamta Mittal [1], Algorithms such as Gradient descent method, Natural Language Processing (NLP), and feed-forward neural network (FNN) are used to create the chatbot. Gradient Descent (GD) is a cost-minimization technique that examines the coefficients of a function (f). It is a key optimisation approach for determining the minimal cost function. The model may be conveniently stored in memory with little noise using the GD technique. Computational linguistic rule-based human language modeling is combined with statistical, deep learning models and machine learning in NLP. These technologies work together to allow computers to analyze human language in the form of text or speech data and comprehend its entire meaning, including the speaker's or writer's purpose and mood. This chatbot answers questions about hospital information, such as specialist availability, OPD hours, room registration, bed capacity, doctor availability and emergency information, among other things. The suggested chatbot acts as if it were a genuine hospital receptionist, assisting users. It offers consumers complete medical support 24/7. In the paper by Rohit Binu Mathew [2], KNN (K-nearest neighbor algorithm) and NLP (Natural Language Processing) algorithms are used to create the chatbot. The K-Nearest neighbor method is one of

the Supervised Learning techniques and is one of the most popular Machine Learning algorithms. KNN works by classifying new data into most similar class label. The created chatbot application is an android application in which the user may tell the chatbot about their symptoms, and the chatbot will then tell them what health measures they should take. In the paper by Harsh Mendapara [3], the backend of the chatbot is written in Python, while the user interface is created using HTML, CSS, and JavaScript. Chatterbot, a natural processing library, is used to communicate between the user and the system. Text analysis is used to apply natural language processing. The application is hosted on a localhost server, which responds to user inquiries with relevant information. On the localhost server, the healthcare assistant's frontend interface is presented, and it is ready to address patient symptoms based on a certain ailment. The health assistant will collect certain personal information from the user at first, which will be saved in the database. User queries are entered into illnesses such as headaches, coughs, and colds. A separate data file is prepared for a doctor's appointment. The chatbot will next ask the user a question in which the user is expected to address health-related issues. If the patient has a high temperature, high bp (blood pressure) or low bp then the chatbot will prescribe the necessary medicine. In the paper by Siddhi Pardeshi [4], Long Short-Term Memory (LSTM), Natural Language Processing (NLP), Hybrid Emotion Inference Model (HEIM), Pattern Matching Algorithm and Naive Bayes Algorithm are



some of the chatbot design techniques covered. Natural language processing allows machines to take in input, break it down, retrieve its meaning, determine suitable action, and respond to users in natural language. Long Short-Term Memory (LSTM) is a Artificial Recurrent Neural Network (RNN). LSTMs are useful not just for processing single data inputs such as photos, but also for processing full sequences of data such as voice or video. The LSTM algorithm's primary tasks are handwriting identification and speech recognition. The most popularly used algorithm in chatbots is pattern matching algorithm. This Algorithm is basically a database that contains questions and corresponding answers. Patterns are like the names to the questions, whereas templates identify responses/answers. The response to this query is made up of Artificial Intelligence Mark-up Language (AIML) tags. Patterns (questions) and templates (answers) are stored in a tree structure. Questions are on the branches, and responses are at the nodes, thus anytime a user asks a question, the query is first searched for an answer term by term, and then the specific answer is fetched from the node. Another most popular algorithm used in chatbots is Naive Bayes. Tokenization comes first in this process, followed by stemming. Tokenization is the process of breaking down a phrase into individual words called tokens. The stems are then added to each token. For example, the sentence "it is a giant lion" is tokenized and then stemmed as "it", "is", "a", "giant" and "lion". The following step is to provide

training data. This information is saved in the form of lists or dictionaries, with class and sentence as properties in the dictionary. In the paper by Lekha Athota [5], N-gram, which is a series of N words, is used to construct the chatbot application. So, for example, "Final demo" is a 2-gram (a bigram), "This is a final demo" is a 4-gram, and "Good to go" is a 3-gram (trigram). The TF-IDF (Term Frequency-Inverse Document Frequency) which works by examining whether the word belongs to a document in a large collection of documents. This can be examined by multiplying two metrics: the word's inverse document frequency over a collection of documents and the number of times a word occurs in a document. It's used to get the keyword out of the user query. To get the best response for the inquiry, each term is weighted down. The Web-interface is designed for users to enter their query. The programme is enhanced with security and effectiveness modifications that ensure user protection and integrity when getting answers to queries. This chatbot assists users with basic health information. When a person initially visits the website, they must register before asking the questions to chatbot. If the answer is not available in the database, the system employs an expert system to respond to the queries. In the paper by Dammavalam Srinivasa Rao [6], an AI chatbot for college activities is developed using Deep Neural networks. The data regarding college activities is being collected in the JSON format and Bag of word technique is used in preprocessing of data. Gradient Descent is used for optimizing the model to process the patterns



and give best possible response to the question asked by the user. pytsx3 python library is used for speech recognition to enable users to give input questions using voice. The model accuracy is found to be around 93 percentage for 1200 epochs of training the model.

3. EXISTING SYSTEM

In all the existing systems, the scope is divided and they provide very few features at a time. Few chatbots only provide appointment booking functionality only and also may not include voice inputs from the users. Some bots provide disease diagnosis but can't provide medication and navigation through a complex hospital website. There are smaller number of chatbots integrated to hospital website that provides all the necessary contents and features.

4. PROPOSED SYSTEM

The proposed system focuses on integrating all basic features in one place in an application and powering it with an AI chat bot further adds new functionalities like easy navigation, access to data on doctors, diagnostics information, symptom analysis, precautionary or instant medication suggestions and appointment booking, all these in a single application. Further, considering people who cannot write fluently, those with special needs and those in emergency situations, both voice and text input formats are accepted by the chatbot. We are building the website using Flask, which contains Login and registration page, dashboard of website, appointment booking and viewing pages and also the animated chatbot button at the end of every page of the website. Speech enabled chatbots

provide higher level of interactivity and usability. User can either give their input using text or speech and similarly chatbot is able to give its response by either text or voice. In our project, this process of conversion between text and speech is done by using speech_recognition and pytsx3 python modules. a) Voice Input by User (Speech to Text): Using systems inbuilt microphone live audio input can be transcribed using Google's Web Speech API (recognize_google()).By using adjust for ambient noise function we can set the engine to listen to ambient noise for some time period(here 2 seconds) and adjust energy threshold accordingly. If speech Recognizer unable to detect the speech correctly, respective error messages will be given as response. b) Voice Output by Bot (Text to Speech): Pytsx3 is a Text to Speech Conversion Python Library. Using pytsx3.init() an engine instance will be created for which we can set various properties like voice rate, volume level and also voices (male or female). We can directly pass the text that need to be converted to voice to this engine and output will be voice saying the text accordingly. User gives a question to interact with the chatbot. After that, a model developed with LSTM analyses the user query. LSTMs are a kind of recurrent neural network that, as opposed to just passing its result into the following section of the network, plays out a progression of math tasks to work on its memory. There are four "gates" in an LSTM. They are forget gate, remember gate, learn gate, and output gate.



5. SYSTEM DESIGN

Doctor Patient Portal

In this project as per your requirements we have developed 3 modules

- 1) Admin module: admin can login to system by using username and password as 'admin' and 'admin'. After login admin can add Doctor Details and give login data to doctor. Admin can view doctor details, appointments, organ donor details, feedback and patient details
- 2) Patient Module: patient can signup with the application and then login to application. Patient can view their own details, book and cancel appointments, view appointments, donate or search organs and can give feedback
- 3) Doctor Module: doctor can login to system by using login details given by admin and then can view won profile, view appointments and generate prescription.

Each doctor will work in 3 hours slot and in 3 hour 6 patients can be taken and after 6 appointments system will display 'Already Booked' message.

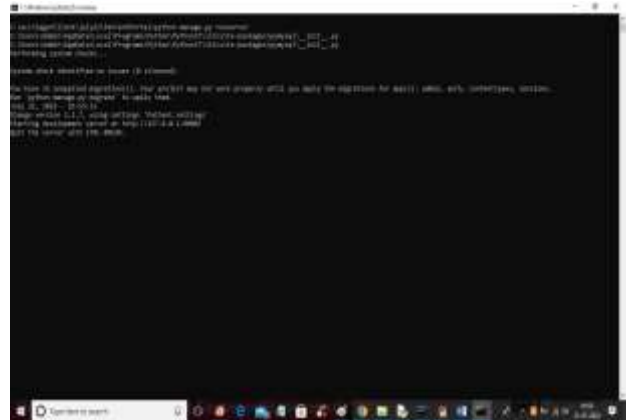
To run code follow below steps

Install python 3.7.0 and then open command prompt and install packages using below commands

```
pip install PyMySQL==0.9.3
```

```
pip install Django==2.1.7
```

Now double click on 'runServer.bat' to start DJANGO server and get below screen



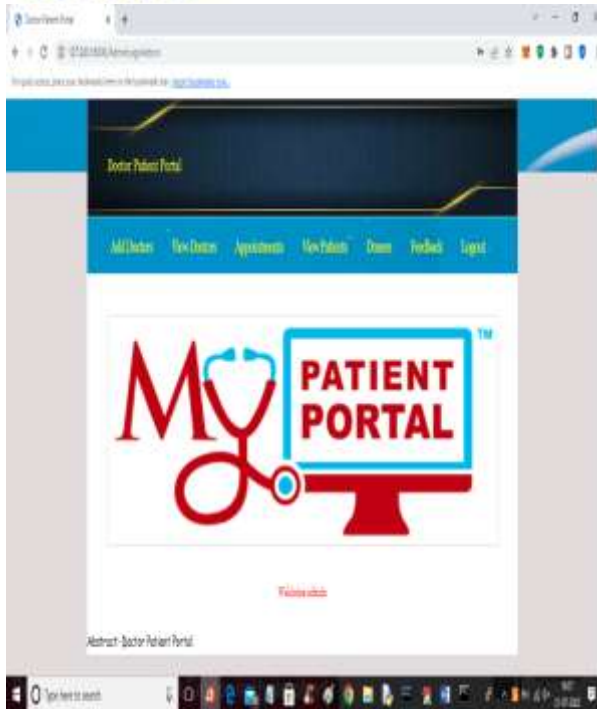
In above screen server is started and now open browser and enter URL as 'http://127.0.0.1:8000/index.html' and press enter key to get below screen



In above screen click on 'Admin' link to get below login screen



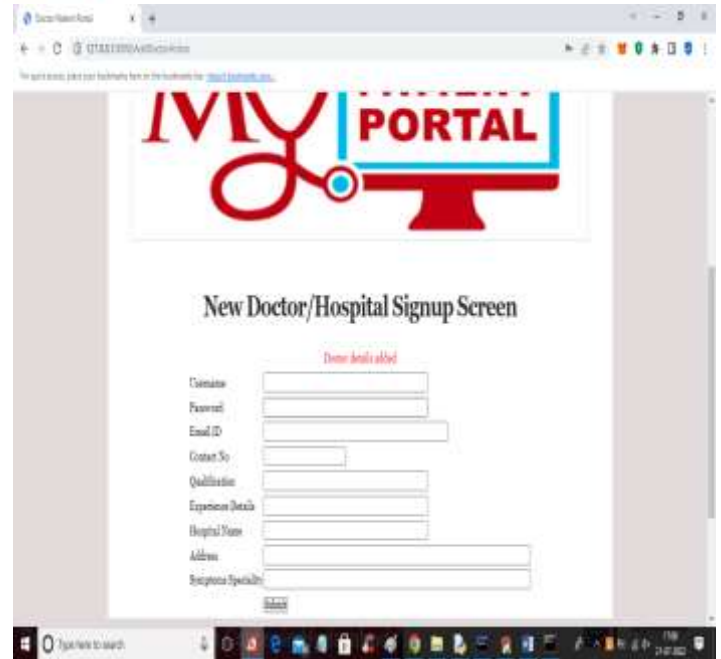
In above screen admin is login and after login will get below admin home screen



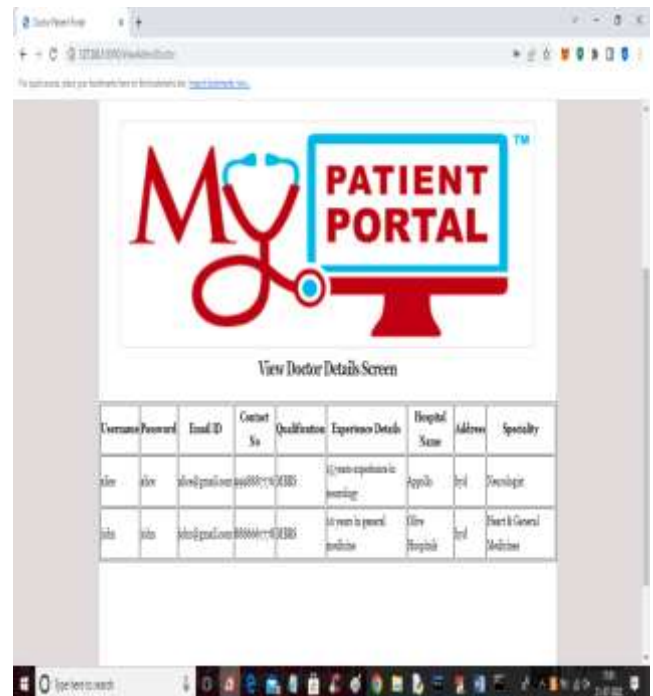
In above screen admin can click on 'Add Doctor' link to get below screen



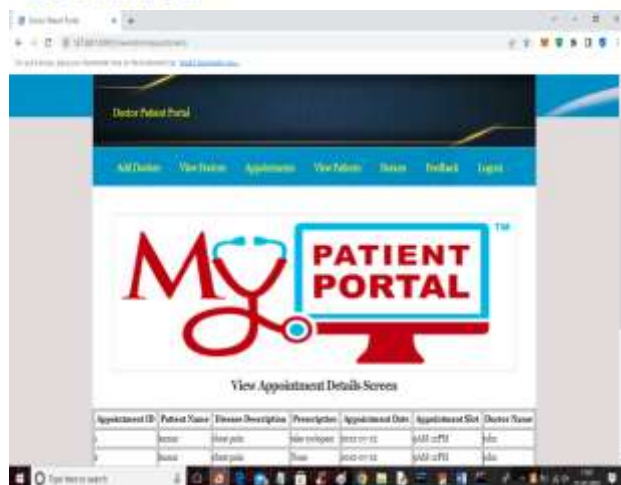
In above screen admin is adding doctor details and then press button to get below output



In above screen doctor details added and now admin can click on 'View Doctors' link to view all available doctor details



In above screen admin can see all doctor details and now click on 'Appointments' link to view all appointments



In above screen admin can see appointments details and if prescription not given by doctor yet then it will display message as 'None' else display given prescription. Now click on 'View Patients' link to view all patients details

6. CONCLUSION

The main objective of our hospital management system chatbot is to automate repeated tasks in a user-friendly manner such that it will provide hospital employees to focus on important tasks and also to enable fast response for customer instead of waiting for employee to solve their queries as user can interact with bot anytime. Enabling Speech recognition in our chatbot also helps customers to have a simple and fast conversation. The user interactive UI provides better navigation through the website. We have tested our application by trying various kinds of profiles. The results were satisfactory.

7. REFERENCES

[1] Mittal M, Battineni G, Singh D, Nagarwal T, Yadav P. Web-based chatbot for Frequently Asked Queries (FAQ) in

Hospitals. *J Taibah Univ Med Sc*2021;16(5):740e746.

[2] Mathew, Rohit Binu; Varghese, Sandra; Joy, Sera Elsa; Alex, Swanthana Susan (2019). IEEE 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) - Tirunelveli, India (2019.4.23- 2019.4.25)] 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI) - Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning, 851–856.

[3] Harsh Mendapara, Suhas Digole, Manthan Thakur, Anas Dange, AI Based Healthcare Chatbot System by Using Natural Language Processing, *International Journal of Scientific Research and Engineering Development*, Volume 4, Issue 2, Mar-Apr 2021.

[4] Siddhi Pardeshi, Suyasha Ovhal, Pranali Shinde, Manasi Bansode, Anandkumar Birajdar, A survey on Different Algorithms used in Chatbot, *International Research Journal of Engineering and Technology*; Volume: 07 Issue: 05, May 2020;

[5] Athota, Lekha; Shukla, Vinod Kumar; Pandey, Nitin; Rana, Ajay (2020). Chatbot for Healthcare System Using Artificial Intelligence, IEEE 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) - Noida, India (2020.6.4-2020.6.5)] 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) -, 619–622.

[6] Dammavalam Srinivasa Rao, K. Lakshman Srikanth, J. Noshitha Padma



Pratyusha, M. Sucharitha, M. Tejaswini and T. Ashwini, "Development of Artificial Intelligence Based Chatbot Using Deep Neural Network", In: Raju Pal and Praveen Kumar Shukla (eds), SCRS Conference Proceedings on Intelligent Systems, SCRS, India, 2021, pp. 143-151.

[7] Rashmi Dharwadkar, Neeta A. Deshpande "A Medical Chatbot", International Journal of Computer Trends and Technology, 2018

[8] Jitendra Chaudhary, Vaibhav Joshi, Atharv Khare, Rahul Gawali, Asmita Manna "A Comparative Study of Medical Chatbots", International Research Journal of Engineering and Technology (IRJET), 2021

[9] A. Ait-Mlouk and L. Jiang, "KBot: A Knowledge Graph Based ChatBot for Natural Language Understanding Over Linked Data," in IEEE Access, vol. 8, pp. 149220-149230, 2020.

[10] Rarhi, Krishnendu & Bhattacharya, Abhishek & Mishra, Abhishek & Mandal, Krishnasis. (2017). Automated Medical Chatbot. SSRN Electronic Journal.

COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING CRIME HOTSPOTS

Nookala Devika (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

Crime prediction is of great significance to the formulation of policing strategies and the implementation of crime prevention and control. Machine learning is the current mainstream prediction method. However, few studies have systematically compared different machine learning methods for crime prediction. This paper takes the historical data of public property crime from 2015 to 2018 from a section of a large coastal city in the southeast of China as research data to assess the predictive power between several machine learning algorithms. Results based on the historical crime data alone suggest that the LSTM model outperformed KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks. In addition, the built environment data of points of interests (POIs) and urban road network density are input into LSTM model as covariates. It is found that the model with built environment covariates has better prediction effect compared with the original model that is based on historical crime data alone. Therefore, future crime prediction should take advantage of both historical crime data and covariates associated with criminological theories. Not all machine learning algorithms are equally effective in crime prediction.

1. INTRODUCTION

Spatiotemporal data related to the public security have been growing at an exponential rate during the recent years. However, not all data have been effectively used to tackle real-world problems. In order to facilitate crime prevention, several scholars have developed models to predict crime [1]. Most used historical crime data alone to calibrate the predictive models. The research on crime prediction currently focuses on two major aspects: crime risk area prediction [2], [3] and crime hotspot prediction [4], [5].

The crime risk area prediction, based on the relevant influencing factors of criminal activities, refers to the correlation between criminal activities and physical environment, which both derived from the "routine activity theory" [6]. Traditional crime risk estimation methods usually detect crime hotspots from the historical distribution of crime cases, and assume that the pattern will persist in the following time periods [7]. For example, considering the proximity of crime places and the aggregation of crime elements, the terrain risk model tends to use crime-related environmental factors and crime history data, and is relatively effective for long-term, stable crime hotspot prediction [2].

Many studies have carried out empirical research on crime prediction in different time periods, combining demographic and economic statistics data, land use data, mobile phone data and crime history data. Crime hotspot prediction aims to predict the likely location of future crime events and hotspots where the future events would concentrate [8]. A commonly used method is kernel density estimation [9][12]. A model that considers temporal or spatial autocorrelations of past events performs better than those that fail to account for the autocorrelation [13]. Recently machine learning algorithms have gained popularity. The most popular methods include K-Nearest Neighbor(KNN), random forest algorithm, support vector machine (SVM), neural network and Bayesian model etc. [6]. Some compared the linear methods of crime trend prediction [14], some compared Bayesian model and BP neural network [15], [16], and others compared the spatiotemporal kernel density method with the random forest method in different periods of crime prediction [12].

Among these algorithms, KNN is an efficient supervised learning method algorithm [17], [18]. SVM is a popular machine learning model because it can not only implement classification and regression tasks, but also detect outliers [4], [19]. Random forest algorithm has been proven to have strong non-linear relational data processing ability and high prediction accuracy in multiple fields [20][23]. Naïve Bayes (NB) is a classical classification algorithm, which has only a few parameters and it is not sensitive to missing data [15], [24]. Convolutional neural networks (CNN) has strong expansibility, and can enhance its expression ability with a very deep layer to deal with more complex classification problems [25], [26].

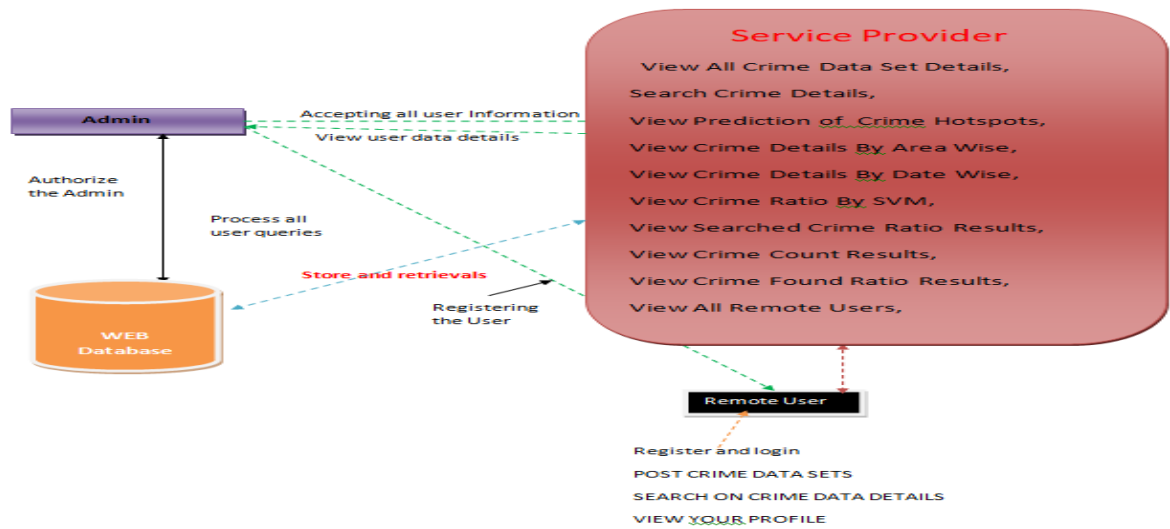
Long Short-Term Memory (LSTM) neural network extracts time-series features from features, and has a significant effect on processing data with strong time series trends [27][29]. This paper will focus on the comparison of the above six machine learning algorithms, and

recommend the best performing one to demonstrate the predictive power with and without the use of covariates.

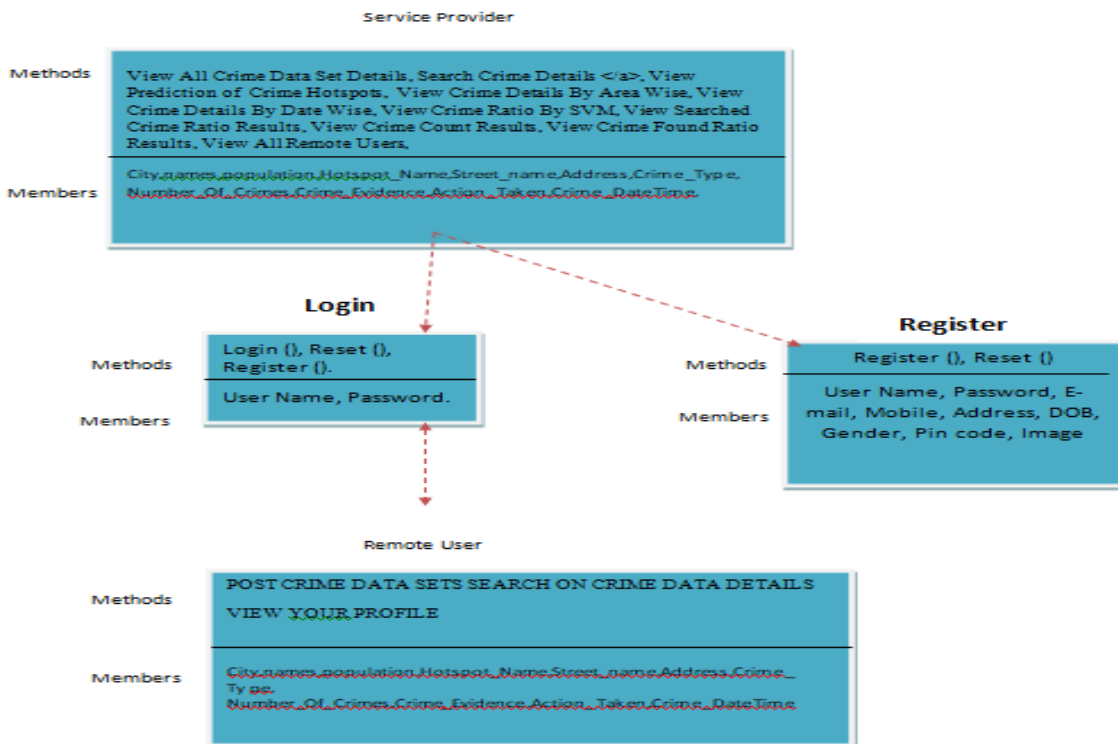
2. EXISTING SYSTEM

- ❖ Routine activity theory [30] was jointly proposed by Cohen and Felson in 1979, and has now been further developed through integration with other theories. This theory believes that the occurrence of most crimes, especially predatory crimes, needs the convergence of the three elements including motivated offenders, suitable targets, and lack of ability to defend in time and space.
- ❖ Rational choice theory [31] was proposed by Cornish and Clarke. The theory holds that the offender's choices in terms of location, goals, methods be explained by the rational balance of effort, risk and reward. Crime pattern theory [32] integrates the routine activities theory and the rational choice theory, which more closely explains the spatial distribution of criminal events. People form "cognitive map" and "activity space" through daily activities. At the same time, potential offenders also need to use their cognitive maps and choose specific locations for crimes in a relatively familiar space. When committing a crime, the offender tends to avoid those places they don't know but to choose the places where the "criminal opportunity overlaps with cognitive space" based on their rational ability. The reason why these places become crime hotspots is that they have the obvious characteristics of "producing" or "attracting" crime. Therefore, the environmental factors of the places need to be considered besides historical crime data for the prediction of crime hotspots.

Architecture Diagram



➤ **Class Diagram :**



PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some

entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4. CONCLUSIONS

In this paper, six machine learning algorithms are applied to predict the occurrence of crime hotspots in a town in the southeast coastal city of China. The following conclusions are drawn: 1) The prediction accuracies of LSTM model are better than those of the other models. It can better extract the pattern and regularity from historical crime data. 2) The addition of urban built environment covariates further improves the prediction accuracies of the LSTM model. The prediction results are better than those of the original model using historical crime data alone. Our models have improved prediction accuracies, compared with other models. In empirical research on the prediction of crime hotspots, Rummens et al. used historical crime data at a grid unit scale of 200 m²200 m, using three models of logistic regression, neural network, and the combination of logistic regression and neural network [41]. In the biweekly forecast, the highest case hit rate for the two robbery type is 31.97%, and the highest grid hit rate is 32.95%; Liu et al. Used the random forest model to predict the hot spots in multiple experiments in two weeks under the research scale of 150m²150m [23]. The average case hit rate of the model was 52.3%, and the average grid hit rate was 46.6%. The case hit rate of the LSTM model used in this paper was 59.9%, and the average grid hit rate was 57.6%, which was improved compared with the previous research results, For the future research, there are still some aspects to be improved. The first is the temporal resolution of the prediction. Felson et al. revealed that the crime level changes with time [43] Some studies have shown that it is useful to check the variation of risks during the day [44]. We chose two weeks as the prediction window. It does not capture the impact of crime changes within a week, let alone the change within a day. The sparsity of data makes the prediction of crime event difficult if the prediction window is narrowed down to day of a week or hour within a day. There is no viable solution to this challenging problem at this time. The second is the spatial resolution of the grid. In this paper, the grid size is 150m²150m. Future research will assess the impact of changing grid sizes on prediction accuracy. Third, the robustness and generality of the findings of this paper needs to be tested in other study areas. Nonetheless, the findings of this research have proven to be useful in a recent hotspot crime prevention experiment by the local police department at the study size.

5. REFERENCES

- [1] U. Thongsatapornwatana, "A survey of data mining techniques for analyzing crime patterns," in Proc. 2nd Asian Conf. Defence Technol. (ACDT), Jan. 2016, pp. 123128.

- [2] J. M. Caplan, L. W. Kennedy, and J. Miller, "Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting," *Justice Quart.*, vol. 28, no. 2, pp. 360381, Apr. 2011.
- [3] M. Cahill and G. Mulligan, "Using geographically weighted regression to explore local crime patterns," *Social Sci. Comput. Rev.*, vol. 25, no. 2, pp. 174193, May 2007.
- [4] A. Almehmadi, Z. Joudaki, and R. Jalali, "Language usage on Twitter predicts crime rates," in *Proc. 10th Int. Conf. Secur. Inf. Netw. (SIN)*, 2017, pp. 307310.
- [5] H. Berestycki and J.-P. Nadal, "Self-organised critical hot spots of criminal activity," *Eur. J. Appl. Math.*, vol. 21, nos. 45, pp. 371399, Oct. 2010.
- [6] K. C. Baumgartner, S. Ferrari, and C. G. Salfati, "Bayesian network modeling of offender behavior for criminal proling," in *Proc. 44th IEEE Conf. Decis. Control, Eur. Control Conf. (CDC-ECC)*, Dec. 2005, pp. 27022709.
- [7] W. Gorr and R. Harries, "Introduction to crime forecasting," *Int. J. Fore- casting*, vol. 19, no. 4, pp. 551555, Oct. 2003.
- [8] W. H. Li, L. Wen, and Y. B. Chen, "Application of improved GA-BP neural network model in property crime prediction," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 42, no. 8, pp. 11101116, 2017.
- [9] R. Haining, "Mapping and analysing crime data: Lessons from research and practice," *Int. J. Geogr. Inf. Sci.*, vol. 16, no. 5, pp. 203507, 2002.
- [10] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Secur. J.*, vol. 21, nos. 12, pp. 428, Feb. 2008.
- [11] S. Chainey and J. Ratcliffe, "GIS and crime mapping," *Soc. Sci. ComputRev.*, vol. 25, no. 2, pp. 279282, 2005.
- [12] L. Lin, W. J. Liu, and W. W. Liao, "Comparison of random forest algorithm and space-time kernel density mapping for crime hotspot prediction," *Prog. Geogr.*, vol. 37, no. 6, pp. 761771, 2018.
- [13] C. L. X. Liu, S. H. Zhou, and C. Jiang, "Spatial heterogeneity of microspatial factors' effects on street robberies: A case study of DP Peninsula," *Geograph. Res.*, vol. 36, no. 12, pp. 24922504, 2017.
- [14] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255260, Jul. 2015.

- [15] X. Zhao and J. Tang, "Modeling temporal-spatial correlations for crime prediction," in Proc. Int. Conf. Inf. Knowl. Manag. Proc., vol. F1318, 2017, pp. 497506.
- [16] A. Babakura, M. N. Sulaiman, and M. A. Yusuf, "Improved method of classification algorithms for crime prediction," in Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST), 2015, pp. 250255.
- [17] Q. Zhang, P. Yuan, Q. Zhou, and Z. Yang, "Mixed spatial-temporal characteristics based crime hot spots prediction," in Proc. IEEE 20th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD), May 2016, pp. 97101.

FOUREYE DEFENSIVE DECEPTION AGAINST ADVANCED PERSISTENT THREATS VIA HYPERGAME THEORY

Pala Lakshmi Prasanna (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Defensive deception techniques have emerged as a promising proactive defense mechanism to mislead an attacker and thereby achieve attack failure. However, most game-theoretic defensive deception approaches have assumed that players maintain consistent views under uncertainty. They do not consider players' possible, subjective beliefs formed due to asymmetric information given to them. In this work, we formulate a hypergame between an attacker and a defender where they can interpret the same game differently and accordingly choose their best strategy based on their respective beliefs. This gives a chance for defensive deception strategies to manipulate an attacker's belief, which is the key to the attacker's decision making. We consider advanced persistent threat (APT) attacks, which perform multiple attacks in the stages of the cyber kill chain where both the attacker and the defender aim to select optimal strategies based on their beliefs. Through extensive simulation experiments, we demonstrated how effectively the defender can leverage defensive deception techniques while dealing with multi-staged APT attacks in a hypergame in which the imperfect information is reflected based on perceived uncertainty, cost, and expected utilities of both attacker and defender, the system lifetime (i.e., mean time to security failure), and improved false positive rates in detecting attackers

1. INTRODUCTION

The key purpose of a defensive deception technique is to mislead an attacker's view and make it choose a suboptimal or poor action for the attack failure [33]. When both the attacker and defender are constrained in their resources, strategic interactions can be the key to

beat an opponent. In this sense, non-game-theoretic defense approaches have inherent limitations due to lack of efficient and effective strategic tactics. Forms of deception techniques have been discussed based on certain classifications, such as hiding the truth vs. providing false information or passive vs. active for increasing attackers' ambiguity or confusion [3, 9].

Game theory has been substantially used for dynamic decision making under uncertainty, assuming that players have consistent views. However, this assumption fails as players may often subjectively process asymmetric information available to them [22]. Hyper game theory [5] is a variant of game theory that provides a form of analysis considering each player's subjective belief, misbelief, and perceived uncertainty and accordingly their effect on decision making in choosing a best strategy [22]. This paper leverages hyper game theory to resolve conflicts of views of multiple players as a robust decision making mechanism under uncertainty where the players may have different beliefs towards the same game. Hyper game theory models players, such as attackers and defenders in cyber security to deal with advanced persistent threat (APT) attacks. We dub this effort Foureye after the Foureye butterfly fish, demonstrating deceptive defense in nature [40].

To be specific, we identify the following nontrivial challenges in obtaining a solution. First of all, it is not trivial to derive realistic game scenarios and develop defensive deception techniques to deal with APT attacks beyond the reconnaissance stage. This aspect has not been explored in the state-of-the-art. Second, quantifying the degree of uncertainty in the views of attackers and defenders is challenging, although they are critical because how each player frames a game significantly affects its strategies to take. Third, given a number of possible choices under dynamic situations, dealing with a large number of solution spaces is not trivial whereas the deployment and maintenance of defensive deception techniques is costly in contested environments. We partly addressed these challenges in our prior work in [12]; however, its contribution is very limited in considering a small-scale network and a small set of strategies with a highly simplified probability model developed using Stochastic Petri Network.

To be specific, this paper has the following **new key contributions**:

_ We modeled an attack-defense game under uncertainty based on hypergame theory where an attacker and a defender have different views of the situation and are uncertain about strategies taken by their opponents.

_ We reduced a player's action space by using a sub game determined based on a set of strategies available where each sub game is formulated based on each stage of the cyber kill chain (CKC) based on a player's belief under uncertainty.

_ We considered multiple defense strategies, including defensive deception techniques whose performance can be significantly affected by an attacker's belief and perceived uncertainty, which impacts its choice of a strategy.

_ We modeled an attacker's and a defender's uncertainty towards its opponent (i.e., the defender and the attacker, respectively) based on how long each player has monitored the opponent and its chosen strategy. To the best of our knowledge, prior research on hyper game theory uses a predefined constant probability to represent a player's uncertainty. In this work, we estimated the player's uncertainty

based on the dynamic, strategic interactions between an attacker and a defender.

_ We conducted comparative performance analysis with or without a defender using defensive deception (DD) strategies and with or without perfect knowledge available towards actions taken by the opponent. We measured the effectiveness and efficiency of DD techniques in terms of a system's security and performance, such as perceived uncertainty, hyper game expected utility, action cost, mean time to security failure (MTTSF or system lifetime), and improved false positive rate (FPR) of an intrusion detection by the DD strategies taken by the defender.

2. EXISTING SYSTEM

Garg and Grosu [15] proposed a game-theoretic deception framework in honeynets with imperfect information to find optimal actions of an attacker and a defender and investigated the mixed strategy equilibrium. Carroll and Grosu [10] used deception in attacker-defender interactions in a signaling game based on perfect Bayesian equilibria and hybrid equilibria. They considered defensive deception techniques, such as honeypots, camouflaged systems, or normal systems. Yin et al. [41] considered a Stackelberg attack-defense game where both players make decisions based on their perceived observations and identified an optimal level of deceptive protection using fake resources.

Casey et al. [11] examined how to discover Sybil attacks based on an evolutionary signaling game where a defender can use a fake identity to lure the attacker to facilitate cooperation. Schlenker et al. [32] studied a sophisticated and naïve APT attacker in the reconnaissance stage to identify an optimal defensive deception strategy in a zero-sum Stackelberg game by solving a mixed integer linear program.

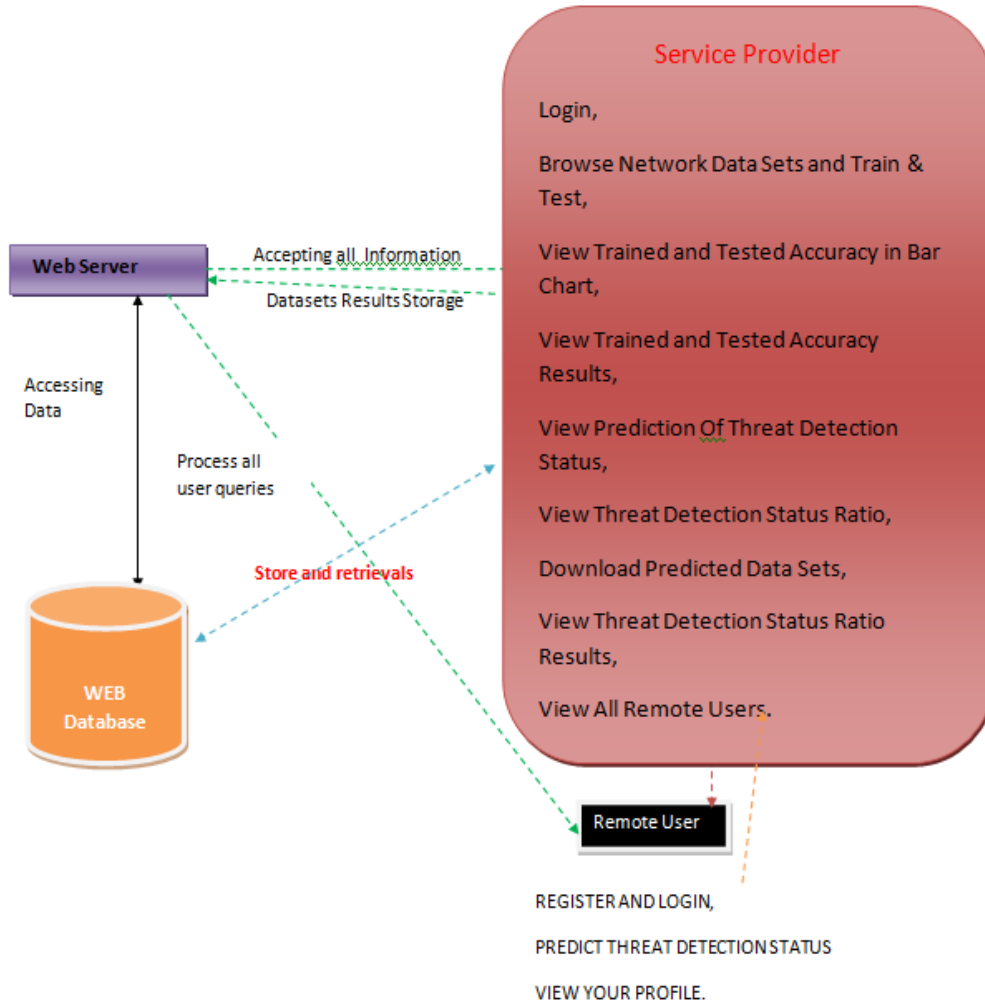
Unlike the above works cited [10, 11, 15, 32, 41], our work used hypergame theory which offers the powerful capability to model uncertainty, different views, and bounded rationality by different players. This way reflects more realistic scenarios between the attacker and defender.

Hypergame theory has emerged to better reflect realworld scenarios by capturing players' subjective and imperfect belief, aiming to mislead them to adopt uncertain or non-optimized strategies. Although other game theories deal with uncertainty by considering probabilities that a certain event may happen, they assume that all players play the same game [34]. Hypergame theory has been used to solve decision-making problems in military and adversarial environments House and Cybenko [20], Vane [37], Vane and Lehner [39]. Several studies [16, 17] investigated how players' beliefs evolve based on hypergame theory by developing a misbelief function measuring the differences between a player's belief and the ground truth payoff of other players' strategies. Kanazawa et al. [21] studied an individual's belief in an evolutionary hypergame and how this belief can be modelled by interpreter functions. Sasaki [31] discussed the concept of subjective rationalizability where an agent believes that its action is a best response to the other agent's choices based on its perceived game.

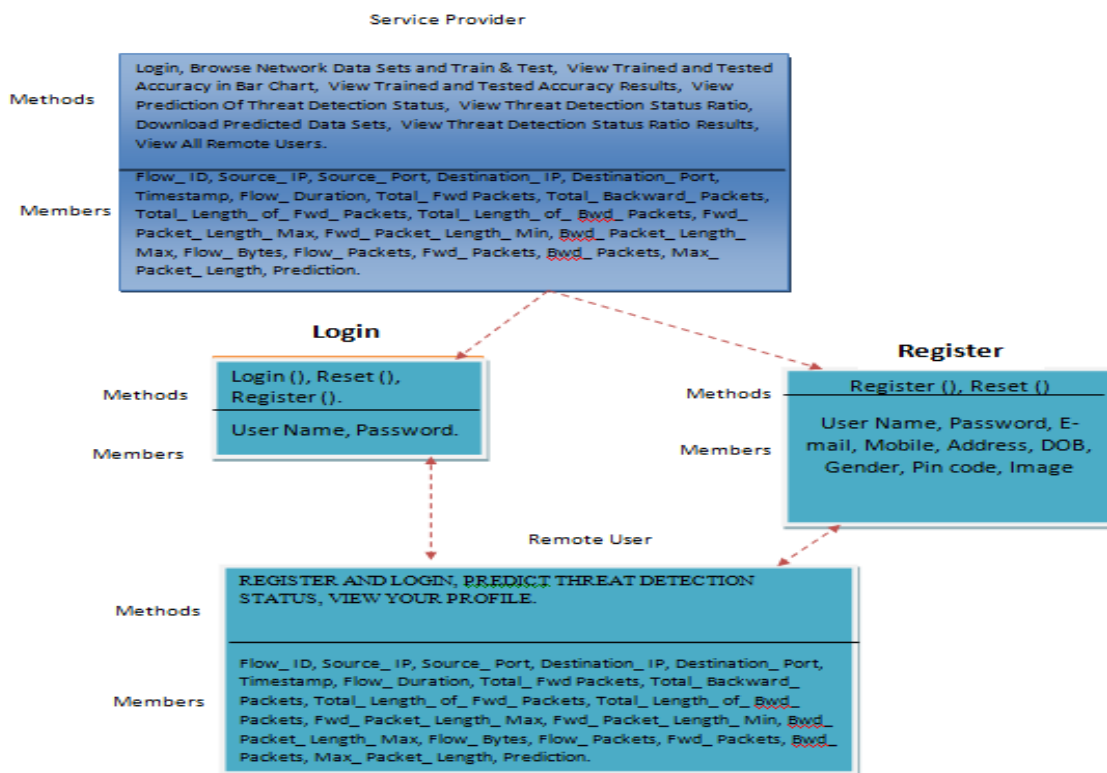
Putro et al. [30] proposed an adaptive, genetic learning algorithm to derive optimal strategies by players in a hypergame. Ferguson-Walter et al. [13] studied the placement of decoys based on a hypergame. This work developed a game tree and investigated an optimal move for both an attacker and defender in an adaptive game. Aljefri et al. [2] studied a first level hypergame involving misbeliefs to resolve conflicts for two and then more decision makers. Bakker et al. [4]

modeled a repeated hypergame in dynamistochastic setting against APT attacks primarily in cyberphysicalsystems.

Architecture Diagram



➤ **Class Diagram :**



3. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the

Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4. CONCLUSION

From this study, we obtained the following key findings:

_ An attacker's and defender's perceived uncertainty can be reduced when defensive deception (DD) is used. This is because the attacker perceives more knowledge about the system as it performs attacks as an inside attacker. On the other hand, the defender's uncertainty can be reduced by collecting more attack intelligence by using DD while allowing the attacker to be in the system.

_ Attack cost and defense cost are two critical factors in determining HEUs (hyper game expected utilities). Therefore, high DHEU (defender's HEU) is not necessarily related to high system performance in MTTSF (mean time to security failure) or TPR (true positive rate) which can also be a key indicator of system security. Therefore, using DD under imperfect information (IPI) yields the best performance in MTTSF (i.e., the longest system lifetime) while it gives the minimum DHEU among all schemes.

_ DD can effectively increase TPR of the NIDS in the system based on the attack intelligence collected through the DD strategies.

This work bring up some important directions for future research by: (1) considering multiple attackers arriving in a system simultaneously in order to consider more realistic scenarios; (2) estimating each player's belief based on machine learning in order to more correctly predict a next move of its opponent; (3) dynamically adjusting a risk threshold, i.e., Eq. (6), depending on a system's security state; (4) introducing a recovery mechanism to restore a compromised node to a healthy node allowing the recovery delay; (5) developing an intrusion

response system that can reassess a detected intrusion in order to minimize false positives while identifying an optimal response strategy to deal with intrusions with high urgency; and (6) considering another intrusion prevention mechanism, such as moving target defense, as one of the defense strategies.

5. REFERENCES

- [1] “Common vulnerability scoring system (CVSS).” [Online]. Available: <https://www.first.org/cvss/>
- [2] Y. M. Aljefri, M. A. Bashar, L. Fang, and k. W. Hipel, “First-level hypergame for investigating misperception in conflicts,” *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 48, no. 12, pp. 2158–2175, 2017.
- [3] H. Almeshekah and H. Spafford, “Cyber security deception,” in *Cyber Deception*. Springer, 2016, pp. 25–52.
- [4] C. Bakker, A. Bhattacharya, S. Chatterjee, and D. L. Vrabie, “Learning and information manipulation: Repeated hypergames for cyber-physical security,” *IEEE Control Systems Letters*, vol. 4, no. 2, pp. 295–300, 2019.
- [5] P. G. Bennett, “Toward a theory of hypergames,” *Omega*, vol. 5, no. 6, pp. 749–751, 1977.
- [6] E. Bertino and N. Islam, “Botnets and Internet of Things security,” *Computer*, vol. 50, no. 2, pp. 76–79, Feb. 2017.
- [7] M. Boussard, D. T. Bui, L. Ciavaglia, R. Douville, M. L. Pallec, N. L. Sauze, L. Noirie, S. Papillon, P. Peloso, and F. Santoro, “Software-defined LANs for interconnected smart environment,” in *2015 27th Int’l Teletraffic Congress*, Sep. 2015, pp. 219–227.
- [8] U. Brandes, “A faster algorithm for betweenness centrality,” *Jour. mathematical sociology*, vol. 25, no. 2, pp. 163–177, 2001.
- [9] J. W. Caddell, “Deception 101-primer on deception,” DTIC Document, Tech. Rep., 2004.
- [10] T. E. Carroll and D. Grosu, “A game theoretic investigation of deception in network security,” *Security and Communication Networks*, vol. 4, no. 10, pp. 1162–1172, 2011.



ADAPTIVE DIFFUSION OF SENSITIVE INFORMATION IN ONLINE SOCIAL NETWORKS

Pala Gayathri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

The cascading of sensitive information such as private contents and rumors is a severe issue in online social networks. One approach for limiting the cascading of sensitive information is constraining the diffusion among social network users. However, the diffusion constraining measures limit the diffusion of non-sensitive information diffusion as well, resulting in the bad user experiences. To tackle this issue, in this paper, we study the problem of how to minimize the sensitive information diffusion while preserve the diffusion of non-sensitive information, and formulate it as a constrained minimization problem where we characterize the intention of preserving non-sensitive information diffusion as the constraint. We study the problem of interest over the fully-known network with known diffusion abilities of all users and the semi-known network where diffusion abilities of partial users remain unknown in advance. By modeling the sensitive information diffusion size as the reward of a bandit, we utilize the bandit framework to jointly design the solutions with polynomial complexity in the both scenarios. Moreover, the unknown diffusion abilities over the semi-known network induce it difficult to quantify the information diffusion size in algorithm design. For this issue, we propose to learn the unknown diffusion abilities from the diffusion process in real time and then adaptively conduct the diffusion constraining measures based on the learned diffusion abilities, relying on the bandit framework. Extensive experiments on real and synthetic datasets demonstrate that our solutions can effectively constrain the sensitive information diffusion, and enjoy a 40% less diffusion loss of non-sensitive information comparing with four baseline algorithms.

1. INTRODUCTION

The prevalence of online social networks such as Facebook, Twitter and Wechat facilitates the information diffusion among users, and thus enables the efficient promotion of positive informations, e.g., products, news, innovations [1]- [8]. Although such efficient diffusion can easily lead to large-scale diffusion called information cascading, the unconstrained cascading behavior could meanwhile cause the sensitive information to be incautiously diffused over the network [9]- [20]. Here the sensitive information refers to any kind of information that needs to be prohibited from cascading such as rumors, personal contents, and trade secrets. The cascading of such sensitive information may cause the risk of leaking users' privacies or arising panics among publics [9]- [20]. With this concern, several social network medias (e.g., Facebook, Twitter) have claimed authorities to block account of users and delete some posts or tweets when they violate relevant rules about privacies or securities [9] [21] [22].



Thus network managers are able to take measures to prohibit the cascading of sensitive information. The existing attempts that share the closest correlation with prohibiting sensitive information diffusion belong to the rumor influence minimization [9]- [20], whose current strategies can mainly be classified into two aspects. The first is diffusing the truths over network to counteract rumors [12]- [14]. However, diffusing truths is only suitable for constraining the rumors, while is not suitable for constraining the diffusion of the other kinds of sensitive informations, including personal informations, trade secrets, and etc. The second is temporarily blocking a number of users with high diffusion abilities [9] [10] [15] [16] or blocking a number of social links among users [17]- [20] in hope of minimizing the diffusion of a rumor. Although such strategy is effective for preventing rumors about some significant events like earthquakes, terrorist attacks and political elections, it is unrealistic for network managers to adopt this strategy on constraining the diffusion of sensitive informations with various contents that widely exist in our daily lives. If network managers take such measure, it is required to block a much larger size of users or links. Then two critical problems arise.

Firstly, blocking too many users or social links will degrade user experiences and may arouse complaints for the right violation. Secondly, blocking users or social links for restraining rumors also brings the loss of the diffusion of positive informations, say information loss, which is not beneficial to the viral marketers that utilize information cascading to promote products [1]- [6], [23] [24]. Regarding the limitations of existing solutions, in this paper, we take the first look into limiting the cascading of sensitive informations while preserving the diffusion of non sensitive ones to lower the information loss. Considering the randomness of the users accepting informations diffused from their social neighbors, we adopt the widely used random diffusion model that each user diffuses information to his social neighbor successfully with a diffusion probability via the social link between them.

Then our technical 1 objective is adjusting the diffusion probabilities via social links to minimize the diffusion size of sensitive informations, under the constraint of keeping the value of the sum of diffusion probabilities via all social links. Corresponding to the reality, we consider a case where some advertisements in viral marketing and some rumors simultaneously diffuse over an online social network. In this case, decreasing diffusion probabilities models the measures such as deleting partial posts or fanpages reposted by users [25] [26], while the measures for increasing diffusion probabilities include sticking and adding pushes or deliveries of the posts reposted by given users [16] [27].

Then, if network managers decrease the diffusion probability from a user holding rumors, the advertisements diffused from the user will inevitably be constrained as well. Thus, for lowering the diffusion loss of the advertisements and preserving the global diffusion ability of the whole network on diffusing non-sensitive informations, a natural approach is increasing the diffusion probabilities from one or more other users which hold the advertisements. We study the problem of interest on both fully-known and semi-known networks which are the two main scenarios considered in current studies on information diffusion [1]- [16]. Over the fully known network, we assume network managers know the diffusion abilities of all users. The examples for the fully-known network lie on the social networks for enterprises (e.g., Skype) or special interest groups (SIGs) (e.g., Douban1). As



the full topology of a local social network, which consists of the staff of a same enterprise or the members in a same SIG, is available to network managers, it is feasible to quantify the diffusion abilities of all users.

On the contrast, the semi known network here refers to the case that diffusion abilities of partial users remain unknown in advance. For example, the data of Facebook was reported to be utilized to influence the 2016 election in the US, which then led to a severe trust crisis for Facebook. Thus, due to the privacy concern and potential side effect, even for network managers, it is difficult to obtain the full topology of some global large scale social networks like Facebook, Wechat. Unless the full network topology is known, we cannot evaluate the diffusion abilities of all users. Over the fully-known network, although we can determine the diffusion probability variations via social links through solving a constrained minimization problem, the huge size of social links in current large scale networks leads to the high complexity of the problem. Moreover, the unknown diffusion abilities of partial users over the semi-known network induce it infeasible to directly solve the constrained minimization problem for minimizing the diffusion size of sensitive informations.

To tackle the above challenges, we utilize the constrained combinatorial multi-arm bandit framework to jointly design our solutions over the fully-known and semi-known networks, where we take the diffusion size of sensitive informations as the reward of a bandit and model the probability variations as the arms in bandit. With this mapping, we determine the probability variations through a constrained arms picking process with the aim of minimizing the obtained rewards. Through incorporating the constraint of diffusion probability variations into the construction of the arms of bandit, we relax the problem of interest into an unconstrained minimization problem when determining the diffusion probability variations based on the arms. This enables us to determine the probability variations via social links with high efficiency.

Furthermore, for coping with the unknown diffusion abilities over the semi-known network, we propose to iteratively learn the unknown diffusion abilities through learning the reward distributions of the arms based on the rewards obtained from previously picked arms, and then determine the diffusion probability variations based on the learned reward distributions of arms. Our main contributions are summarized as follows: (1) We take the first look into minimizing the diffusion size of sensitive informations while preserving the diffusion non-sensitive ones. We formulate the problem of interest into a constrained minimization problem where we characterize the intention of preserving non-sensitive information diffusions as the constraint. (2) We propose an efficient bandit based framework to jointly explore the solutions over the fully known and semi known networks within polynomial running time.

Moreover, we design the distributed implementation scheme of our solutions for the further improvement of time efficiency. (3) We further extend our bandit based solution into a “learning- determining” manner for addressing the challenge of unknown diffusion abilities in semi-known networks. We theoretically prove that the regret bound of our solution is sub-linear to the diffusion time, indicating that the probability variations returned by our solution approximates to the optimal one with the increase of diffusion time. (4) We perform extensive experiments on both real and synthetic social network datasets.



2. EXISTING SYSTEM

Kempe et al. [23] first propose two classic diffusion models: Independent Cascading (IC) model and linear threshold (LT) model. In the IC model, each user has a single chance to successfully diffuse the information to his neighbors with a given probability after this user having received the information. While in the LT model, a user would get the information if a certain fraction of his neighbors have received the information. Since then, a great deal of works study the Influence Maximization (IM) problem, which focuses on efficiently selecting the optimal seed users to trigger a diffusion process in hope of maximizing the final information diffusion size [1]. Recently, due to the high cost of seeding influential users, Shi et al. [3] propose to let influential users repost the required information while seed the ordinary users for lowering the cost of IM campaign. Similar to the multi-round setting in this paper, the seed selection for maximizing the information diffusion in multiple time rounds is considered in [2] [40]. Moreover, considering the widespread interactions between the cyber (online) and physical (offline) worlds, offline events are utilized in [7] to further improve the performance of IM.

On the contrast of the IM problem, there are also abundant researches focusing on minimizing the influence of rumors. One strategy for rumor influence minimization is diffusing the truths over network to counteract rumors [12]- [14]. Specifically, the competitive linear threshold (CLT) model that characterizes the competing diffusion of truth and rumor is introduced in [12]. Then He et al. [12] and Chen et al. [14] propose to select a set of seed users to maximize the diffusion of truths under the CLT model. Chen et al. [13] extends the IC model to describe the diffusion of positive informations under the effect of negative information, and studies how to maximize the positive information diffusion. However, such clarifying measure cannot be used to constrain the diffusion of private sensitive informations such as personal informations, trade secrets.

Another class of rumor blocking measures focuses on blocking a certain number of influential users [9] [15] or social links [17]- [20]. On one hand, Song et al. [15] propose to temporarily block a number of users with high diffusion abilities to reduce the diffusion of rumors before a deadline. With the consideration of user experiences, Wang et al. [9] study the online rumor blocking problem that periodically blocking a fraction of users during the rumor diffusion, and set a threshold to controls the blocking time of each user. Further, for coping with the unforeseen events in rumor diffusion, the adaptive blocking strategy is proposed in [10]. On the other hand, considering that straightforwardly blocking users is not desirable, [17]- [20] propose to block a given number of social links for minimizing the diffusion of rumors. However, as we illustrated before, this kind of measures may incur much information diffusion loss, if being adopted to constrain the diffusion of the sensitive informations considered in this paper. In addition, taking measures to constrain or promote information diffusion is also related to the studies about the effect of human behaviors on diffusion [29] [41] [42].

Disadvantages

The system is less effective due to lack of Constraining sensitive Information diffusion.



The system doesn't effective due to lack of Mapping Adaptive Diffusion in Fully-known Network into Bandit.

3. PROPOSED SYSTEM

The system takes the first look into minimizing the diffusion size of sensitive informations while preserving the diffusion of non-sensitive ones. We formulate the problem of interest into a constrained minimization problem where we characterize the intention of preserving non-sensitive information diffusions as the constraint.

The system proposes an efficient bandit based framework to jointly explore the solutions over the fully-known and semiknown networks within polynomial running time. Moreover, we design the distributed implementation scheme of our solutions for the further improvement of time efficiency.

The system further extend our bandit based solution into a "learning- determining" manner for addressing the challenge of unknown diffusion abilities in semi-known networks. We theoretically prove that the regret bound of our solution is sub-linear to the diffusion time, indicating that the probability variations returned by our solution approximates to the optimal one with the increase of diffusion time.

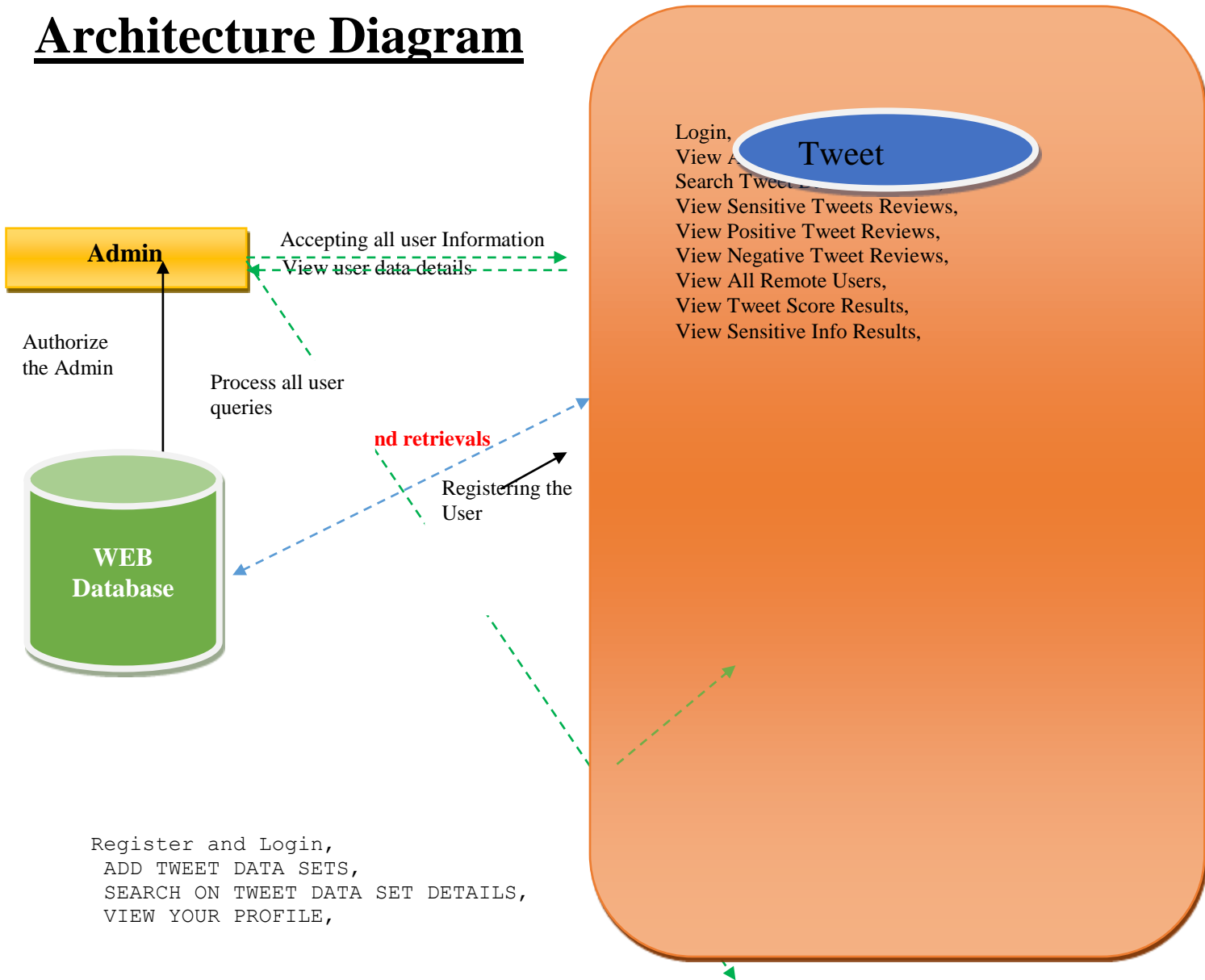
The system performs extensive experiments on both real and synthetic social network datasets. The results demonstrate that the proposed algorithms can effectively constrain the diffusion of sensitive informations, and more importantly, enjoy a superiority over four baselines in terms of 40% less information diffusion loss..

Advantages

The system is more effective due to ADAPTIVE DIFFUSION IN FULLY-KNOWN NETWORKS.

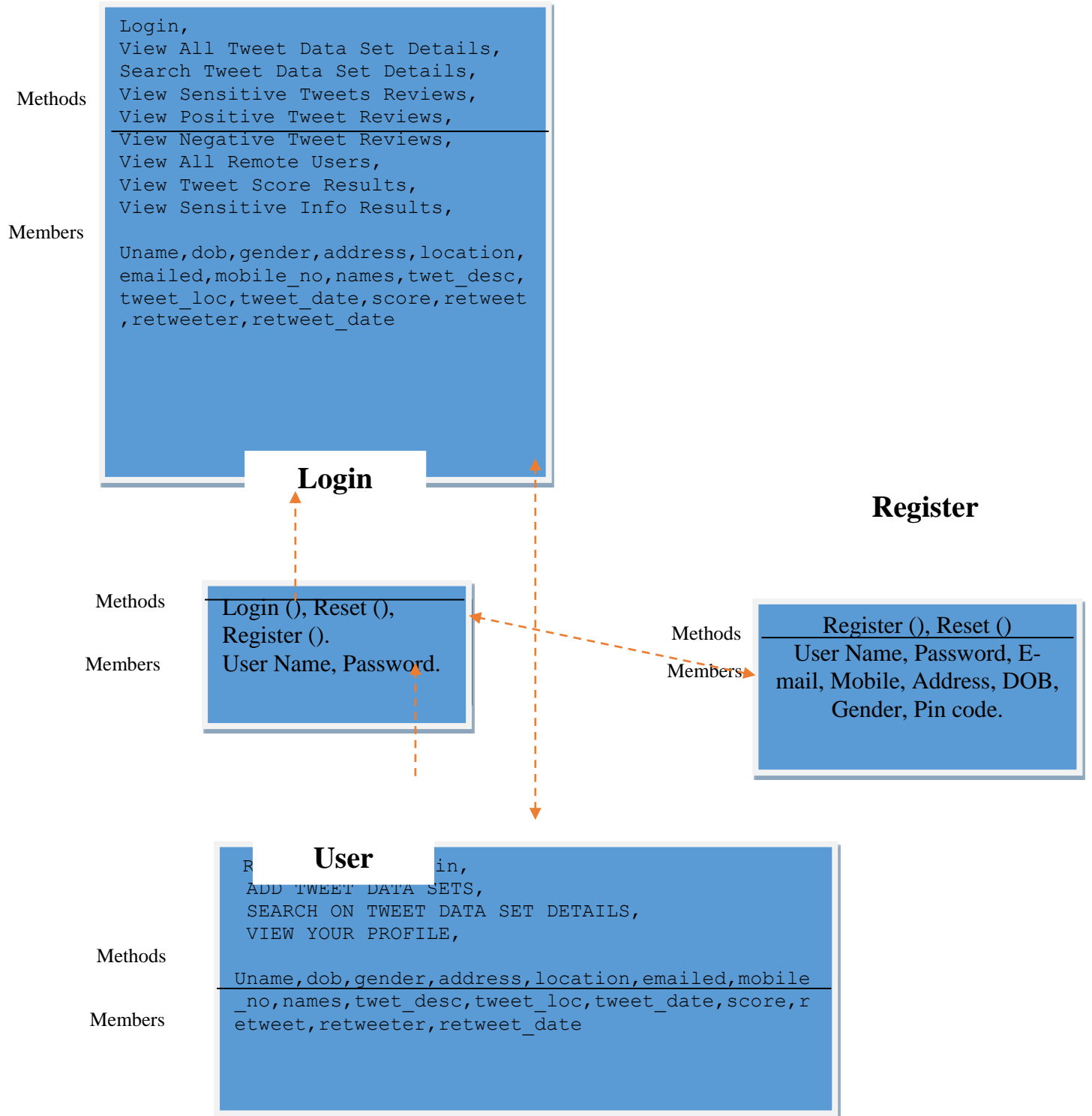
The system is more effective due to Constraining sensitive Information diffusion.

Architecture Diagram



➤ **Class Diagram :**

Tweet Server



4. CONCLUSIONS

In this paper, we study the problem of constraining the diffusion of sensitive informations in social networks while preserving the diffusion of non-sensitive informations. We model the



diffusion constraining measures as the variations of diffusion probabilities via social links, and model the problem of interest as adaptively determining the probability variations through a constrained minimization problem in multiple rounds. We utilize the CCMAB framework to jointly design our solutions in the fully-known and semi known networks. Over the fully-known network, we propose the CCMAB based algorithm ADFN to efficiently determine the probability variations via social links. Over the semi-known network, for tackling the challenge of unknown diffusion abilities of partial users, we propose the algorithm ADSN to iteratively learn the unknown diffusion abilities and determine the probability variations based on the learned diffusion abilities in each round. The analysis of regret bound and extensive experiments have been conducted to justify the superiority of our solutions.

In addition, in the current work, we define the constraint of maintaining the sum of diffusion probabilities via edges in the objective problem, for the aim of preserving the global diffusion ability of the whole network on diffusing nonsensitive informations. In the future work, we will explore other relevant solutions such as simultaneously minimizing the sensitive information diffusion and maximizing the nonsensitive information diffusion.

5. REFERENCES

- [1] Y. Li, J. Fan, Y. Wang, and K. L. Tan, "Influence maximization on social graphs: A survey", in IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 30, no. 10, pp. 1852-1872, 2018.
- [2] L. Sun, W. Huang, P. S. Yu, and W. Chen, "Multi-round influence maximization", in Proc. ACM SIGKDD, 2018.
- [3] Q. Shi, C. Wang, J. Chen, Y. Feng, and C. Chen, "Post and repost: A holistic view of budgeted influence maximization", in Neurocomputing, vol. 338, pp. 92-100, 2019.
- [4] X. Wu, L. Fu, Y. Yao, X. Fu, X. Wang, and G. Chen, "GLP: a novel framework for group-level location promotion in Geo-social networks", in IEEE/ACM Transactions on Networking (TON), vol. 26, no. 6, pp. 1-14, 2018.
- [5] Y. Lin, W. Chen, and J. C. Lui, "Boosting information spread: An algorithmic approach", in Proc. IEEE ICDE, 2017.
- [6] Y. Zhang, and B. A. Prakash, "Data-aware vaccine allocation over large networks", in ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 10, no. 2, article 20, 2015.
- [7] Q. Shi, C. Wang, J. Chen, Y. Feng, and C. Chen, "Location driven influence maximization: Online spread via offline deployment", in Knowledge-Based Systems, vol. 166, pp. 30-41, 2019.
- [8] H. T. Nguyen, T. P. Nguyen, T. N. Vu, and T. N. Dinh, "Outward influence and cascade size estimation in billion-scale networks, in Proc. ACM SIGMETRICS, 2017.
- [9] B. Wang, G. Chen, L. Fu, L. Song, and X. Wang, "Drimux: Dynamic rumor influence minimization with user experience in social networks", in IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 29, no. 10, pp. 2168-2181, 2017.



[10] Q. Shi, C. Wang, D. Ye, J. Chen, Y. Feng, and C. Chen, “Adaptive Influence Blocking: Minimizing the Negative Spread by Observation-based Policies”, in Proc. IEEE ICDE, 2019.

EMPLOYEE PROMOTION SYSTEM

Palla Santhosh^{*1}

^{*1}B.V. Raju College MCA Department, Adikavi Nannaya University, Bhimavaram,
Andhra Pradesh, India.

DOI : <https://www.doi.org/10.56726/IRJMETS42657>

ABSTRACT

The project aims to address the issue of employees seeking opportunities outside the company after gaining knowledge from company courses. It emphasizes the importance of providing internal growth opportunities to skilled employees. In this process using python Django programming.

I. INTRODUCTION

The employee promotion system is a crucial aspect of any company's human resource management strategy. It plays a significant role in nurturing talent, retaining skilled employees, and fostering a culture of growth and development within the organization. However, many companies struggle to provide internal growth opportunities, leading to employee attrition and missed potential.

II. METHODOLOGY

The existing system suffers from several disadvantages that can significantly impact the company's success and employee satisfaction. The lack of internal growth opportunities results in a higher turnover rate as skilled employees seek better prospects elsewhere. This leads to increased recruitment and training costs as the company relies on external hires to fill important roles. Additionally, employees may feel undervalued and experience low morale and engagement, which negatively affects productivity and job satisfaction. The absence of a clear career progression path also reduces employee loyalty and makes it difficult for the company to attract top talent. Furthermore, the existing system hampers knowledge transfer and collaboration within the organization, hindering overall learning and innovation. By addressing these shortcomings through the proposed web application, the company can create an environment that fosters employee development, engagement, and retention while maximizing the utilization of knowledge gained from company courses.

III. SYSTEM ANALYSIS

The existing system for employee promotion and internal growth opportunities within companies often lacks a centralized platform or structured process. Employees acquire knowledge through company courses, training programs, or professional development initiatives, but these opportunities may not be effectively utilized to retain and develop skilled employees. The following practices are commonly observed in the existing system:

IV. PROPOSED SYSTEM

The proposed system aims to overcome the limitations of the existing system by providing a comprehensive web application that facilitates internal growth opportunities for employees within the company. By addressing the challenges of employee retention, knowledge utilization, and career advancement, the proposed system seeks to create a more engaging and fulfilling work environment.

The core component of the proposed system is a user-friendly web application that streamlines the process of applying for interviews, scheduling them, and conducting video interviews. Employees can sign up and access their accounts using unique credentials, allowing them to navigate through the application's various features. The application provides a seamless experience for employees to apply for interviews, submit necessary details, and have their data securely stored in a database.

Admin users, responsible for conducting interviews, have dedicated access to employee data through the application. They can review the list of employees who have applied for interviews and have the option to accept or reject interview requests. Once an interview request is accepted, the admin can provide details such as the scheduled date and mode of the interview. The employee then receives an interview link to join the interview room, which can be created by the admin through the application.

The proposed system also emphasizes the importance of knowledge transfer and collaboration within the organization. By effectively utilizing the knowledge gained from company courses, employees can easily access and share valuable insights through the web application. This promotes efficient knowledge sharing, collaboration, and overall organizational learning.

The advantages of the proposed system include enhanced employee retention, leveraging employee expertise, cost savings through reduced reliance on external hiring, increased employee morale and engagement, attraction of top talent, efficient knowledge sharing, and improved succession planning. By providing internal growth opportunities and maximizing the utilization of employee knowledge and expertise, the proposed system aims to foster.

In summary, the proposed system's web application offers a user-friendly platform that addresses the limitations of the existing system. It provides a structured process for applying for interviews, conducting video interviews, and promoting internal growth opportunities. By implementing this system, the company can create a more engaging, fulfilling, and collaborative work environment that fosters employee development, satisfaction, and long-term success.

V. INPUT AND OUTPUT DESIGN

Input Design:

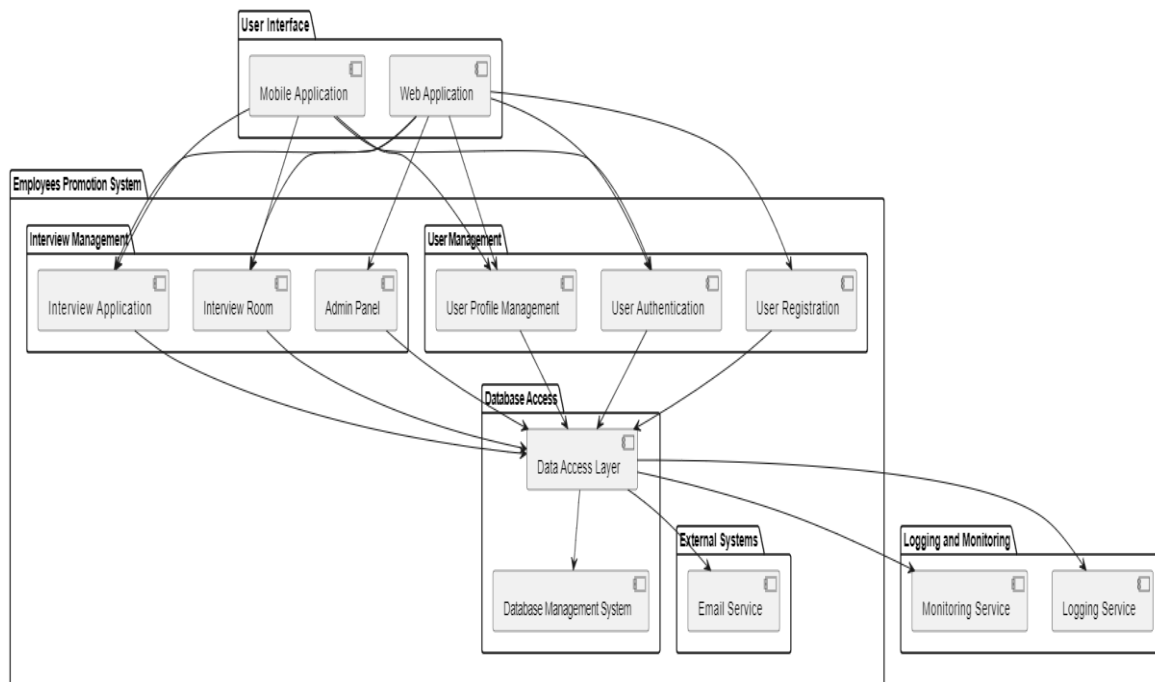
The employee promotion system incorporates various input fields to facilitate its functionalities. During the signup and login process, users are required to provide their credentials, including a user ID and password. These inputs are necessary for creating and accessing user accounts in the system. When employees apply for interviews, they are presented with an application form that includes text input fields for capturing their name, employee ID, contact information, skills, qualifications, desired position, and additional comments. Admin users, responsible for conducting interviews, are required to input their unique credentials to log in and access the employee data. These inputs help ensure secure access to the system's administrative features.

Output design:

The employee promotion system provides informative and user-friendly outputs to guide users through the process. Upon successful login, employees are directed to the home page, which features navigation links such as "apply interview," "admins (interviewers)," and "create interview room." These outputs enable easy navigation and access to the relevant functionalities. When employees submit their interview application, the system generates a confirmation message indicating the successful submission of the application. Admin users can view a list of employees who have applied for interviews, and the system displays their details, such as name, employee ID, contact information, skills, and desired position. When an admin accepts an interview request, the system generates a confirmation message with the scheduled date and mode of the interview. The employee receives an interview link from the admin, providing them with the necessary information to join the interview room. Throughout the process, the system also displays error messages for any invalid or incomplete inputs, ensuring that users are notified of any issues and can take appropriate actions.

I hope this separation into distinct paragraphs for input design and output design provides a clearer understanding of how the system handles user inputs and delivers relevant outputs.

VI. ARCHITECTURE DIAGRAM



All hypertext links and section bookmarks will be removed from papers during the processing of papers for publication. If you need to refer to an Internet email address or URL in your paper, you must type out the address or URL fully in Regular font.

VII. CONCLUSION

In conclusion, the proposed web application for the employee promotion system offers a comprehensive solution to address the limitations of the existing system. By providing internal growth opportunities, the application aims to enhance employee satisfaction, leverage their expertise, reduce recruitment costs, and foster a culture of continuous learning within the organization. The user-friendly interface and structured processes of the web application make it easier for employees to apply for interviews, schedule and conduct video interviews, and access internal career development resources. This system has the potential to create a positive work environment that promotes employee development, engagement, and long-term loyalty.

Moving forward, there are several avenues for future work and improvements to the proposed employee promotion system. One area of focus is performance optimization, where continuous efforts can be made to ensure smooth functionality even with a growing number of users and increasing data volume. This can involve optimizing database queries, implementing caching mechanisms, and leveraging cloud-based infrastructure to handle scalability. Additionally, rZpersonalized career development recommendations can be implemented based on employees' skills, interests, and aspirations. By utilizing advanced algorithms and machine learning techniques, the system can match employees with suitable career paths and suggest relevant training programs. Integration with existing learning management systems (LMS) or the creation of an integrated LMS within the application can enhance the employee training and development process. This integration would enable employees to access relevant training materials, track their progress, and earn certifications directly through the promotion system. Furthermore, incorporating performance evaluation features and feedback mechanisms within the web application can provide a comprehensive assessment of employees' progress and offer timely feedback. This can include regular performance reviews, 360-degree feedback, and goal-setting functionalities.

Integrating the employee promotion system with other HR systems, such as payroll and employee records, can streamline data management and ensure seamless information flow across different departments. This integration would enhance the overall HR processes and provide a unified view of employee information for better decision-making. Adding gamification elements, such as badges, leaderboards, and rewards, can also enhance employee engagement and motivation within the application. By incorporating game-like features,

employees can be incentivized to actively participate in the promotion process and engage in continuous learning and development.

Additionally, developing a mobile application version of the web application would provide employees with greater flexibility and accessibility. A mobile app would enable employees to access the promotion system, apply for interviews, and receive notifications on the go, enhancing the user experience and convenience. This expansion into mobile platforms can cater to the preferences of employees who rely heavily on mobile devices for work-related activities.

By pursuing these future developments and enhancements, the employee promotion system can continue to evolve and adapt to meet the changing needs of the organization and its employees. This ongoing improvement will contribute to a more robust and effective system that fosters employee growth, engagement, and organizational success.

ACKNOWLEDGEMENT

I would like to express my profound gratitude to Mr./Mrs. V Bhaskara Murthy Sir__ (name of the HOD), of __M.C.A__ (designation and department name) department, and to the completion of my project titled __Employee Promotion System__.

VIII. REFERENCES

- [1] Introduction to computer science and programming using python in:
<https://www.youtube.com/watch?v=rfscVS0vtbw>
- [2] Django Backend programming in <https://www.youtube.com/watch?v=jBzwzrDvZ18>
- [3] Front End Html, Css, Bootstrap, Javascript in Udemy Online Courses
- [4] Readyment Templates in Codepen.com

UNSUPERVISED DOMAIN ADAPTATION FOR CRIME RISK PREDICTION ACROSS CITIES

Parvathala Baby Venkata Naga Padma (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

Crime risk prediction is crucial for city safety and residents' life quality. However, without labeled data, it is challenging to predict crime risk in cities. Due to municipal regulations and maintenance costs, it is not trivial for many cities to collect high-quality labeled crime data. In particular, some cities have lots of labeled data while others may have few. It has been possible to develop a crime prediction model for a city without labeled crime data by learning knowledge from a city with abundant data. Nevertheless, the inconsistency of relevant context data between cities exacerbates the difficulty of this prediction task. To this end, this article proposes an effective unsupervised domain adaptation model (UDAC) for crime risk prediction across cities while addressing the contexts' inconsistency issue. More specifically, we first identify several similar source city grids for each target city grid. Based on these source city grids, we then construct auxiliary contexts for the target city, to make contexts consistent between the two cities. A dense convolutional network with unsupervised domain adaptation is designed to learn high-level representations for accurate crime risk prediction and simultaneously learn domain-invariant features for domain adaptation. The effectiveness of our model is verified through extensive experiments using three real-world datasets.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, different type of algorithms is trained to make classifications or predictions, and to uncover key insights in this project. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics.

Machine learning algorithms build a model based on this project data, known as training data, in order to make predictions or decisions without being explicitly programmed to do so. Machine

learning algorithms are used in a wide variety of datasets, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks.

1. INTRODUCTION

CRIME continuously threatens urban safety and undermines citizens' life quality. According to [1], there have been 435 mass shooting events happened in the United States during the year 2019, resulting in 517 dead, 1648 wounded, severe property loss, and inestimable grief. Thus, sensing crime risk is important for individuals and society, to prevent and reduce potential crime events. Fortunately, the availability of various urban data in some cities (e.g., Chicago) fosters unprecedented opportunities for researchers to explore crime-related problems, such as crime hotspot detection [2], [3], crime classification [4], [5], [6], crime rate inference [7], [8], and crime count prediction [4], [5], [9], [10], [11]. An amount of urban data has been investigated to be helpful for performance improvement of crime-related studies. For example, the occurrence of crime events may be affected by human mobility, that more crowd of people may bring an increasing possibility of larceny.

Nevertheless, due to the uneven development level of cities, a number of cities do not disclose data to the public for some possible reasons, i.e., the high cost of data collection and maintenance, the absence of clear-cut regulations, and increasing privacy concerns. Thus, residents need sufficient experience to sense whether there will be risk. But not all residents have such local experience, and this brings more challenges for newcomers, e.g., tourists. Recently, transfer learning [12], [13] provides a new paradigm that enables us to use learned knowledge from a data-rich city (*source city*) to solve similar tasks in a data-scarce city (*target city*), e.g., chain store site recommendation and crowd flow prediction [14], [15]. Therefore, we attempt to resort to unsupervised transfer learning to explore crime risk prediction in cities without labeled crime data.

2. EXISTING SYSTEM

Our work is related to previous studies on crime prediction and unsupervised domain adaptation. In this section, we briefly introduce some related work from these two categories.

There have been a few studies on urban crime prediction in the past decades. Identifying relevant external features for crime prediction study is significant. Ranson [16] analyzed meteorological

data and found that these data may be relevant to crime, e.g., weather and temperature information.

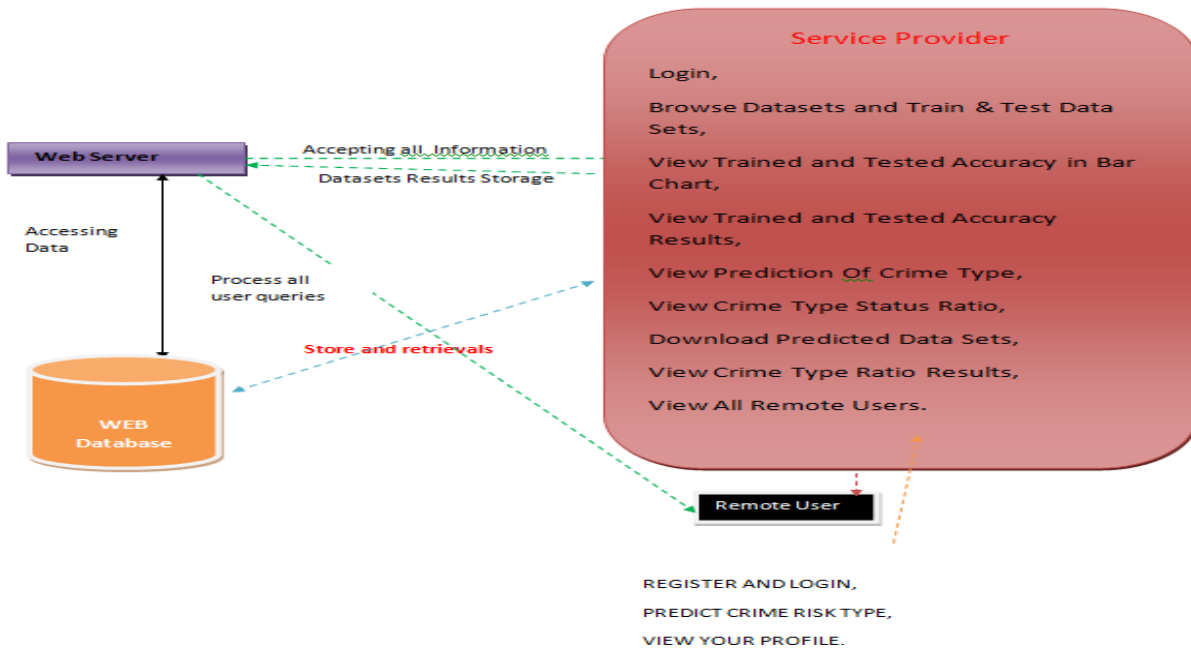
Zhou *et al.* [17] conducted a fine-grained study to understand crime leveraging various urban data, including meteorological data, POI distribution, and taxi trips' data. They found valuable correlation between these data and crime. With features analyzed, designing effective models to achieve accurate prediction has been popular. A lot of spatio-temporal prediction models have been proposed over the past few decades capturing spatial and temporal dependencies to solve various tasks, such as traffic prediction and inference [18], [19], [20], social event prediction [21], air quality prediction [22], and logistics management optimization [23].

However, these spatio-temporal prediction models paid little effort on the prediction tasks without labeled data. For crime prediction, a large amount of data, e.g., taxi trip, Twitter, demographic, and Foursquare data, have been used in various methods (e.g., linear models, count models, and machine learning models) to improve prediction performance [9], [24],[25]. Huang *et al.* [4] proposed a hierarchical recurrent neural network with an attention layer to capture dynamic patterns and learn temporal relevance for future crime occurrences' prediction, using crime data, POI, and 311 public service complaint data. Yang *et al.* [3] leveraged Twitter and POI data into multiple machine learning models (e.g., random forest and decision tree) to predict crime hotspots in NYC.

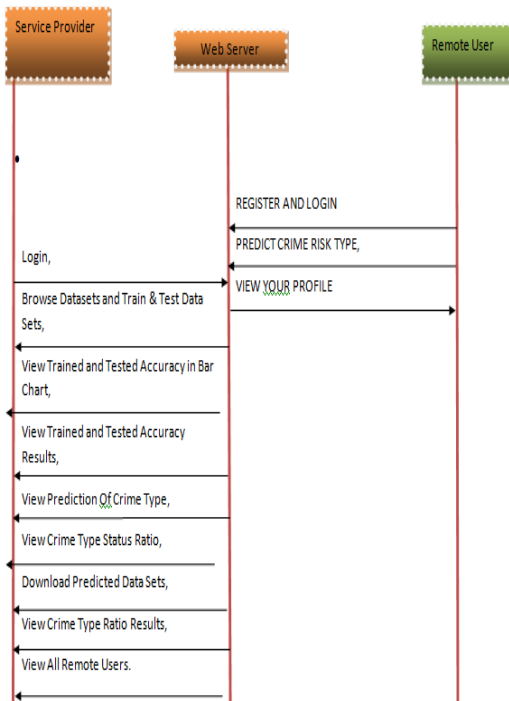
Yi *et al.* [26] proposed an integrated model using a clustered continuous conditional random field (CCRF) method to extract spatio-temporal features and improve future crime prediction performance. They further incorporated long short-term memory (LSTM) units into the aforementioned CCRF method to learn nonlinear relationship between the input and the output, and stacked denoising autoencoder to learn pairwise interactions between spatial regions [10].

Zhou *et al.* [27] proposed a hierarchical framework for road-level crime prediction, which first established a pattern using spatio-temporal features to estimate crime prior knowledge and then update crime prediction results incorporating recurrence crime features. They further investigated crime dynamics from the perspective of influence propagation and proposed a zero-inflated negative binomial regression model to predict future road-level crime risk [28].

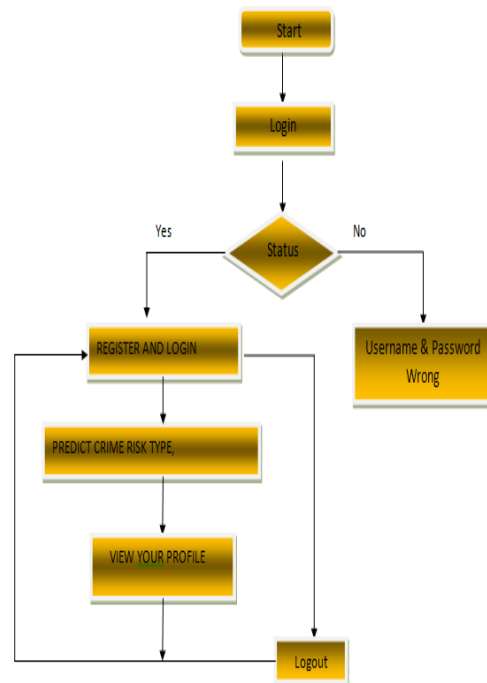
Architecture Diagram



Sequence Diagram



Flow Chart : Remote User



3. CONCLUSION

In this article, to solve the prediction tasks with unlabeled data, we propose an unsupervised domain adaptation method. This method can learn knowledge from a source city with labeled data and transfer it to the target city while addressing the context inconsistency issue between cities, to predict crime risk. We partition cities into multiple equal-sized grids and identify several similar source city grids for each target grid. Based on these pairs, we construct auxiliary features for the target city, to address the contexts' inconsistency problem across cities. We then propose a dense-convolutional network-based unsupervised domain adaptation module to learn knowledge from the source city and apply it to the target city for future crime risk prediction. Domain-invariant features are learned to facilitate knowledge transfer. Extensive experiments are conducted to verify the effectiveness of our method using real-world data from NYC, Chicago, and LA. The experimental results reveal the superiority of our proposed method over various state-of-the-art comparison methods.

In the future, we intend to improve our work from several perspectives. First, we plan to explore prediction performance when more serious contexts' inconsistency exists between the source and target cities. Second, we plan to investigate a fine-grained unsupervised crime risk prediction, such as predicting crime risk in roads. This would be more challenge due to severe data scarcity problem.

4. REFERENCES

- [1] (2020). GunViolenceArchive. *List of Mass Shootings in the 70 United States in 2019*. [Online]. Available: <https://www.gunviolencearchive.org/reports/mass-shooting>
- [2] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Syst.*, vol. 61, pp. 115–125, May 2014.
- [3] D. Yang, T. Heaney, A. Tonon, L. Wang, and P. Cudré-Mauroux, "CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion," *World Wide Web*, vol. 21, pp. 1323–1347, Sep. 2018.
- [4] C. Huang, J. Zhang, Y. Zheng, and N. V. Chawla, "DeepCrime: Attentive hierarchical recurrent networks for crime prediction," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, 2018, pp. 1423–1432.

- [5] L. Xia *et al.*, “Spatial–temporal sequential hypergraph network for crime 716 prediction with dynamic multiplex relation learning,” in *Proc. 30th Int. 717 Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1631–1637. 718
- [6] P. Das, A. K. Das, J. Nayak, D. Pelusi, and W. Ding, “Incremental classi719 fier in crime prediction using bi-objective particle swarm optimization,” *720 Inf. Sci.*, vol. 562, pp. 279–303, Jul. 2021.
- [7] H. Wang, D. Kifer, C. Graif, and Z. Li, “Crime rate inference with big data,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 635–644.
- [8] H. Wang, H. Yao, D. Kifer, C. Graif, and Z. Li, “Non-stationary model for crime rate inference using modern urban data,” *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 180–194, Jun. 2019.
- [9] L. Vomfell, W. K. Härdle, and S. Lessmann, “Improving crime count forecasts using Twitter and taxi data,” *Decis. Support Syst.*, vol. 113, pp. 73–85, Sep. 2018.
- [10] F. Yi, Z. Yu, F. Zhuang, and B. Guo, “Neural network based continuous conditional random field for fine-grained crime prediction,” in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 10–16.

EMOTION RECOGNITION BY TEXTUAL TWEETS CLASSIFICATION USING VOTING CLASSIFIER LR-SGD

Patti Jagadeesh (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The proliferation of user-generated content on social media has made opinion mining an arduous job. As a microblogging platform, Twitter is being used to collect views about products, trends, and politics. Sentiment analysis is a technique used to analyze the attitude, emotions and opinions of different people towards anything, and it can be carried out on tweets to analyze public opinion on news, policies, social movements, and personalities. By employing Machine Learning models, opinion mining can be performed without reading tweets manually. Their results could assist governments and businesses in rolling out policies, products, and events. Seven Machine Learning models are implemented for emotion recognition by classifying tweets as happy or unhappy. With an in-depth comparative performance analysis, it was observed that proposed voting classifier(LR-SGD) with TF-IDF produces the most optimal result with 79% accuracy and 81% F1 score. To further validate stability of the proposed approach on two more datasets, one binary and other multi-class dataset and achieved robust results.

1. INTRODUCTION

Automatic emotion recognition, pattern recognition and computer vision have become significantly important in Artificial Intelligence lately with applications is a wide range of areas. Recently, social media platforms such as Twitter have generated enormous amounts of structured, unstructured and semi-structured data. One of the most recent example is COVID-19 infodemic that shows misinformation in social media can be far more important and devastating than a disaster such as a pandemic.

There is a need to analyse to accurately assign sentiment classes on a large scale. To perform such tasks, accurate NLP techniques and machine learning (ML) models for text classification are required. Twitter provides an opportunity to its users to analyse its data on a large and broader

point of view. Efficient methods are important to automatically label text data due to its noisy nature. In the past many studies have been performed on Twitter sentiment classification [1]. As Twitter is very fast and an efficient micro-blogging examination that facilitates the end users to transmit small posts are said to be tweets. Twitter is a highly demanding app in the world and is a successful platform in social media.

Free account can be created by using Twitter that can provide an enormous audience potential. With the purpose of business and marketing, Twitter can be proved as the best platform, through which one can get in touch with very rich and famous personalities like stars and celebrities, so their purchasing can be very charming for them as well as for advertisers. Using Twitter, every celebrity is linked with fans as well as to



grant a communication to followers. Such a platform is one of the superlative approaches for lovers as well. But, it has a short note range; only 140 letters for each post and it can type a post or link on the website since it has no cost and also open as the advertisements as well. There is no problem with clusters of personal ads which are similar to other social networking sites. It is quick because as a tweet is posted on Twitter, the public who is subsequent to respective business will get it without delay.

Companies and advertisers can compose utilization of this source to check the diverse operational point of views which are very considerable. With help of this, they will obtain an immediate response from their followers. Remarkably, a lot of businesses with the intention of purchase, Twitter followers increase their deals. Twitter facilitates the followers by making them identify regarding fresh business, products, services, websites, blogs, eBooks etc. Consequently, Twitter clients might tick lying on link and also optimistically endow in a manufactured goods or examine the products presented and to get share in pro_t. It is extremely effortless to utilize as people can follow to get the news and updates, as organizations can tweet or re-tweet, they can mark favorite or selected people to send the tweets, also know how to propel the posts plus to be able to endow their money and instance through it. Academy, Industry, super bowls and Grammy Awards of such major Sports and Entertainment events generate a lot of buzz in the global world by using it.

2. EXISTING SYSTEM

Sarlan *et al.* [2] established a sentiment analysis through extracting number of

tweets with the help of prototyping and the results organized customers' views via tweets into positive and negative. Their research divided into two phrases. The first part is based on literature study which involves the Sentiment analysis techniques and methods that nowadays are used. In the second part, the application necessities and operations are described preceding to its development.

In another research Alsaeedi and Zubair Khan [3] analyzed various kinds of sentiment analysis that is applied on to Twitter dataset and its conclusions. The distinct approaches and conclusions of algorithm performance were compared. Methods were used which were supervised ML based,, lexicon-based, ensemble methods. Authors used four methods that were Twitter sentiment Analysis using Supervised ML Approaches; Twitter sentiment Analysis using Ensemble Approaches. Twitter sentiment Analysis is using lexicon based Approaches.

Lexicon based approaches have been explored by many researchers for emotion classification. Bandhakavi *et al.* [4] performed emotion-based feature extraction using domain specific lexicon generation.

Disadvantages

The existing model which is ensemble of LR and SGD is not applied on both dataset and the results.

Voting Classifier(VC) is not a cooperative learning which engages multiple individual classifiers.

3. PROPOSED SYSTEM

In the proposed system, different techniques have been used for methodology in ML for its objectives. Versatile experiments were examined using different methods and techniques.

Multiple classifiers applied on the dataset, but the Voting classifier is an ensemble of Logistic Regression and Stochastic Gradient Descent outperforms than all other ML models in terms of accuracy, recall, precision and F1-score.

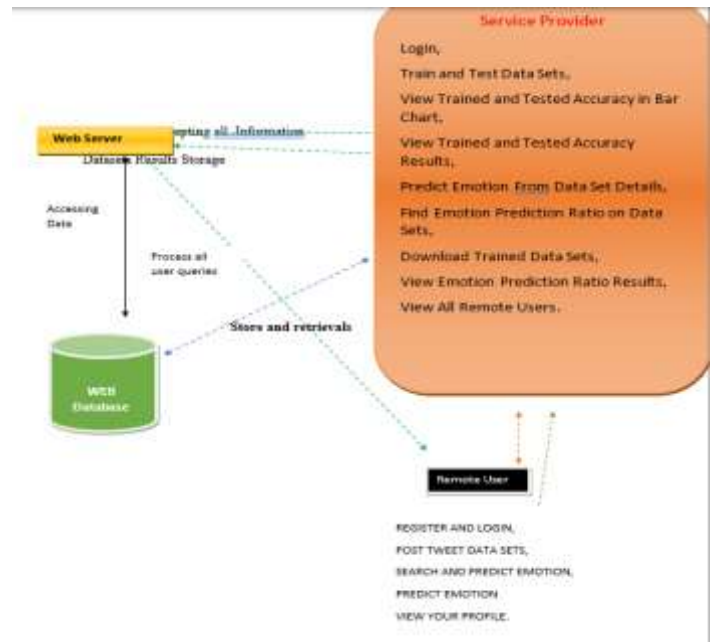
Twitter dataset used in this experiment is scrapped from Kaggle repository. First the dataset is pre-processed by removing unwanted data. Then, the data was split into two sets: training set and testing set. The training set was given the percentage of 70% while the test set portion is 30%. After that feature engineering techniques are applied on the training set. Multiple machine learning classifiers are trained on the training set and tested using the test set. The evaluation parameters used in this experiment are: (a) Accuracy (b) Recall (c) Precision (d) F1-score.

Advantages

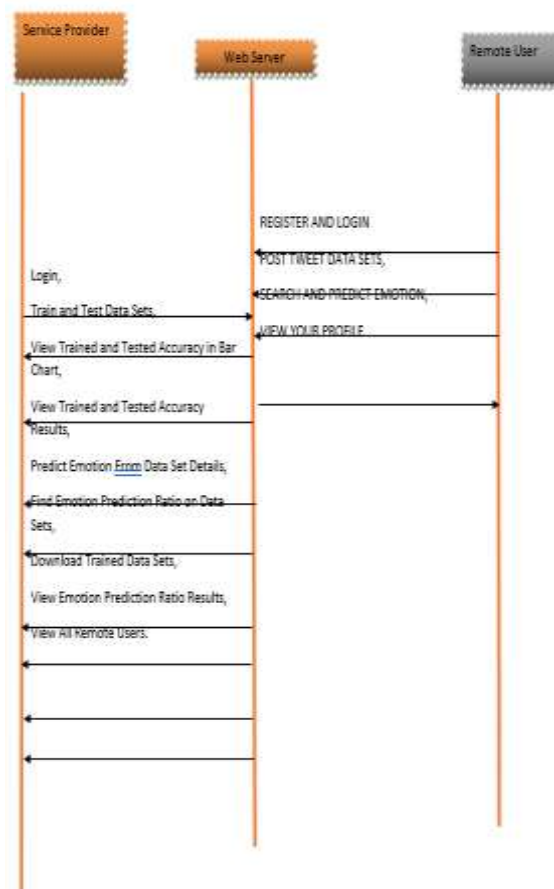
The proposed system presents a voting classifier (LR-SGD) and aims to estimate the performance of famous ML classifiers on twitter datasets.

Data Visualization helps to understand the hidden patterns lying inside the dataset. It helps to qualitatively get more details about the dataset by visualizing the characteristics of the attributes.

4. ARCHITECTURE DIAGRAM



Sequence Diagram



5. CONCLUSIONS

This paper proposed a novel combination of LR and SGD as a voting classifier for



emotion recognition by classifying tweets as happy or unhappy. Our experiments showed that one can improve the performance of models by recognizing patterns efficiently and through effective averaging combination of models. Experiments are conducted to test seven machine learning models that are; (1) SVM, (2) RF, (3) GBM, (4) LR, (5) DT, (6) NB and (7) VC(LR-SGD). This study also employed two feature representation techniques TF and TF-IDF. The results showed that all models performed well on tweet dataset but our proposed voting classifier VC(LR-SGD) outperforms by using both TF and TF-IDF among all. Proposed model achieves the highest results using TF-IDF with 79% Accuracy, 84% Recall and 81% F1-score. The proposed model is further validated on two more dataset and achieved robust results. The future work will compare more feature engineering techniques and explore more combinations of ensemble models to improve the performance. In addition, new techniques will be investigated to deal with sarcastic comments.

6. REFERENCES

- [1] N. F. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decis. Support Syst.*, vol. 66, pp. 170_179, Oct. 2014.
- [2] C. Kariya and P. Khodke, "Twitter sentiment analysis," in *Proc. Int. Conf. Emerg. Technol. (INCET)*, Jun. 2020, pp. 212_216.
- [3] A. Alsaeedi and M. Zubair, "A study on sentiment analysis techniques of Twitter data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 2, pp. 361_374, 2019.
- [4] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, "Lexicon based feature extraction for emotion text classification," *Pattern Recognit. Lett.*, vol. 93, pp. 133_142, Jul. 2017.
- [5] J. Capdevila, J. Cerquides, J. Nin, and J. Torres, "Tweet-SCAN: An event discovery technique for geo-located tweets," *Pattern Recognit. Lett.*, vol. 93, pp. 58_68, Jul. 2017.
- [6] T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu, and J. Planes, "An argumentative approach for discovering relevant opinions in Twitter with probabilistic valued relationships," *Pattern Recognit. Lett.*, vol. 105, pp. 191_199, Apr. 2018.
- [7] W. Chen, Y. Zhang, C. K. Yeo, C. T. Lau, and B. S. Lee, "Unsupervised rumor detection based on users' behaviors using neural networks," *Pattern Recognit. Lett.*, vol. 105, pp. 226_233, Apr. 2018.
- [8] H. Hakh, I. Aljarah, and B. Al-Shboul, "Online social media-based sentiment analysis for us airline companies," in *New Trends in Information Technology*. Amman, Jordan: Univ. of Jordan, Apr. 2017.
- [9] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Inf. Sci.*, vol. 181, no. 6, pp. 1138_1152, Mar. 2011.
- [10] M. Umer, S. Sadiq, M. Ahmad, S. Ullah, G. S. Choi, and A. Mehmood, "A novel stacked CNN for malarial parasite detection in thin blood smear images," *IEEE Access*, vol. 8, pp. 93782_93792, 2020.

FEATURE EXTRACTION FOR CLASSIFYING STUDENT BASED ON THERE ACADEMIC PERFORMANCE

Pecchetti Jyothi Kiran (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT Developing tools to support students and learning in a traditional or online setting is a significant task in today's educational environment. The initial steps towards enabling such technologies using machine learning techniques focused on predicting the student's performance in terms of the achieved grades. The disadvantage of these approaches is that they do not perform as well in predicting poor-performing students. The objective of our work is two-fold. First, in order to overcome this limitation, we explore if poorly performing students can be more accurately predicted by formulating the problem as binary classification. Second, in order to gain insights as to which are the factors that can lead to poor performance, we engineered a number of human interpretable features that quantify these factors. These features were derived from the students' grades from the University of Minnesota, an undergraduate public institution. Based on these features, we perform a study to identify different student groups of interest, while at the same time, identify their importance. Keywords academic student success, classification, feature importance

1. INTRODUCTION

Higher educational institutions constantly try to improve the retention and success of their enrolled students. According to the US National Center for Education Statistics [8], 60% of undergraduate students on four-year degrees will not graduate at the same institution where they started within the first six years. At the same time, 30% of college freshmen drop out after their first year of college. As a result, colleges look for ways to serve students more efficiently and effectively. This is where data mining is introduced to provide some solutions to these problems. Educational data mining and learning analytics have been developed to provide tools for supporting the learning process, like monitor and measure student progress, but also, predict success or guide intervention strategies. Most of the existing approaches focus on identifying students at risk who could benefit from further assistance in order to successfully complete a course or activity. A fundamental task in this process is to actually

predict the student's performance in terms of grades. While reasonable prediction accuracy has been achieved [14, 10], there is a significant weakness of the models proposed to identify the poor-performing students [18]. Usually, these models tend to be over-optimistic for the performance of students, as the majority of the students do well, or have satisfactory enough performance. In this paper, we investigate the problem of predicting the performance of a student in the end of the semester before he/she actually takes the course. In order to focus on the poor-performing students, who are the ones that need these systems the most, the prediction problem is formulated as a classification task, where two groups of students are formed according to their course performance. We essentially identify two complementary groups of students, the ones that are likely to successfully complete a course or activity, and the ones that seem to struggle. After identifying the latter group, we can provide additional resources and support to enhance their likelihood of success. However, "success" and "failure" can be relative or not. For example, a B- grade might be considered a bad grade for an excellent student, while being a good grade for a very weak student. We investigated different ways to define groups of students taking a course: failing students, students dropping the class, students performing worse than expected and students performing worse than expected, while taking into consideration the difficulty of a course. In order to gain more insight into the learning process and its most important characteristics, we have created features that capture possible factors that influence the grades at the end of the semester. Using these features, we present a comprehensive study to answer the following questions: which features are good indicators of a student's performance? which features are the most important? The findings are interesting, as different features are the most important for different classification tasks.

2. RELATED WORK

As we are interested in estimating next-term student performance, we will review the related work in this area of research. The binary classification has been used in various educational problems, like predicting if a student will drop out from high school [6] or to predict if a student will pass a module in a distance learning setting [7]. Multi-label classification has been applied to provide a qualitative measure of students' performance. In [17], decision tree and naive Bayes classifiers are used with data from a survey. Attributes collected by a learning management system have been employed to estimate the outcome as Fail, Pass, Good and Excellent [16], or to classify students [12]. Some approaches [11, 9] test different ways to label the student performance, with two (pass or fail) or more labels. The majority of

the aforementioned approaches are small-scale studies, that are applied to a limited number of courses. In recent years, influenced by advances in the recommender systems, big data approaches have been also utilized in the area of learning analytics. Initially, the term “next-term grade prediction” was introduced by Sweeney et al. [18] in the context of higher education, and it refers to the problem of predicting the grades for each student in the courses that he/she will take during the next semester. Models based on SVD and factorization machines (FM) were tested. In another approach [15], the previous performance of students controls the grade estimation in two different ways while building latent models. In [19], some additional state-of-the-art methods were used, as well as, a hybrid of FM and random forests (RF). The data used are the historical grades and additional content features, representing student, course and instructor characteristics. At the same setting, [14] and [10] developed course-specific methods to perform next-term grade prediction based on linear regression and matrix factorization. All these methods assign a specific numerical grade to each student’s attempt to take a course. A limitation identified in these approaches was that the developed models perform poorly for failing students. In [5], failing students have been completely removed from the dataset. As this is the subpopulation of students that needs additional support the most, it is very important for a model to be able to accurately identify these students at risk. This work is a more general study of the factors that influence the student performance, in a very large scale. The only observed data that we have available are the students’ grades at the end of the semester. In our approach, we formulated this problem as a binary classification task, in order to detect the different group of students. In other words, we keep the classification methodology, but apply it on the context of big data.

3. DATASET

First, we will clarify the use of some terms in the current context. An instance refers to the performance of a student, s , in a course, c , at the end of the semester. All the courses that a student took in past semesters, before taking course c , are the prior courses, denoted by $C_{s,allprior}$. The set of courses for a single semester x is denoted as $C_{s,x}$. Additionally, for a course c there might exist a stated set of courses that are required for a student to take before attempting c . We refer to this set as the prerequisite courses. Every course x worths a specified number of credits, cr_x . An undergraduate student enrolled to a college or university has to take some courses each semester, and receive a satisfactory grade in order to successfully complete them. Depending on the student’s degree program, these courses might be required, electives, or simply courses that the student takes for his/her own advancement,

intellectual curiosity, or enjoyment. If a student withdraws from a course after the first two weeks of classes, it is denoted by the letter ‘W’ in the student’s transcript. The original dataset was obtained from the University of Minnesota and it spans over 13 years. We removed any instances with a letter grade not in the A–F grading scale (A, A–, B+, B, B–, C+, C, C–, D+, D, F). Statistics about the grades in the dataset are shown in Fig. 1. In our dataset, the letter grade A is the most common. We extract features for the instances occurring during the last 10 fall and spring semesters. Given a semester, we utilize all the students that had taken the course before, and for each student taking a course, we extract a set of features. Additionally, we generate features for the instances awarded with the letter W, but we do not utilize them in any other way during the feature extraction process. These will be used only when trying to predict the students that drop-out from a course.

4. EXTRACTED FEATURES

Having as input the historical grading data, we derived different features to capture possible factors for a student’s poor performance. The features can be separated into three distinct categories: the student-specific (independent from course c), course-specific features (independent from students) and student- and course-specific features (they are a function of both s and c). All extracted features are described in Tables 1, 2, where related features are grouped together into eight different subcategories. The keywords on bold are used to indicate the corresponding group of features later. Note that for each $\{s, t, c\}$, where student s took course c in semester t , we generate a different set of features. Every set of features characterize a student’s attempt to take course c at the specific point of his/her studies. These features are either numerical, categorical or indicator variables. For indicator features, we use the values of 0 or 1. The categorical features are encoded via a numerical value. For example, the feature about the current semester is categorical, and the values $\{\text{fall, spring, summer}\}$ are transformed to $\{0,1,2\}$, respectively.

5. CLASSIFICATION PROBLEMS

5.1 Classification tasks Our motivation was to identify groups of students that need further assistance and guidance in order to successfully complete a course. These students could benefit from informed interventions. We consider this to be a binary classification problem, where these students form one of the classes and the remaining students form the other class. We consider different ways of measuring when a student does not do well in a course to deal with the performance measurement challenges we mentioned earlier. Unsatisfactory

performance can occur when the earned grade represents a performance that is below the student's potential. We considered the following four ways for labelling, resulting to these absolute and relative classification tasks: 1. Failing student performance, i.e., letter grades D and F (denoted as the Fgr task). 2. The letter grade W (denoted as the Wgr task). This represents the instances when the student dropped the course. This behavior is worrisome as it shows that either the student was not interested in the course anymore or he/she expects to perform poorly. 3. Student performance that is worse than expected, i.e., the grade achieved is more than two letter grades lower than the student's GPA (denoted as the RelF task). 4. Student performance that is worse than expected while taking into consideration the difficulty of the course (denoted as the RelCF task). The difficulty of a course is expressed by the average grade achieved by the students that took the course in prior offerings. A positive instance is when the grade achieved is more than two letter grades lower than the average of the student's GPA and the course's prior average grade. Statistics for the different classification tasks can be found at Table 3. As discussed at the related work section, it is easier to predict the successful students. In order to have a better understanding of the relative difficulty of this task compared with the four tasks mentioned above, we also examined the task of predicting the students that completed a course with the grade A (denoted as the Agr task).

5.2 Methods compared

In order to support students that need help to successfully complete a course, we will use classification techniques to identify them from the rest of the students. The instances of interest will be labeled as 1, and the rest as 0. The problem can be described as follows. We are given a set of training examples that are in the form (x, y) and we want to learn their structure. We assume that there is some unknown function $y = f(x)$, that corresponds the feature vector x to a value y . In our case, $y = \{0, 1\}$. A classifier is an hypothesis about the true function f . Given unseen values of x , it predicts the corresponding y values. We tested the following classifiers [4], using scikit-learn library in Python [13]: Decision Tree (DT) [2] and Linear Support Vector Machine (SVM) [3] as base classifiers, and Random Forest (RF) [1] and Gradient Boosting (GB) [4] as ensemble classifiers. While using DT, the classification process is modeled as a series of hierarchical decisions on the features, forming a tree-like structure. In other words, we ask a series of questions about the features of an instance, and based on the answer, we may ask more questions, until we reach to a conclusion about the class label of that instance. The goal is to get a split that allow us to make a confident prediction. Consider the m -dimensional space that is defined by the feature vectors x , of length m . There, every training instance corresponds to a single point. A Linear SVM looks for a decision boundary between two classes, a hyperplane that bisects the data

with the largest possible margin between the two different classes. The margin on each side of the hyperplane is the area with no data points in it. Ensemble methods try to increase the prediction accuracy by combining the results from multiple base classifiers. RF is a class of ensemble methods that uses decision trees as weak learners. Randomness has been explicitly inserted in the model building process, as every splitting criterion considers only a subset of features, randomly selected from the feature vector of x , to select the best split. Once we build all the trees, the majority class is reported. In boosting, a weight is associated with each training instance. Using the same algorithm, classifiers are training on a weighted training set to focus on hard-to-classify instances. At the end of each iteration, the weights of instances with high misclassification error are relatively increased for future iterations. In GB for binary classification, a single regression tree is built, where in each splitting criterion, only a subset of the features is considered. Once the tree is built, then, the corresponding weight of the classifier in the current iteration is estimated.

6. CONCLUSIONS

The purpose of this paper is to accurately identify students that are at risk. These students might fail the class, drop it, or perform worse than they usually do. We extracted features from historical grading data, in order to test different simple and sophisticated classification methods based on big data approaches. The best performing methods are the Gradient Boosting and Random Forest classifiers, based on AUC and F1 score metrics. We also got interesting findings that can explain the student performance.

7. REFERENCES

- [1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [2] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984. [3] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [4] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.
- [5] Z. Iqbal, J. Qadir, A. N. Mian, and F. Kamiran. Machine learning based student grade prediction: A case study. *arXiv preprint arXiv:1708.08744*, 2017.

- [6] J. E. Knowles. Of needles and haystacks: Building an accurate statewide dropout early warning system in wisconsin. *Journal of Educational Data Mining*, 7(3):18–67, 2015.
- [7] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Predicting students’ performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18(5):411–426, 2004.
- [8] J. McFarland, B. Hussar, C. de Brey, T. Snyder, X. Wang, S. Wilkinson-Flicker, S. Gebrekristos, J. Zhang, A. Rathbun, A. Barmer, et al. Undergraduate retention and graduation rates. In *The Condition of Education 2017*. NCES 2017-144. ERIC, 2017.
- [9] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In *Frontiers in education*, 2003. FIE 2003 33rd annual, volume 1, pages T2A–13. IEEE, 2003.
- [10] S. Morsy and G. Karypis. Cumulative knowledge-based regression models for next-term grade prediction. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 552–560. SIAM, 2017.

Suicidal ideation detection. A review of machine learning and applications

SOWJANYA PEEKA

PG Scholar, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

V. SRI VALLI DEVI

Assistant Professor, Department M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

Abstract

Suicide is a critical issue in modern society. Early detection and prevention of suicide attempts should be addressed to save people's life. Current suicidal ideation detection (SID) methods include clinical methods based on the interaction between social workers or experts and the targeted individuals and machine learning techniques with feature engineering or deep learning for automatic detection based on online social contents. This article is the first survey that comprehensively introduces and discusses the methods from these categories. Domain-specific applications of SID are reviewed according to their data sources, i.e., questionnaires, electronic health records, suicide notes, and online user content. Several specific tasks and data sets are introduced and summarized to facilitate further research. Finally, we summarize the limitations of current work and provide an outlook of further research directions.

Keyword: Suicidal ideation detection, social content, feature engineering, deep learning.

1. INTRODUCTION

1.1 Introduction:

MENTAL health issues, such as anxiety and depression, are becoming increasingly concerned in modern society, as they turn out to be especially severe in developed countries and emerging markets. Severe mental disorders without effective treatment can turn to suicidal ideation or even suicide attempts. Some online posts contain much negative information and generate problematic phenomena, such as cybers talking and cyber bullying.

Consequences can be severe and risky since such lousy information is often engaged in some form of social cruelty, leading to rumors or even mental damage. Research shows that there is a link between cyber bullying and suicide. Victims overexposed to too many negative messages

or events may become depressed and desperate; even worse, some may commit suicide. The reasons that people commit suicide are complicated. People with depression are highly likely to commit suicide, but many without depression can also have suicidal thoughts. According to the American Foundation for Suicide Prevention (AFSP), suicide factors fall under three categories: health factors, environmental factors, and historical factors. Ferrari et al found that mental health issues and substance use disorders are attributed to the factors of suicide.

O'Connor and Nock conducted a thorough review of the psychology of suicide and summarized psychological risks as personality and individual differences, cognitive factors, social factors, and negative life events. Suicidal ideation

detection (SID) determines whether the person has suicidal ideation or thoughts by given tabular data of a person or textual content written by a person. Due to the advances in social media and online anonymity, an increasing number of individuals turn to interact with others on the Internet. Online communication channels are becoming a new way for people to express their feelings, suffering, and suicidal tendencies. Hence, online channels have naturally started to act as a surveillance tool for suicidal ideation, and mining social content can improve suicide prevention.

Strange social phenomena are emerging, e.g., online communities reaching an agreement on self-mutilation and copycat suicide. For example, a social network phenomenon called the “Blue Whale Game”¹ in 2016 uses many tasks (such as self-harming) and leads game members to commit suicide in the end. Suicide is a critical social issue and takes thousands of lives every year. Thus, it is necessary to detect suicidality and prevent suicide before victims end their life. Early detection and treatment are regarded as the most effective ways to prevent potential suicide attempts. Potential victims with suicidal ideation may express their thoughts of committing suicide in fleeting thoughts, suicide plans, and role-playing. SID is to find out these risks of intentions or behaviors before tragedy strikes.

A meta-analysis conducted by McHugh et al. shown statistical limitations of ideation as a screening tool but also pointed out that people’s expression of suicidal ideation represents their psychological distress. Effective detection of early signals of suicidal ideation can identify people with suicidal thoughts and open a communication portal to let social workers

mitigate their mental issues. The reasons for suicide are complicated and attributed to a complex interaction of many factors. To detect suicidal ideation, many researchers conducted psychological and clinical studies and classified responses of questionnaires [. Based on their social media data, artificial intelligence (AI) and machine learning techniques can predict people’s likelihood of suicide, which can better understand people’s intentions and pave the way for early intervention. Detection on social content focuses on feature engineering, sentiment analysis, and deep learning. Those methods generally require heuristics to select features or design artificial neural network (ANN) architectures for learning rich representation.

The research trend focuses on selecting more useful features from people’s health records and developing neural architectures to understand the language with suicidal ideation better. Mobile technologies have been studied and applied to suicide prevention, for example, the mobile suicide intervention application I Bobbly [19] developed by the Black Dog Institute.² Many other suicide prevention tools integrated with social networking services have also been developed, including Samaritans Radar³ and Woebot.⁴ The former was a Twitter plug in that was later discontinued because of privacy issues. For monitoring alarming posts, the latter is a Face book chat bot based on cognitive behavioral therapy and natural language processing (NLP) techniques for relieving people’s depression and anxiety.

Applying cutting-edge AI technologies for SID inevitably comes with privacy issue and ethical concerns. Linthicum et al. put forward three ethical issues, including the influence of bias on machine learning algorithms, the prediction

on time of suicide act, and ethical and legal questions raised by false positive and false negative prediction. It is not easy to answer ethical questions for AI as these require algorithms to reach a balance between competing values, issues, and interests. AI has been applied to solve many challenging social problems. Detection of suicidal ideation with AI techniques is one of the potential applications for social good and should be addressed to improve people's wellbeing meaningfully. The research problems include feature selection on tabular and text data and representation learning on natural language. Many AI-based methods have been applied to classify suicide risks. However, there remain some challenges. There are a limited number of benchmarks for training and evaluating SID. AI-powered models, sometimes, learn statistical clues but fail to understand people's intentions.

Moreover, many neural models are lack of interpretability. This survey reviews SID methods from the perspective of AI and machine learning and specific domain applications with social impact. The categorization from these two perspectives is shown in Fig. 1. This article provides a comprehensive review of the increasingly important field of SID with machine learning methods. It proposes a summary of current research progress and an outlook of future work. The contributions of our survey are summarized as follows. 1) To the best of our knowledge, this is the first survey that conducts a comprehensive review of SID, its methods, and its applications from a machine learning perspective. 2) We introduce and discuss the classical content analysis and modern machine learning techniques, plus their application to questionnaires, EHR data, suicide notes, and online social content. 3) We enumerate

existing and less explored tasks and discuss their limitations.

We also summarize existing data sets and provide an outlook of future research directions in this field. The remainder of this article is organized as follows. Methods and applications are introduced and summarized in Sections II and III, respectively. Section IV enumerates specific tasks and some data sets. Finally, we have a discussion and propose some future directions in Section V.

1.2 Purpose:

Consequences can be severe and risky since such lousy information is often engaged in some form of social cruelty, leading to rumors or even mental damage. Research shows that there is a link between cyber bullying and suicide. Victims overexposed to too many negative messages or events may become depressed and desperate; even worse, some may commit suicide. The reasons that people commit suicide are complicated. People with depression are highly likely to commit suicide, but many without depression can also have suicidal thoughts.

1.3 Scope:

Online social networks (OSNs), such as Facebook,¹ Twitter,² and LinkedIn,³ has created a fruitful environment for the spread of positive information. However, the high openness and autonomy of the OSNs also enable the spread of negative information, such as unsubstantiated rumors, conspiracy theories, and other forms of misinformation.

1.4 Motivation:

According to the American Foundation for Suicide Prevention (AFSP), suicide factors fall under three categories: health factors, environmental factors, and historical factors. Ferrari et al. found that mental health issues and substance use disorders are attributed to the factors of suicide.

1.5 Overview:

Detection of suicidal ideation with AI techniques is one of the potential applications for social good and should be addressed to improve people's wellbeing meaningfully. The research problems include feature selection on tabular and text data and representation learning on natural language. Many AI-based methods have been applied to classify suicide risks. However, there remain some challenges. There are a limited number of benchmarks for training and evaluating SID. AI-powered models, sometimes, learn statistical clues but fail to understand people's intentions.

2. RELATED WORKS

Opinion mining involves several important tasks, including sentiment polarity and intensity assignment. Polarity assignment is concerned with determining whether a text has a positive, negative, or neutral semantic orientation. Sentiment intensity assignment looks at whether the positive/negative sentiments are mild or strong. Given the two phrases "I don't like you" and "I hate you," both would be assigned a negative semantic orientation but the latter would be considered more intense. Effectively classifying sentiment polarities and intensities entails the use of classification methods applied to linguistic features. While several classification methods have been employed for opinion mining, Support Vector Machine (SVM) has outperformed various techniques including Naive Bayes, Decision Trees, Winnow, etc. . The most popular class of features used for opinion mining is n-grams. Various n-gram categories have attained state-of-the-art results. Larger n-gram feature sets require the use of feature selection methods to extract appropriate attribute subsets. Next, we discuss these two areas: n-gram features and feature selection techniques used for Author profiling.

2.1 N-GRAM FEATURES FOR AUTHOR PROFILING

N-gram features can be classified into two categories: fixed and variable. Fixed n-grams are exact sequences occurring at either the character or token level. Variable n-grams are extraction patterns capable of representing more sophisticated linguistic phenomena. A plethora of fixed and variable n-grams have been used for opinion mining, including word, part-of-speech (POS), character, legomena, syntactic, and semantic n-grams. Word n-grams include bag-of-words (BOWs) and higher order word n-grams (e.g., bigrams, trigrams). Word n-grams have been used effectively in several studies [28]. Typically, unigrams to trigrams are used, though 4-grams have also been employed. Word n-grams often provide a feature set foundation, with additional feature categories added to them. Given the pervasiveness of adjectives and adverbs in opinion-rich text, POS tag, n-grams are very useful for sentiment classification. Additionally, some studies have employed word plus part-of-speech (POS Word) n-grams. These n-grams consider a word along with its POS tag in order to overcome word-sense disambiguation in situations where a word may otherwise have several senses . For example, the phrase "quality of the" can be represented with the POS Word trigram "quality-noun of prep the-det." Character n-grams are letter sequences. For example, the word "like" can be represented with the following two and three letter sequences "li, ik, ke, lik, ike." While character n-grams were previously used mostly for style classification, they have recently been shown to be useful in related affect classification research attempting to identify emotions in text. Legomena n-grams are collocations that replace once (hapax legomena) and twice occurring words (dis

legomena) with “HAPAX” and “DIS” tags [2], [38].

2.2 FEATURE SELECTION FOR AUTHOR PROFILING

Prior sentiment classification studies have placed limited emphasis on feature selection techniques, despite their benefits. Feature selection can potentially improve classification accuracy, narrow in on a key feature subset of sentiment discriminators, and provide greater insight into important class attributes. There are two categories of feature selection methods, both of which have been used in prior Author profiling work: univariate and multivariate.

Univariate methods consider attributes individually. Examples include information gain, chi-squared, log likelihood, and occurrence frequency. Although univariate methods are computationally efficient, evaluating individual attributes can also be disadvantageous since important attribute interactions are not considered. It is also easier to interpret the contribution of individual attributes using univariate methods. Most opinion mining studies have used univariate feature selection methods such as minimum frequency thresholds and the log-likelihood ratio [12], [27], [39]. Information gain (IG) [44], [45] has also been shown to work well for various text categorization tasks, including Author profiling [3]. Tsutsumi et al. [35] used the Chi-Squared test to select features for text sentiment classification.

2.3 OTHER FEATURE SELECTION METHODS

2.4 In addition to prior sentiment feature selection methods, it is important to briefly discuss multivariate and hybrid methods used in related tasks. Principal component analysis (PCA) has been used considerably for dimensionality reduction in text style classification

problems. Recently, many powerful dimensionality reduction techniques have also been applied to non text feature selection problems

3. EXISTING SYSTEM

Traditional suicide detection relies on clinical methods, including self-reports and face-to-face interviews. Venek *et al.* [9] designed a five-item ubiquitous questionnaire for the assessment of suicidal risks and applied a hierarchical classifier on the patients’ response to determine their suicidal intentions. Through face-to-face interaction, verbal and acoustic information can be utilized. Scherer [23] investigated the prosodic speech characteristics and voice quality in a dyadic interview to identify suicidal and non suicidal juveniles. Other clinical methods examine the resting state heart rate from converted sensing signals and classify the functional magnetic resonance imaging-based neural representations of death- and life-related word and event-related instigators converted from EEG signals. Another aspect of clinical treatment is the understanding of the psychology behind suicidal behavior, which, however, relies heavily on the clinician’s knowledge and face-to-face interaction. Suicide risk assessment scales with clinical interview can reveal informative cues for predicting suicide conducted an interview and survey study in Weibo, a Twitter-like service in China, to explore the engagement of suicide attempters with intervention by direct messages.

3.1 Disadvantages:

In the existing work, the system is Traditional suicide detection which relies on clinical methods, including self-reports and face-to-face interviews.

This system is analyzed word frequencies in suicide notes using a fuzzy cognitive map to discern causality which is less effective.

4. PROPOSED SYSTEM

To the best of our knowledge, this is the first survey that conducts a comprehensive review of SID, its methods, and its applications from a machine learning perspective.

The proposed system introduces and discusses the classical content analysis and modern machine learning techniques, plus their application to questionnaires, EHR data, suicide notes, and online social content.

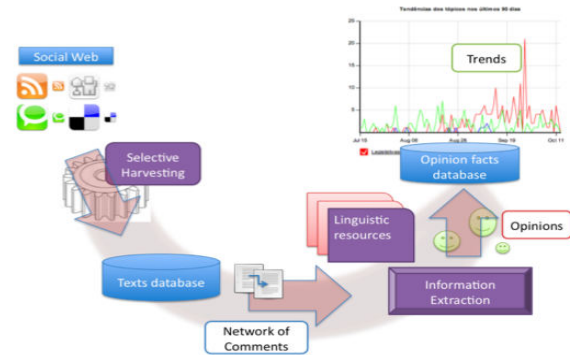
The proposed system enumerates existing and less explored tasks and discusses their limitations. We also summarize existing data sets and provide an outlook of future research directions in this field.

4.1 Advantages

The popularization of machine learning has facilitated research on SID from multimodal data and provided a promising way for effective early warning.

2) Massive data mining and machine learning algorithms have achieved remarkable outcomes by using DNNs.

5. ARCHITECTURE



6. Modules Description:

Tweet Server

In this module, the Server has to login by using valid user name and password. After login successful he can perform some operations such as View All Users, Add Filter, View All Friend Request and Response, View All Users Tweets, View Tweets All Topic & Comments, View All Suicide-related and non Suicide-related Posts, View Suicide-related posts Results, View Tweet Topics Rank Results.

Friend Request & Response

In this module, the admin can view all the friend requests and responses. Here all the requests and responses will be displayed with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then the status will be changed to accepted or else the status will remain as waiting.

User

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user

can perform some operations like View My Profile, Search Friends And Request, Friend Requests By Me, Friend Requests By Others, All My Friends, View My Friends Tweets and Re Tweet, Create Tweets, All My Tweets with Ranks.

7 CONCLUSION AND FUTURE ENHANCEMENTS

Suicide prevention remains an essential task in our modern society. Early detection of suicidal ideation is an important and effective way to prevent suicide. This survey investigates existing methods for SID from a broad perspective that covers clinical methods, such as patient–clinician interaction and medical signal sensing; textual content analysis, such as lexicon-based filtering and word cloud visualization; feature engineering, including tabular, textual, and affective features; and deep learning-based representation learning, such as CNN and LSTM-based text encoders. Four main domain-specific applications on questionnaires, EHRs, suicide notes, and online user content are introduced. Psychological experts have conducted most work in this field with statistical analysis and computer scientists with feature engineering-based machine learning and deep learning-based representation learning. Based on current research, we summarized existing tasks and further proposed new possible tasks. Last but not least, we discuss some limitations of current research and propose a series of future directions, including utilizing emerging learning techniques, interpretable intention understanding, temporal detection, and proactive conversational intervention. Online social content is very likely to be the main channel for SID in the future. Therefore, it is essential to develop new methods, which can heal the schism between clinical mental health detection and automatic machine detection, to detect

online texts containing suicidal ideation in the hope that suicide can be prevented.

9. BIBLIOGRAPHY

- [1] S. Belguith, N. Kaaniche, A. Jemai, M. Laurent, and R. Attia, “PAbAC: A privacy preserving attribute based framework for fine grained access control in clouds,” in Proc. 13th Int. Joint Conf. e-Bus. Telecommun., 2016, pp. 133–146.
- [2] US Department of Health and Human Services Report. Accessed: Jan. 7, 2018.
[Online]. Available: https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf
- [3] P. J. Denning and D. E. Denning, “Discussing cyber attack,” Commun. ACM, vol. 53, no. 9, pp. 29–31, 2010.
- [4] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, “Inferring internet denial-of-service activity,” ACM Trans. Comput. Syst., vol. 24, no. 2, pp. 115–139, 2006.
- [5] K. G. Anagnostakis, S. Sidiroglou, P. Akritidis, K. Xinidis, E. P. Markatos, and A. D. Keromytis, “Detecting targeted attacks using shadow honeypots,” in Usenix Secur., 2005.
- [6] S. Balram and M. Wilscy, “User traffic profile for traffic reduction and effective botnet C&C detection,” IJ Netw. Secur., vol. 16, no. 1, pp. 46–52, 2014.
- [7] G. Fedynyshyn, M. C. Chuah, and G. Tan, “Detection and classification of different botnet C&C channels,” in Autonomic and Trusted Computing (Lecture Notes in Computer Science), vol. 6906, J. M. A. Calero, L. T. Yang, F. G. MÆrmol, L. J. G. Villalba, A. X. Li, and Y. Wang, Eds. Berlin, Germany: Springer, 2011

[8] P. Wurzinger, L. Bilge, T. Holz, J. Goebel, C. Kruegel, and E. Kirda, “Automatically generating models for botnet detection,” in *Computer Security—ESORICS (Lecture Notes in Computer Science)*, vol. 5789, M. Backes and P. Ning, Eds. Berlin, Germany: Springer, 2009.

[9] F. Giroire, J. Chandrashekar, N. Taft, E. Schooler, and D. Papagiannaki, “Exploiting temporal persistence to detect covert botnet channels,” in *Recent Advances Intrusion Detection (Lecture Notes in Computer Science)*, vol. 5758, E. Kirda, S. Jha, and D. Balzarotti, Eds. Berlin, Germany: Springer, 2009.

[10] G. Gu, J. Zhang, and W. Lee, “BotSniffer: Detecting botnet command and control channels in network traffic,” in *Proc. 15th Annu. Netw. Distrib. Syst. Secur. Symp.* Dayton, OH, USA: Wright State Univ., 2008.

LIVER DISEASE PREDICTION SYSTEM USING MACHINE LEARNING TECHNIQUES

Peluri Govardhan (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

In this paper we are going to discuss how to predict risk of liver disease for a person, based on the blood test report results of the user. In this paper, the risk of liver disease was predicted using various machine learning algorithms. The final output was predicted based on the most accurate machine learning algorithm. Based on the accurate model we designed a system which asks a person to enter the details of his/her blood test report. Then the system uses the most accurate model which is trained to predict, whether a person has risk of liver disease or not.

Keywords—Machine learning, Liver disease, Confusion matrix, Use case diagram, back propagation algorithm

1. INTRODUCTION

With a growing trend of sedentary and lack of physical activities, diseases related to liver have become a common encounter nowadays. In rural areas the intensity is still manageable, but in urban areas, and especially metropolitan areas the liver disease is a very common sighting nowadays. Liver diseases cause millions of deaths every year. Viral hepatitis alone causes 1.34 million deaths every year. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. An early diagnosis of liver problems will increase patient's survival rate. Liver failures are at high rate of risk among Indians. It is expected that by 2025 India may become the World Capital for Liver Diseases. The widespread occurrence of liver infection in India is contributed due to deskbound lifestyle, increased alcohol consumption and smoking. There are about 100 types of liver infections. With

such alarming figures, it is necessary to have a concern towards tackling these diseases. After all, we cannot expect a developed and prosperous nation, with unhealthy youths. In this project we have taken UCI ILPD Dataset which contains 10 variables that are age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos and contains 415 as liver disease patients and 167 as non liver disease patients. As we got through the next parts of this paper we will explain what process as taken place for the selection of best model and building neccessary sytem for the prediction of liver disease.

The major outcomes that can be expected through this project are:

- Increased convenience for predicting a liver disease
- Reduction in number of deaths due to liver diseases
- More accurate diagnosis of liver disease by the doctors

2. LITERATURE SURVEY

A. Naive Bayes Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. They are among the simplest Bayesian network models, but coupled with kernel density estimation, they can achieve higher accuracy levels. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.

B. ANN Artificial Neural networks (ANN) or neural networks are computational algorithms. They intend to simulate the behavior of biological systems composed of "neurons". ANNs are computational models inspired by an animal's central nervous systems. They are capable of machine learning as well as pattern recognition. These are present as systems of interconnected "neurons" which can compute values from inputs. A neural network is an oriented graph. It consists of nodes which in the biological analogy represent neurons, connected by arcs. It corresponds to dendrites and synapses. Each arc is associated with a weight while at each node. To do the prediction, we need to apply the values received as input by the node and define activation

function along the incoming arcs, adjusted by the weights of the arcs. An ANN is trained based on backpropagation algorithm.

3. SYSTEM DESIGN

A. Proposed system

The system being proposed here uses concept of machine learning, and the models are first trained, then tested. Finally the most accurate model will predict the final result. At first, the system asks you to enter your details including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Values of last eight parameters mentioned here, can be known by blood test report of the user. After taking these inputs from the user, the system compares the data input with the training dataset of most accurate model and then predicts the result accordingly as risk or no risk.

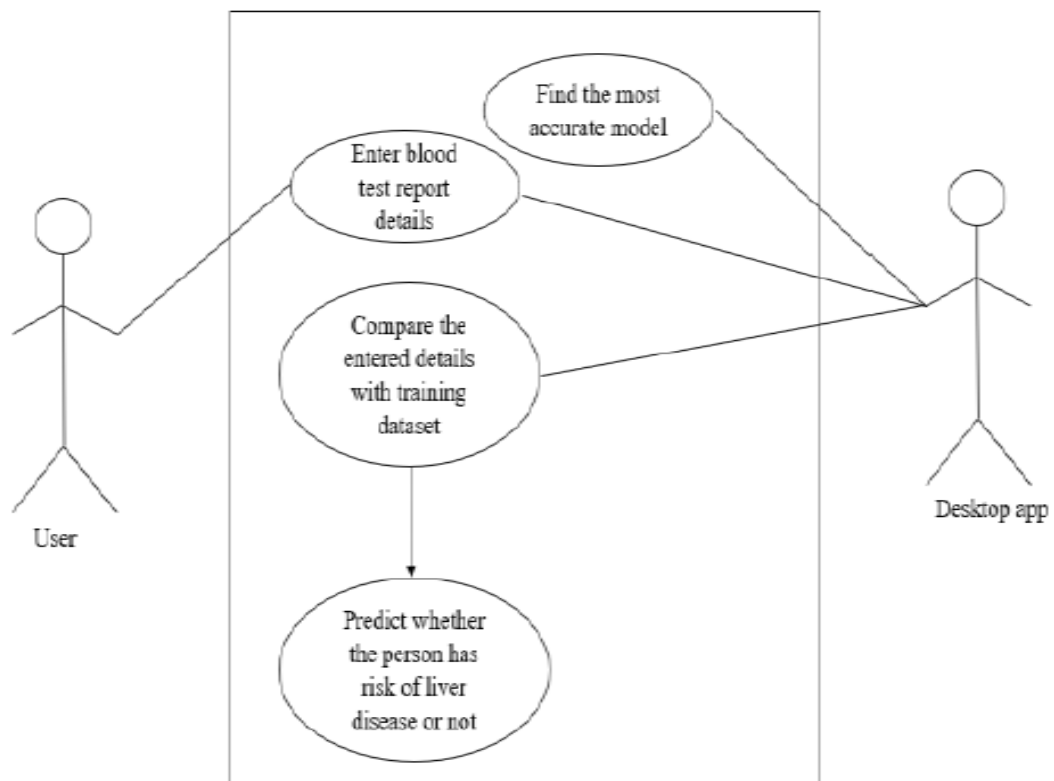


Fig1: Use case diagram

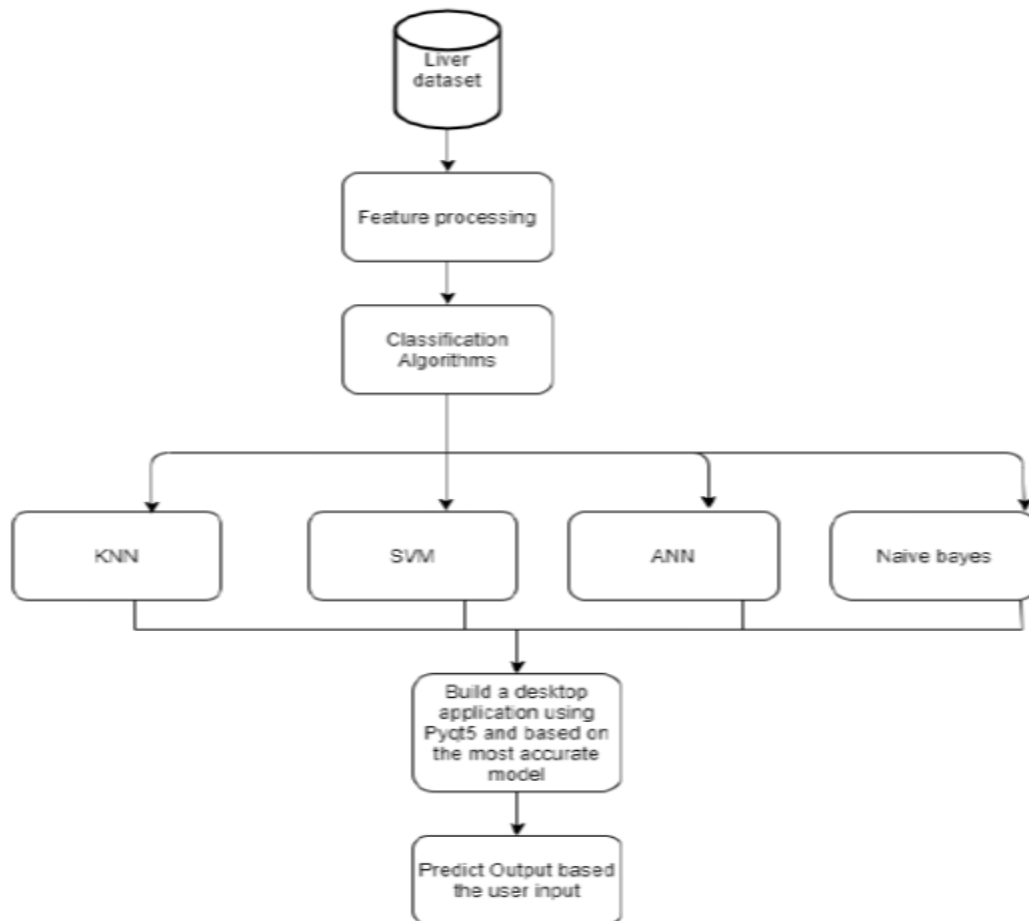


Fig 2: Work flow diagram

4. SYSTEM DESIGN

A. Proposed system The system being proposed here uses concept of machine learning, and the models are first trained, then tested. Finally the most accurate model will predict the final result. At first, the system asks you to enter your details including age, gender, total Bilirubin, direct Bilirubin, total proteins, albumin, A/G ratio, SGPT, SGOT and Alkphos. Values of last eight parameters mentioned here, can be known by blood test report of the user. After taking these inputs from the user, the system compares the data input with the training dataset of most accurate model and then predicts the result accordingly as risk or no risk. The system has following advantages: 1. *No medical expertise required*: You don't need to have any knowledge of medical science and liver diseases to predict the liver disease using this application. All you need to do is enter the details being asked, which are already

present in the blood test report(some like age, gender are already known) and then you will get the results of prediction.

2. *High accuracy*: The system predicts the results with 100 % accuracy for the dataset that we have used while creating this application. While the accuracy might be different in some cases, it will still be high enough to be trustworthy at a large scale.

3. *Immediate results*: The results here are predicted within seconds of entering the details. You don't need to wait for a doctor to come, unlike in traditional method.

5. CONCLUSION AND FUTURE WORKS

Diseases related to liver and heart are becoming more and more common with time. With continuous technological advancements, these are only going to increase in the future. Although people are becoming more conscious of health nowadays and are joining yoga classes, dance classes; still the sedentary lifestyle and luxuries that are continuously being introduced and enhanced; the problem is going to last long. So, in such a scenario, our project will be extremely helpful to the society. With the dataset that we used for this project, we got 100 % accuracy for SVM model, and though it might be difficult to get such accuracies with very large datasets, from this project's results, one can clearly conclude that we can predict the risk of liver diseases with accuracy of 90 % or more.

Today almost everybody above the age of 12 years has smartphones with them, and so we can incorporate these solutions into an android app or ios app. Also it can be incorporated into a website and these app and website will be highly beneficial for a large section of society.

6. REFERENCES

- [1] Biomarkers for prediction of liver fibrosis in patients with chronic alcoholic liver disease written by Sylvie Naveau and Bruno Runyard.
- [2] Strategic analysis in prediction of liver disease using classification algorithms written by Piyush Kr Shukla and Binish Khan.

[3] Software based prediction of liver disease with feature selection and classification techniques witten Jagdeep Singh, Sandeep Bagga and Ranjodh Kaur.

[4] Liver disease prediction using SVM and Naïve Bayes algorithm written by S Dhayanand.

[5] Prediction and analysis of liver diseases using data mining written Shambel Kefelgen, Pooja Kamat.



AGRICULTURAL CROP RECOMMENDATIONS BASED ON PRODUCTIVITY AND SEASON

Penmatsa Gayathri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract— As we know the fact that, India is the second largest population country in the world and majority of people in India have agriculture as their occupation. Farmers are growing same crops repeatedly without trying new variety of crops and they are applying fertilizers in random quantity without knowing the deficient content and quantity. So, this is directly affecting on crop yield and also causes the soil acidification and damages the top layer. So, we have designed the system using machine learning algorithms for betterment of farmers. Our system will suggest the best suitable crop for particular land based on content and weather parameters. And also, the system provides information about the required content and quantity of fertilizers, required seeds for cultivation. Hence by utilizing our system farmers can cultivate a new variety of crop, may increase in profit margin and can avoid soil pollution.

Keywords—Machine Learning, Crop prediction, Decision tree, SVM, Rainfall prediction, Crop recommendation;

1. INTRODUCTION

Agriculture is one of the important occupation practiced in India. It is the broadest economic sector and plays a most important role in the overall development of the country. More than 60% of the land in the country is used for agriculture in order to suffice the needs of 1.3 billion people. Thus adopting new agriculture technologies is very important. This will lead the farmers of our country towards profit [1]. Prior crop prediction and yield prediction was performed on the basis of farmers experience on a particular location. They will prefer the prior or neighborhood or more trend crop in the surrounding region only for their land and they don't have enough of knowledge about soil nutrients content such as nitrogen, phosphorus, potassium in the land. Being this as the current situation without the rotation of the crop and apply an inadequate amount of nutrients to soil it

leads to reduce in the yield and soil pollution (soil acidification) and damages the top layer. Considering all these problems takes into the account we designed the system using a machine learning for betterment of the farmer. Machine learning (ML) is a game changer for agriculture sector. Machine learning is the part of artificial intelligence, has emerged together with big data technologies and high-performance computing to create new opportunities for data intensive science in the multi-disciplinary agritech domain. In the Agriculture field machine learning for instance is not a mysterious trick or magic, it is a set of well define model that collect specific data and apply specific algorithms to achieve expected results [7]. The designed system will recommend the most suitable crop for particular land. Based on weather parameter and soil content such as Rainfall, Temperature, Humidity and pH.

They are collected from V C Farm Mandya, Government website and weather department. The system takes the required input from the farmers or sensors such as Temperature, Humidity and pH. This all inputs data applies to machine learning predictive algorithms like Support Vector Machine (SVM) [5] and Decision tree [6] to identify the pattern among data and then process it as per input conditions. The system recommends the crop for the farmer and also recommends the amount of nutrients to be add for the predicted crop. The system has some other specification like displaying approximated yield in q/acre, required seed for cultivation in kg/acre and the market price of the crop.

2. LITERATURE SURVEY

Ashwani kumar Kushwaha[2] describes crop yield prediction methods and a suggest suitable crop so that it will improve the profit for the farmer and quality of the agriculture sector. In this paper for crop yield prediction they obtain large volume data, it's been called as big data (soil and weather data) using Hadoop platform and agro algorithm. Hence based repository data will predict the suitability crop for particular condition and improvement crop quality. Girish L [3] describe the crop yield and rain fall prediction using a machine learning method. In this paper they gone through a different machine learning approaches for the prediction of rainfall and crop yield and also mention the efficiency of a different machine learning algorithm like liner regression, SVM, KNN method and decision tree. In that algorithm they conclude that SVM have the highest efficiency for rainfall prediction. Rahul katarya [4] describes the different machine

learning methods used for accelerating crop yield. In this paper they gone through different artificial intelligence techniques such as machine learning algorithm, big data analysis for precision agriculture. They explain about crop recommender system using KNN, Ensemble-based Models, Neural networks, ...etc.

3. PROPOSED SYSTEM

The Proposed system will predict the most suitable crop for particular land based on soil contents and weather parameters such as Temperature, Humidity, soil PH and Rainfall.

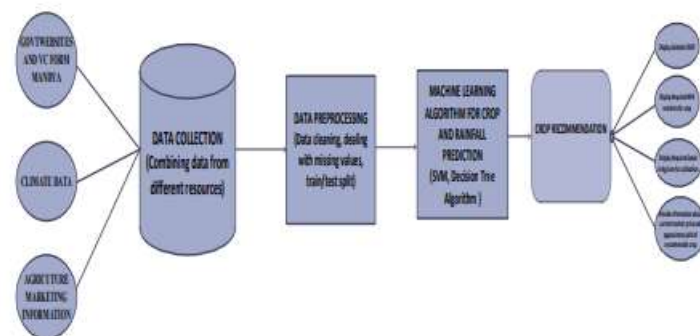


Figure 1. Architecture of the proposed system.

The Architecture of the proposed system consists of various blocks as shown in the fig (1) as follows

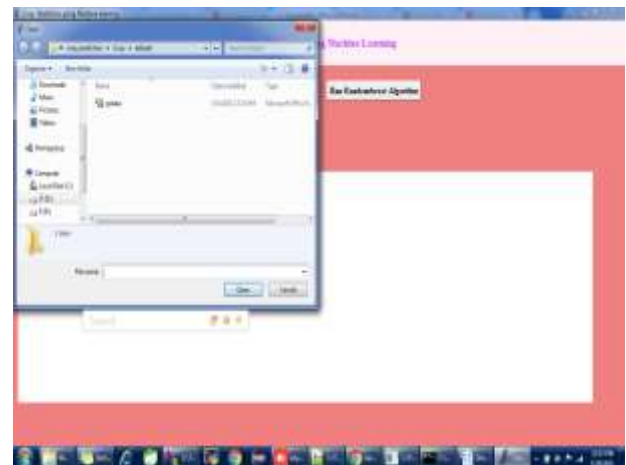
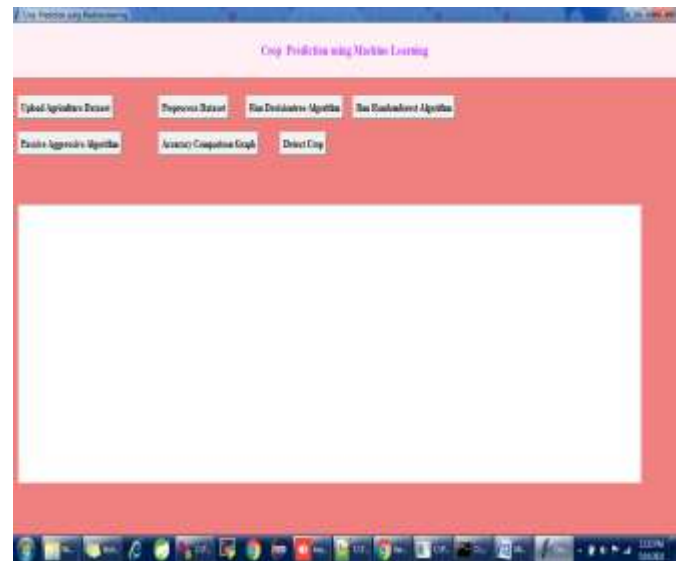
3.1. Data Collection: - Data collection is the most efficient method for collecting and measure the data from different resources like govt websites, VC Form Mandya, APMC website etc. To get an approximate dataset for the system. This dataset must contain the following attributes i)Soil PH ii) Temperature iii) Humidity iv) Rainfall v) Crop data vi) NPK values, those parameters will consider for crop prediction. For the annual rainfall prediction, we collect previous year rainfall data.

3.2. Data Preprocessing: - After collecting datasets from various resources. Dataset

must be preprocessing before training to the model. The data preprocessing can be done by various stages, begins with reading the collected dataset the process continues to data cleaning. In data cleaning the datasets contain some redundant attributes, those attributes are not considering for crop prediction. So, we have to drop unwanted attributes and datasets containing some missing values we need to drop these missing values or fill with unwanted nan values in order to get better accuracy. Then define the target for a model. After data cleaning the dataset will be split into training and test set by using sklearn library.

3.3. Machine Learning Algorithm for Prediction: - Machine learning predictive algorithms has highly optimized estimation has to be likely outcome based on trained data. Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future. In our system we used supervised machine learning algorithm having subcategories as classification and regression. Classification algorithm will be most suitable for our system. ➤ Rainfall prediction: -SVM algorithm. ➤ Crop prediction: - Decision tree algorithm

4. RESULTS





5. CONCLUSION

Presently our farmers are not effectively using technology and analysis, so there may be a chance of wrong selection of crop for cultivation that will reduce their income. To reduce those type of loses we have developed a farmer friendly system with GUI, that will predict which would be the best suitable crop for particular land and this system will also provide information about required nutrients to add up, required seeds for cultivation, expected yield and market price. So, this makes the farmers to take right decision in selecting the crop for cultivation such that agricultural sector will be developed by innovative idea.

FUTURE SCOPE: We have to collect all required data by giving GPS locations of a land and by taking access from Rain forecasting system of by the government, we can predict crops by just giving GPS location. Also, we can develop the model to avoid over and under crisis of the food.

6. REFERENCES

[1] Prof. D.S. Zingade ,Omkar Buchade ,Nilesh Mehta ,Shubham Ghodekar

,Chandan Mehta “Crop Prediction System using Machine Learning”.

[2] Ashwani kumar Kushwaha, Swetabhattacharya “crop yield prediction using agro algorithm in hatooop”.

[3] Girish L, Gangadhar S, Bharath T R, Balaji K S, Abhishek K T “Crop Yield and Rainfall Prediction in Tumakuru District using Machine Learning”.

[4] Rahul Katarya, Ashutosh Raturi, Abhinav Mehndiratta, Abhinav Thapper “Impact of Machine Learning Techniques in Precision Agriculture”.

[5] Pijush Samui, Venkata Ravibabu Mandla, Arun Krishna and Tarun Teja “Prediction of Rainfall Using Support Vector Machine and Relevance Vector Machine”.

[6] Himani Sharma, Sunil Kumar “A Survey on Decision Tree Algorithms of Classification in Data Mining”.

[7] Pavan Patil, Virendra Panpatil, Prof. Shrikant Kokate “Crop Prediction System using Machine Learning Algorithms”.

TRUSTWORTHINESS ASSESSMENT OF USERS IN SOCIAL REVIEWING SYSTEMS

Pennedi Sruthi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Social Networks represent a cornerstone of our daily life, where the so-called social reviewing systems (SRSs) play a key role in our daily lives and are used to access data typically in the form of reviews. Due to their importance, social networks must be trustworthy and secure, so that their shared information can be used by the people without any concerns, and must be protected against possible attacks and misuses. One of the most critical attacks against the reputation system is represented by mendacious reviews. As this kind of attacks can be conducted by legitimate users of the network, a particularly powerful solution is to exploit trust management, by assigning a trust degree to users, so that people can weigh the gathered data based on such trust degrees. Trust management within the context of SRSs is particularly challenging, as determining incorrect behaviors is subjective and hard to be fully automatized. Several attempts in the current literature have been proposed; however, such an issue is still far from being completely resolved. In this study, we propose a solution against mendacious reviews that combines fuzzy logic and the theory of evidence by modeling trust management as a multicriteria multiexpert decision making and exploiting the novel concept of time-dependent and content-dependent crown consensus. We empirically proved that our approach outperforms the main related works approaches, also in dealing with sockpuppet attacks.

1. INTRODUCTION

AS WELL known, the online social networks [1] are Internet-enabled applications used by people to establish social relations with the other individuals sharing similar personal interests

and/or activities. Apart from exchanging personal data, such as photographs or videos, mainly all these applications allow their users to share comments and opinions on specific topics, so as to suggest objects or places of interest (e.g., Trip Advisor, Foursquare, etc.) or to provide social environments able to facilitate particular tasks (e.g., the search of a job as in LinkedIn, the answer to research questions as in Research Gate, purchases on Amazon, etc.). Due to this comment/opinion sharing, these social applications, which we will refer to as social reviewing systems (SRSs) have been extensively used when people need to make daily decisions, increasing their popularity. As a concrete example, most of us access to a preferable SRS before choosing a restaurant or buying something so as to get reviews and feedback. People are progressively and symbiotically dependent on them as proved by the advanced opinion modeling and analysis, exploiting the impact of neighbors on user preferences or approaching the existing information overload in SRS, such as [2], [3]. For this reason, the trustworthiness of SRS is particularly important, and a key concern for effective opinion dynamics and trust propagation within a community of users [4]. In fact, SRSs suffer from forged messages and camouflaged/fake users that are able to avoid individuals take the right decision. This may raise several issues about privacy and security [5], mainly due to the fact that several personal and sensitive information are shared, and leaked, throughout SRS [6], [7], and that a person may choose to hide its true self and intentions behind a totally false virtual identity [8] or a Bot (short for software robots) may mimic human behavior in SRS [9]. In addition, threats in SRS, such as data leaks, phishing bait, information tampering, and so on, are never limited to a given social actor, but spread across the network like an infection by obtaining victims among the friends of the infested actors. So, an SRS provider needs to provide proper protection means to guarantee its trustworthiness.

Some works in the current literature, such as [10], mostly deal only with forging messages as this can be easily resolved by using cryptography. However, the second kind of malicious behavior caused by camouflaged/fake users is still an open issue. During the last decade, several solutions have been proposed in order to deal with the problem of camouflaged/fake users [11]–[13]. The issue of providing privacy has led to the adoption of access control means, while counteracting forging nodes/identities and social links/connections demanded authentication of users and exchanged messages [14], [15]. Mostly, such mechanisms

aim at approaching external attackers or intruders, while thwarting legitimate participants in the SRS acting in a malicious way is extremely challenging. A naive way to protect against malicious individuals is to have users being careful when choosing with whom to have a relationship. Two users in social networks may have various kinds of relationships: 1) in Face book-like systems users can indicate others as “friends,” or 2) in Instagram-like systems a user can “follow” others. However, users are typically not so careful when accepting received joining requests, and selecting other users to be connected with is typically extremely difficult (as malicious users are also experts in camouflaging themselves). Despite the relationships among the social actors within an SRS should be based on the direct knowledge in the real life of the people behind such actors (such as former classmates, colleagues, or member of the same family or group of friends), the majority of the relationships are typically made without such a face-toface knowledge but among users that have never been met in person. *Trust management* is among the most popular solution to fight against such inside attackers [16]. It consists to assign a “trust” value to users based on the direct analysis of their behaviors or indirect trust relationship among social actors. To this aim, it is a soft secure measure implying the revocation of a social link toward those actors with a low trust value, or to strengthen the protection measures for those actors exhibiting a low trust degree, by limiting the data/functionalities that they can have access to. Despite being a powerful protection means [17], trust management is not explicitly provided by the main SRS platforms, due to the issues related to its automatic computation.

2. EXISTING SYSTEM

Yu *et al.* [38] described an approach for computing user trustworthiness by leveraging on the “familiarity” and “similarity” concepts and considering the influence of user actions on the trustworthiness computation. The aim of this methodology is to detect malicious users-based also on a security queue to record users’ historical trust information. Afterward, Yu *et al.* [39] proposed an approach based on deep learning techniques in conjunction with user trustworthiness characterization for configuring privacy settings for social image sharing. In addition, a two-phase trust-based approach based on deep learning techniques has also been proposed by Deng *et al.* [40] for social network recommendation, so as to determine the users’

interests and their trusted friends' interests together with the impact of community effect for recommendations.

Rayana and Akoglu [27] presented a system, namely, *SpEagle*, that uses metadata (i.e., text, timestamp, and rating) in conjunction with relational data to spot suspicious users and reviews.

Other related approaches exploit reviews' evaluation for detecting and/or characterizing spam in social media. Shehnepoor *et al.* [28] proposed a framework named *NetSpam* that models reviews in online social media, as a case of heterogeneous networks, by using spam features for detection purposes. Ye *et al.* [41] described an approach based on the temporal analysis by monitoring selected indicative signals of opinion spams over the time, for detecting and characterizing abnormal events in real time.

A system based on four integrated components, specifically: 1) a reputation-based component; 2) a credibility classifier engine; 3) a user experience component; and 4) a featureranking algorithm, has been designed and implemented by

Alrubian *et al.* [42] for assessing information credibility on Twitter. In [43], the *CommTrust* framework has been introduced for trust evaluations by mining feedback comments. More in detail, it is based on a multidimensional trust model for computing reputation scores from user feedback comments, which are analyzing combining natural language processing techniques, opinion mining, and topic modeling.

Furthermore, another framework, namely, *LiquidCrowd*, has been proposed by Castano *et al.* [44] exploiting consensus and trustworthiness techniques for managing the execution of collective tasks. Kumar *et al.* [45] proposed a system, namely, *FairJudge*, to identify fraudulent users based on the mutually recursive

definition of the following three metrics: 1) the user trustworthiness in rating products; 2) the rating reliability; and 3) the goodness of a product. Moreover, Kumar *et al.* [46] described a system for identifying fraudulent users based on six axioms to define the interdependency among three intrinsic quality metrics concerning a user, reliability and goodness of a product by combining network and behavior properties. Hooi *et al.* [47] developed an algorithm, called

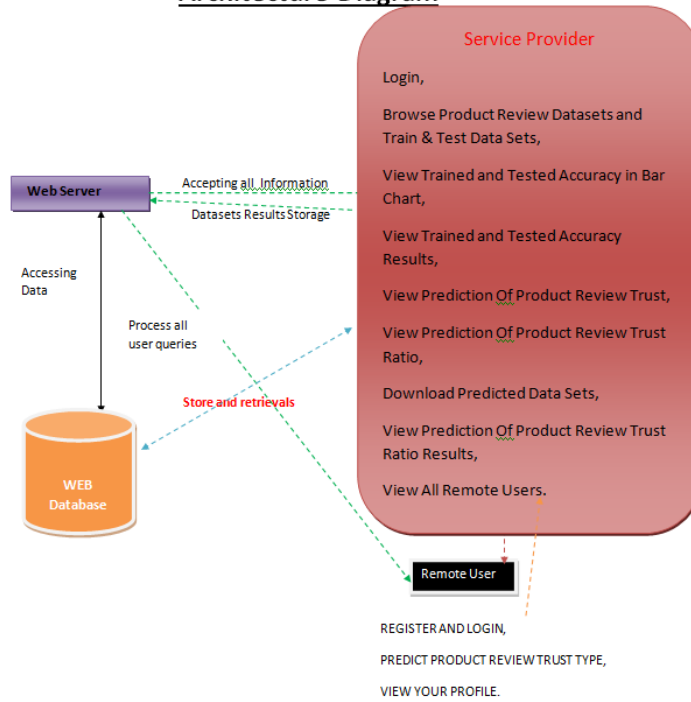
FraudAR, aiming at being resistant to the camouflage attacks, for identifying fake reviews and users. Furthermore, *Birdnest*, an approach combining Bayesian model of user rating behavior and a likelihood-based suspiciousness metric [normalized expected surprise total (NEST)], has been proposed in [48].

Liu *et al.* [32] investigated the sockpuppet attacks on reviewing A system based on four integrated components, specifically: 1) a reputation-based component; 2) a credibility classifier engine; 3) a user experience component; and 4) a feature ranking algorithm, has been designed and implemented by Alrubian *et al.* [42] for assessing information credibility on Twitter. In [43], the *CommTrust* framework has been introduced for trust evaluations by mining feedback comments. More in detail, it is based on a multidimensional trust model for computing reputation scores from user feedback comments, which are analyzing combining natural language processing techniques, opinion mining, and topic modeling. Furthermore, another framework, namely, *LiquidCrowd*, has been proposed by Castano *et al.* [44] exploiting consensus and trustworthiness techniques for managing the execution of collective tasks.

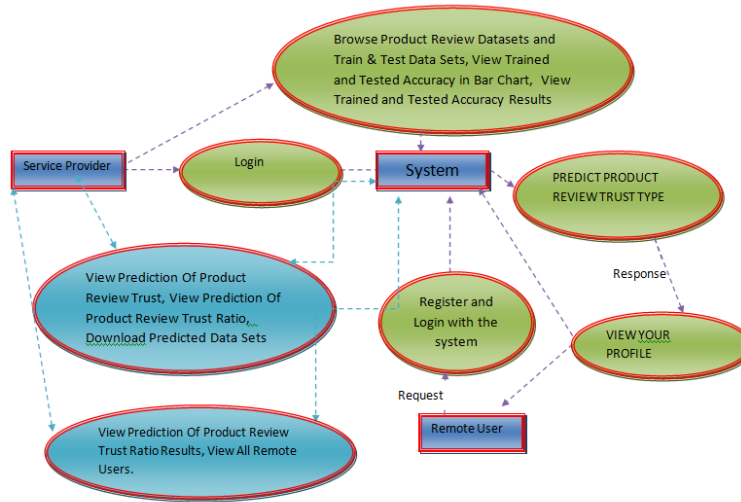
Kumar *et al.* [45] proposed a system, namely, *FairJudge*, to identify fraudulent users based on the mutually recursive definition of the following three metrics: 1) the user trustworthiness in rating products; 2) the rating reliability; and 3) the goodness of a product. Moreover, Kumar *et al.* [46] described a system for identifying fraudulent users based on six axioms to define the interdependency among three intrinsic quality metrics concerning a user, reliability and goodness of a product by combining network and behavior properties.

Hooi *et al.* [47] developed an algorithm, called *FraudAR*, aiming at being resistant to the camouflage attacks, for identifying fake reviews and users. Furthermore, *Birdnest*, an approach combining Bayesian model of user rating behavior and a likelihood-based suspiciousness metric [normalized expected surprise total (NEST)], has been proposed in [48]. Liu *et al.* [32] investigated the sockpuppet attacks on reviewing systems by proposing a fraud detection algorithm, called RTV, that introduces trusted users and also considers reviews left by verified users.

Architecture Diagram



➤ **Data Flow Diagram :**



SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

4. CONCLUSION

This study proposed a solution to the problem of trust management within the context of the social networks, where it is important to deal with the subjectivity of the detection of malicious behaviors and the need of objectivity in order to design an automatic process to assign trust degrees to users based on their activity in the social network. To this aim, we have approached the vagueness and subjectivity in the review analysis from the social network by means of the fuzzy theory. We have leveraged on the theory of evidence so as to devise a MCME-DM process to aggregate the judgments from multiple perspectives and optimize the trust estimation. We have performed a realistic experimental campaign considering the YELP and Amazon dataset and showed that aggregating the output of multiple criteria allows achieving higher accuracy in detecting malicious reviews. We have also compared our approach against the main related works in the existing literature and showed that our approach obtained better efficacy by using 80% and 100% of the considered dataset.

As future work, we plan to investigate more in detail the influence of common attacks toward a recommendation system so as to enhance the security of such a solution, in addition to the study of the privacy concerns of such systems, by considering the key legal frameworks, such as the The EU General Data Protection Regulation (GDPR). Moreover, the main critics to D-S

aggregation are to return counterintuitive results when combining unreliable evidences [61] and/or conflicting evidences from independent sources [62]. In order to improve the detection of a potential problem in the aggregation process, special formulations of the mass functions and other concepts of the D-S theory emerged over the last decade, such as the evolutionary combination rule (ECR) in [63]. We have left as future work the investigation of this approach within the context of our work.

5. REFERENCES

- [1] M. Faloutsos, T. Karagiannis, and S. Moon, "Online social networks," *IEEE Netw.*, vol. 24, no. 5, pp. 4–5, Sep/Oct. 2010.
- [2] J. Castro, J. Lu, G. Zhang, Y. Dong, and L. Martinez, "Opinion dynamics-based group recommender systems," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 48, no. 12, pp. 2394–2406, Dec. 2018.
- [3] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 50, no. 10, pp. 3804–3816, Oct. 2020.
- [4] R. Ureña, G. Kou, Y. Dong, F. Chiclana, and E. Herrera-Viedma, "A review on trust propagation and opinion dynamics in social networks and group decision making frameworks," *Inf. Sci.*, vol. 478, pp. 461–475, Apr. 2019.
- [5] Y. Xiang, E. Bertino, and M. Kutyłowski, "Security and privacy in social networks," *Concurrency Comput. Practice Exp.*, vol. 29, no. 7, 2017, Art. no. e4093.
- [6] D. Irani, S. Webb, K. Li, and C. Pu, "Modeling unintended personal information leakage from multiple online social networks," *IEEE Internet Comput.*, vol. 15, no. 3, pp. 13–19, May/Jun. 2011.
- [7] A. Nosko, E. Wood, and S. Molema, "All about me: Disclosure in online social networking profiles: The case of Facebook," *Comput. Human Behav.*, vol. 26, no. 3, pp. 406–418, 2010.
- [8] K. Krombholz, D. Merkl, and E. Weippl, "Fake identities in social media: A case study on the sustainability of the Facebook business model," *J. Service Sci. Res.*, vol. 4, no. 2, pp. 175–212, 2012.
- [9] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, 2016.

AUTOMATED DETECTING SPAMMERS IN SOCIAL MEDIA

Penumala Vamsi Krishna (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Twitter is one of the most popular micro blogging services, which is generally used to share news and updates through short messages restricted to 280 characters. However, its open nature and large user base are frequently exploited by automated spammers, content polluters, and other ill-intended users to commit various cybercrimes, such as cyber bullying, trolling, rumor dissemination, and stalking. Accordingly, a number of approaches have been proposed by researchers to address these problems. However, most of these approaches are based on user characterization and completely disregarding mutual interactions. In this paper, we present a hybrid approach for detecting automated spammers by amalgamating community based features with other feature categories, namely metadata-, content-, and interaction-based features. The novelty of the proposed approach lies in the characterization of users based on their interactions with their followers given that a user can evade features that are related to his/her own activities, but evading those based on the followers is difficult. Nineteen different features, including six newly defined features and two redefined features, are identified for learning three classifiers, namely, random forest, decision tree, and Bayesian network, on a real dataset that comprises benign users and spammers. The discrimination power of different feature categories is also analyzed, and interaction- and community-based features are determined to be the most effective for spam detection, whereas metadata-based features are proven to be the least effective.

1. INTRODUCTION

TWITTER, a microblogging service, is considered a popular online social network (OSN) with a large user base and is attracting users from different walks of life and age groups. OSNs enable users to keep in touch with friends, relatives, family members, and people with similar interests, profession, and objectives. In addition, they allow users to interact with one another and form communities. A user can become a member of an OSN by registering and providing details, such as name, birthday, gender, and other contact information. Although a

large number of OSNs exist on the web, Facebook and Twitter are among the most popular OSNs and are included in the list of the top 10 websites¹ around the worldwide.

A. OSN and the Social Spam Problem

Twitter, which was founded in 2006, allows its users to post their views, express their thoughts, and share news and other information in the form of tweets that are restricted to 280 characters. Twitter allows the users to follow their favourite politicians, athletes, celebrities, and news channels, and to subscribe to their content without any hindrance. Through *following* activity, a follower can receive status updates of subscribed account. Although Twitter and other OSNs are mainly used for various benign purposes, their open nature, huge user base, and real-time message proliferation have made them lucrative targets for cyber criminals and social bots. OSNs have been proven to be incubators for a new breed of complex and sophisticated attacks and threats, such as cyberbullying, misinformation diffusion, stalking, identity deception, radicalization, and other illicit activities, in addition to classical cyber attacks, such as spamming, phishing, and drive by download [1], [2]. Over the years, classical attacks have evolved into sophisticated attacks to evade detection mechanisms. A report² submitted to the US Securities and Exchange Commission in August 2014 indicates that approximately 14% of Twitter accounts are actually spambots and approximately 9.3% of all tweets are spam. In social networks, spambots are also known as socialbots that mimic human behaviour to gain trust in a network and then exploit it for malicious activities [3]. Such reports and findings demonstrate the extent of cyber crimes committed by spambots and how OSNs are proving to be a heaven for these bots. Although spammers are less than benign users, they are capable of affecting network structure and trust for various illicit purposes.

The main contributions of this study can be summarized as follows.

- A novel study that uses community-based features with other feature categories, including *metadata*, *content*, and *interaction*, for detecting automated spammers.
- Six new features are introduced and two existing features are redefined to design a feature set with improved discriminative power for segregating benign users and spammers. Among the six new features, one is content based, three are interaction-based, and the remaining two are community-based. Meanwhile, both redefined features are content-based. When defining

interaction-based features, focus should be on the *followers* of a user, rather than on the ones he/she is *followings*.

- A detailed analysis of the working behavior of automated spammers and benign users with respect to newly defined features. In addition, two-tailed *Z*-test statistical significance analysis is performed to answer the following question: “*is the difference between the working behavior of spammers and benign users in terms of newly defined features a random chance?*”
- A thorough analysis of the discriminating power of each feature category in segregating automated spammers from benign users.

2. EXISTING SYSTEM

Sahami et al. [14] proposed textual and non textual and domain-specific features and learned naive Bayes classifier to segregate spam emails from legitimate ones. Schafer [15] and [16] proposed metadata-based approaches to detect botnets based on compromised email accounts to diffuse mail spams. Spam campaigns on Facebook were analyzed by Gao et al. [10] using a similarity graph based on semantic similarity between posts and URLs that point to the same destination.

Furthermore, they extracted clusters from a similarity graph, wherein each cluster represents a specific spam campaign. Upon analysis, they determined that most spam sources were hijacked accounts, which exploited the trust of users to redirect legitimate users to phishing sites. In [7] and [8], honey profiles were created and deployed on OSNs to observe the behavior of spammer. Both studies presented different sets of features to discriminate benign users from spammers and evaluated them on different sets of OSNs.

Wang [17] used content and graph-based features to classify malicious and normal profiles on Twitter. In contrast to honey profiles, Wang used Twitter API to crawl the dataset. Yang et al. [12], Wang [17], and Ahmed and Abulaish [18], used content- and interaction based attributes for learning classifiers to segregate spammers from benign users on different OSNs.

Yang et al. [12] and Ahmed and Abulaish [18] analyzed the contribution of each feature to spammer detection, whereas Yang et al. [19] conducted an in-depth empirical analysis of the evasive tactics practiced by spammers to bypass detection systems. They also tested the

robustness of newly devised features. In [20], Zhu et al. used a matrix factorization technique to find the latent features from the sparse activity matrix and adopted social regularization to learn the spam discriminating power of the classifier on the Renren network, one of the most popular OSNs in China. Another spammer detection approach in social media was proposed by Tan et al. [21].

Disadvantages

There are no Hybrid techniques to classify different spam's behaviours.

There is no spambot detection techniques.

3. PROPOSED SYSTEM

In the proposed system, the system proposes a hybrid approach for detecting social spam bots in Twitter, which utilizes an amalgamation of metadata-, content-, interaction-, and community-based features. In the analysis of characterizing features of existing approaches, most network-based features are not defined using user followers and underlying community structures, thereby disregarding the fact that the reputation of user in a network is inherited from the followers (rather than from the ones user is following) and community members. Therefore, the system emphasizes the use of followers and community structures to define the network-based features of a user.

The system classifies set of features into three broad categories, namely, metadata, content, and network, wherein the network category is further classified into interaction- and community based features. Metadata features are extracted from available additional information regarding the tweets of a user, whereas content-based features aim to observe the message posting behavior of a user and the quality of the text that the user uses in posts. Network-based features are extracted from user interaction network.

Advantages

A novel study that uses community-based features with other feature categories, including metadata, content, and interaction, for detecting automated spammers.

Used Hybrid technique to classify spammers such as random forest, decision tree, and Bayesian network.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

Operational Feasibility

Economic Feasibility

Technical Feasibility

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility

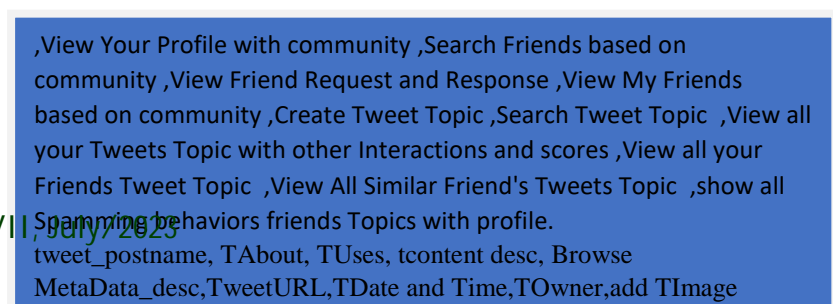
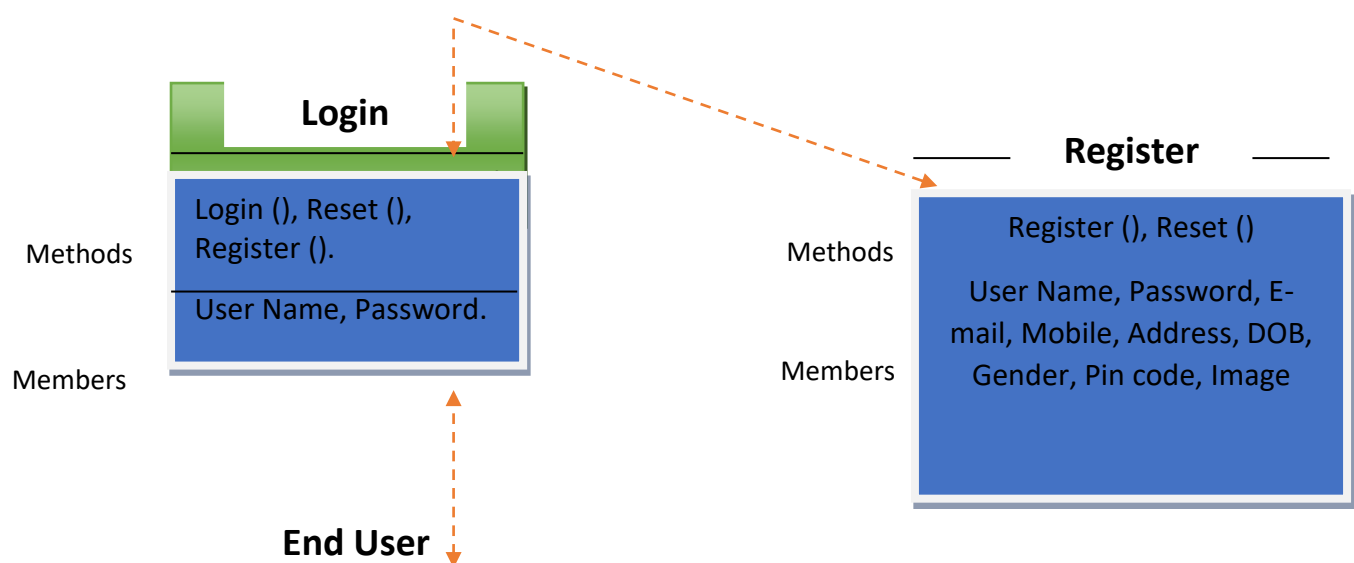
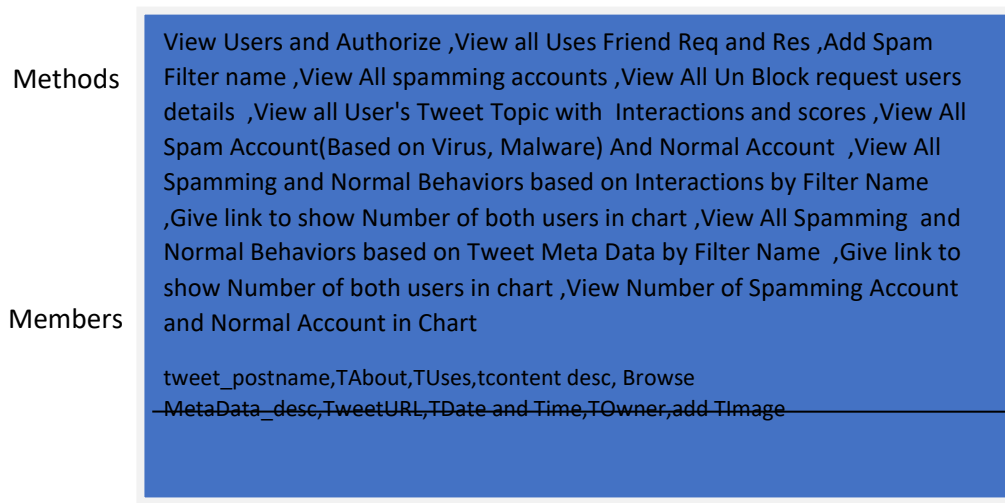
4. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts

Class Diagram :

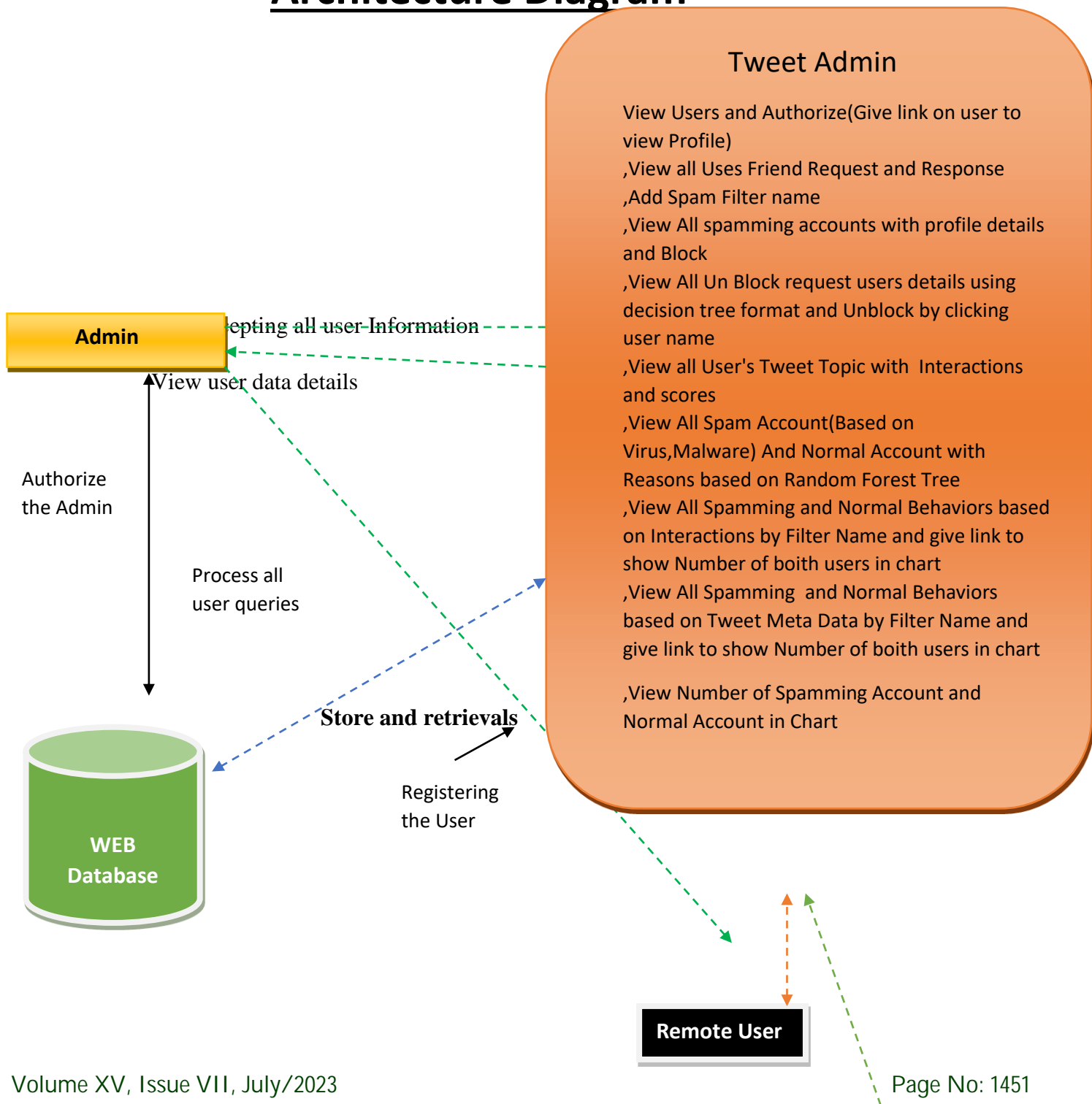
Tweet Admin



Methods

Members

Architecture Diagram



- Register with Community
- ,View Your Profile with community
- ,Search Friends based on community
- ,View Friend Request and Response
- ,View My Friends based on community
- ,Create Tweet Topic
- ,Search Tweet Topic
- ,View all your Tweets Topic with other Interactions and scores
- ,View all your Friends Tweet Topic
- ,View All Similar Friend's Tweets Topic
- ,show all Spamming behaviors friends Topics with profile.

In this paper, we have proposed a novel approach for spammer detection based on *metadata-, content-, and interaction-*

Twitter. Spammers are generally planted in OSNs for varied purposes, but absence of real-life identity hinders them to join the trust network of benign users. Therefore, spammers randomly follow a number of users, but rarely followed back by them, which results in low edge density among their *followers* and *followings*. This type of spammers interaction pattern can be exploited for the development of effective spammers detection systems. Unlike existing approaches of characterizing spammers based on their own profiles, the novelty of the proposed approach lies in the characterization of a spammer based on its neighboring nodes (especially, the followers) and their interaction network. This is mainly due to the fact that users can evade features that are related to their own activities,

but it is difficult to evade those that are based on their followers. On analysis, metadata-based features are found to be least effective as they can be easily evaded by the sophisticated spammers by using random number generator algorithms. On the other hand, both interaction- and community-based features are found to be the most discriminative for spammers detection.

Attaining perfect accuracy in spammers detection is extremely difficult, and accordingly any feature set can never be considered as complete and sound, as spammers keep on changing their operating behavior to evade detection

6. REFERENCES

- [1] M. Tsikerdekis, "Identity deception prevention using common contribution network data," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 188–199, Jan. 2017.
- [2] T. Anwar and M. Abulaish, "Ranking radically influential Web forum users," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 6, pp. 1289–1298, Jun. 2015.
- [3] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "Design and analysis of social botnet," *Comput. Netw.*, vol. 57, no. 2, pp. 556–578, 2013.
- [4] D. Fletcher, "A brief history of spam," *TIME*, Nov. 2, 2009. [Online]. Available: <http://www.time.com/time/business/article/0,8599,1933796,00.html>
- [5] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake OSN accounts by predicting their victims," in *Proc. AISec*, Denver, CO, USA, 2015, pp. 81–89.
- [6] A. A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, and C. Yang, "CATS: Characterizing automation of Twitter spammers," in *Proc. COMSNETS*, Bengaluru, India, Jan. 2013, pp. 1–10.
- [7] K. Lee, J. C. Lee, and S. Webb, "Uncovering social spammers: Socialhoneypots + machine learning," in *Proc. SIGIR*, Geneva, Switzerland, Jul. 2010, pp. 435–442.
- [8] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proc. ACSAC*, Austin, TX, USA, 2010, pp. 1–9.
- [9] H. Yu, M. Kaminsky, P. B. Gibbons, and A. D. Flaxman, "SybilGuard: Defending against sybil attacks via social networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 3, pp. 576–589, Jun. 2008.
- [10] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proc. IMC*, Melbourne, VIC, Australia,



PHISHING WEBSITE DETECTION USING LIGHT GBM AND SVM ALGORITHM

Penumatsa Satya Hanuma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT--Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as using light gbm and svm algorithm.

Keywords: URL, SVM, Light GBM, Cyber security, phishing website.

1. INTRODUCTION

In the once decades, the operation of internet has been increased extensively and makes our live simple, easy and transforms our lives. It plays a major part in areas of communication, education, business conditioning and commerce. A lot of useful data, information and data can be attained from the internet for particular, organizational, profitable and social development. The internet makes it easy to give numerous services through online and enables us to pierce colorful information at any time, from anywhere around the world. Phishing is the act of transferring a indistinguishable dispatch, dispatches or vicious websites to trick the philanthropist / internet druggies into discovering delicate particular information similar as personal identification number (PIN) and word of bank account, credit card information, date of birth or social security figures. Phishing assaults affect hundreds of thousands of internet druggies across the globe.

Individualizes and associations have lost a huge sum of plutocrat and private information through Phishing attacks. Detecting the phishing attack proves to be a challenging task. Tis attack may take a sophisticated form and fool even the savviest users: such as substituting a few characters of the URL with alike unicode characters. By cons, it can come in sloppy forms, as the use of an IP address instead of the domain name. Nonetheless, in the literature, several works tackled the phishing attack detection challenge while using artificial intelligence and data mining techniques [5–9] achieving some satisfying recognition rate peaking at 99.62%. However those systems are not optimal to smartphones and other embed devices because of their complex computing and their high battery usage, since they require as entry complete HTML pages or at least HTML links, tags and webpage JavaScript elements some of those systems uses image processing to achieve the recognition.



Opposite to our recognition system since it is a less greedy in terms of CPU and memory unlike other proposed systems as it needs only six features completely extracted from the URL as input. In this paper, after a summary of this field key researches, we will detail the characteristics of the URL that our system uses to do the recognition. Otherwise we will describe our recognition system, next in the practical part we will test the proposed system while presenting the results obtained. Last but not least we will enumerate the implications and advantages that our system brings as a solution to the phishing attack.

2. OBJECTIVE OF THE PROJECT

Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.

3. RELATED WORK

[1] Andrei Butnaru et al., used a supervised Machine Learning algorithm to block phishing attacks, based on novel mixture

phishing attacks and compare with Google Safe browsers.

[2] Valid Shahrivari et al., proposed a one of the most successful techniques for identifying these malicious works is Machine Learning. It is because of most Phishing attacks have same features which can be noticed by Machine learning techniques. In this many machine learning-based classifiers are used for forecasting the phishing websites. The main advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a forceful tool.

[3] Ammara Zamir et al., proposed a framework for identifying phishing websites using heaping model. Information gain, gain ratio, Relief-F, and recursive feature elimination (RFE) are some of the feature selection algorithms that can be used to analyse Phishing characteristics. The greatest and weakest traits are combined to create two features. Bagging is used in principal component analysis using several Machine learning algorithms, including random forest [RF] and neural network [NN]. Two heaping representations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.

[4] Nguyet Quang Do, Ali Selamat et al., conducted a study on phishing detection and proposed a four different deep learning technique, includes deep neural network (DNN), convolution neural networks (CNN), Long Short-term memory (LSTM),



and gated recurrent unit (GRU). To analyse behaviour of these deep learning architectures, extensive experiments were carried out to examine the impact of parameter tuning on the performance accuracy of the deep learning models. In which each model shows different accuracies from different models.

[5] Ashit Kumar Dutta proposed a URL detection procedure based on Machine Learning methods. An RNN is used for identifying the phishing URL. It is evaluated with 7900 malicious and 5800 genuine sites, respectively. The outcome of this method shows a good concert compare to recent tactics.

4. EXISTING SYSTEM

Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus. Traditionally, phishing attempts were carried out through wide-scale spam campaigns that targeted broad groups of people indiscriminately. The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack. One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate. There are various ML algorithms like SVM, Neural Networks, Random Forest, Decision Tree, XG boost etc. that can be used to classify these URLs. The proposed approach

deals with the Random Forest, Decision Tree classifiers.

5. PROPOSED SYSTEM

Phishing attacks have evolved in terms of sophistication and have increased in sheer number in recent years. This has led to corresponding developments in the methods used to evade the detection of phishing attacks, which pose daunting challenges to the privacy and security of the users of smart systems. This study uses LightGBM and features of the domain name to propose a machine-learning-based method to identify phishing websites and maintain the security of smart systems. Domain name features, often known as symmetry, are the property wherein multiple domain-name-generation algorithms remain constant. The proposed model of detection is first used to extract features of the domain name of the given website, including character-level features and information on the domain name. The features are filtered to improve the model's accuracy and are subsequently used for classification. The results of experimental comparisons showed that the proposed model of detection, which integrates two types of features for training, significantly outperforms the model that uses a single type of feature. The proposed method also has a higher detection accuracy than other methods and is suitable for the real-time detection of many phishing websites.

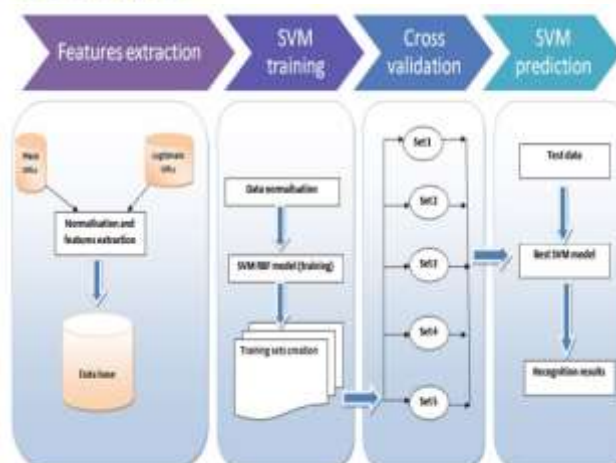


Fig.1. Phishing website process.

6. CONCLUSION

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

Feature Analysis

The features of the domain name used here can be obtained only by using known strings of domain names without obtaining information related to user privacy, such as traffic in the network. Features of the domain name can be divided into two categories according to the acquisition method: features of the characters used in the domain name and features of information on the domain name. The features of information on the domain name can be obtained through the corresponding

website or other query websites to this end, whereas the features of the characters used in the domain name can be obtained through a local feature-extraction algorithm without visiting the website.

7. REFERENCES

- Ms. Sophiya Shikalgar, Mrs. Swati Narwane (2019), Detecting of URL based Phishing Attack using Machine Learning. (vol. 8 Issue 11, November – 2019)
- Rashmi Karnik, Dr. Gayathri M Bhandari, Support Vector Machine Based Malware and Phishing Website Detection.
- Arun Kulkarni, Leonard L. Brown, III², Phishing Websites Detection using Machine Learning (vol. 10, No. 7,2019)
- R. Kiruthiga, D. Akila, Phishing Websites Detection using Machine Learning.
- Ademola Philip Abidoye, Boniface Kabaso, Hybrid Machine Learning: A Tool to detect Phishing Attacks in Communication Networks. (vol. 11 No. 6,2020)
- Andrei Butnaru, Alexios Mylonas and Nikolaos Pitropakis, Article Towards Lightweight URL-Based Phishing Detection.13 June 2021
- Ashit Kumar Dutta (2021), Detecting phishing websites using machine learning technique. Oct 11 2021
- Nguyet Quang Do, Ali Selamat, Ondrej Krejcar, Takeru Yokoi and Hamido Fujita (2021) Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical study.
- Ammara Zamir, Hikmat Ullah Khan and Tassarwar Iqbal, Phishing website detection using diverse machine learning algorithms.



Valid Shahrivari, Mohammad Mahdi Darabi and Mohammad Izadi (2020), Phishing Detection Using Machine Learning Techniques.

A. A. Orunsolu, A. S. Sodiya and A.T. Akinwale (2019), A predictive model for phishing detection.

Wong, R. K. K. (2019). An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning. In Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18-20, 2018, Proceedings (Vol. 11113, p. 199). Springer.

[13] Desai, A., Jatakia, J., Naik, R., & Raul, N. (2017, May). Malicious web content detection using machine learning. In 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1432-1436). IEEE.



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS | ISSN:2320 - 2882

An International Open Access , Peer-reviewed, Refereed Journal

Certificate of Publication

IJCRT | ISSN: 2320-2882 | IJCRT.ORG

The Board of
International Journal of Creative Research Thoughts

Is hereby awarding this certificate to

PICHUKA KEERTHI PRIYA

In recognition of the publication of the paper entitled

EMAIL SPAM DETECTION .

Published In IJCRT(www.ijert.org) & 7.97 Impact Factor by Google Scholar

Volume 8 Issue 8 , Date of Publication: August 2020 2020-08-01

Registration ID :197593




EDITOR IN CHIEF

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS IJCRT
An /international scholar, open Access, Multi-disiplinary , Indexed Journal
Website:www.ijert.org | Email: editor@ijert.org | ESTD:2013

SOCIAL SPAMMER DETECTION VIA CONVEX NONNEGATIVE MATRIX FACTORIZATION

Pokala Alekhya Devi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West
Godavari District, Andhra Pradesh, India, 534202.

V. Sri Valli Devi, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

With the increasing popularity of social network platforms such as Twitter and Sina Weibo, a lot of malicious users, also known as social spammers, disseminate illegal information to normal users. Several approaches are proposed to detect spammers by training a classifier with optimization methods and mainly using content and social following information. Due to the development of spammers' strategies and the courtesy of some legitimate users, social following information becomes vulnerable to fake by spammers. Meanwhile, the possible social activities and behaviors vary significantly among different users, which leads to a large yet sparse feature space to be modeled by existing approaches. To address issues, in this paper, we propose a new approach named CNMFSD for spammer detection in social networks, which exploits both content information and users interaction relationships in an innovative manner. We have empirically validated the proposed method on a real-world Twitter dataset, and experimental results show that the proposed CNMFSD method improves the detection performance significantly compared with baselines.

1. INTRODUCTION

SOCIAL networks, such as Twitter, Face book, and Sina Weibo, are increasingly used to disseminate and share information easily and quickly. However, it is a double-edged sword since the success of social networks also attracts more social spammers [1]. They try to seize our privacy, send us unwanted information, publish malicious content and links [2], and promote commodity information, which thoroughly impacts social stability and organizational management models [3]. According to a study by Nexgate [4], the number of social spammers grows so fast that one in two hundred social messages is spam. Meanwhile, to increase their influence and be undetected, spammers collude with each other to construct the criminal communities [5]. Thus, social spammer detection is a challenging task for researchers.

Successful social spammer detection presents its significance to improve the quality of user experience, and positively impact the overall value of the social systems going forward [6].

In the past decade, researchers have tried different techniques to detect spammers, such as link analysis [7] and content analysis [8], [9]. The methods of content-based detection of spammers mainly focus on analyzing and extracting users' features and then directly applying existing classification approaches such as support vector machines (SVM) to detect spammers [9]–[11]. Recently, more advanced deep learning-based approaches have been proposed to detect social spammers only based on content [12]–[14]. However, with the development of spamming strategies, these methods could not accurately detect spammers with new strategies, only relying on the extracted features. Another category of methods is proposed to detect spammers via social network analysis [15]. These methods assume that spammers cannot establish an arbitrarily large number of social trust relations with legitimate users. The users, who have relatively low social influence or social status in social networks, will be determined as spammers. Unfortunately, only depending on network information, these methods are hard to distinguish between legitimate users and spammers.

Some approaches [16]–[18] have been proposed to detect spammers via both content and network analysis, which identify spammers more accurately than the traditional approaches. The main challenge in detecting social spammers is that the possible social activities and behaviors are more varied and complex, and they constitute a much larger feature space. As a result, spammers are more challenging to detect. Therefore, it is crucial to design more effective methods for extracting users' features. Meanwhile, the reflexive reciprocity [19] indicates that many users simply follow back when they are followed by someone for the sake of courtesy. It is easier for spammers to acquire a large number of follower links in social networks. Thus, with the perceived social influence, they can avoid being detected. However, the interactions between spammers and legitimate users are usually unilateral. In most cases, spammers share a message and then mention (i.e., @) legitimate users. On the contrary, legitimate users constantly interact with legitimate users but have few interactions with spammers. Consequently, it is more reasonable to take the interactions among users into consideration when detecting spammers.

2. EXISTING SYSTEM

Spammers since Heymann et al. [22] firstly surveyed potential solutions and challenges in social spammer detection. Masood et al. [6] elaborated a classification of spammer detection techniques, including fake content, URL-based spam detection, detecting spam in trending topics, and fake user identification. In this paper, we only focus on the binary classification task, i.e., spammer or legitimate user identification.

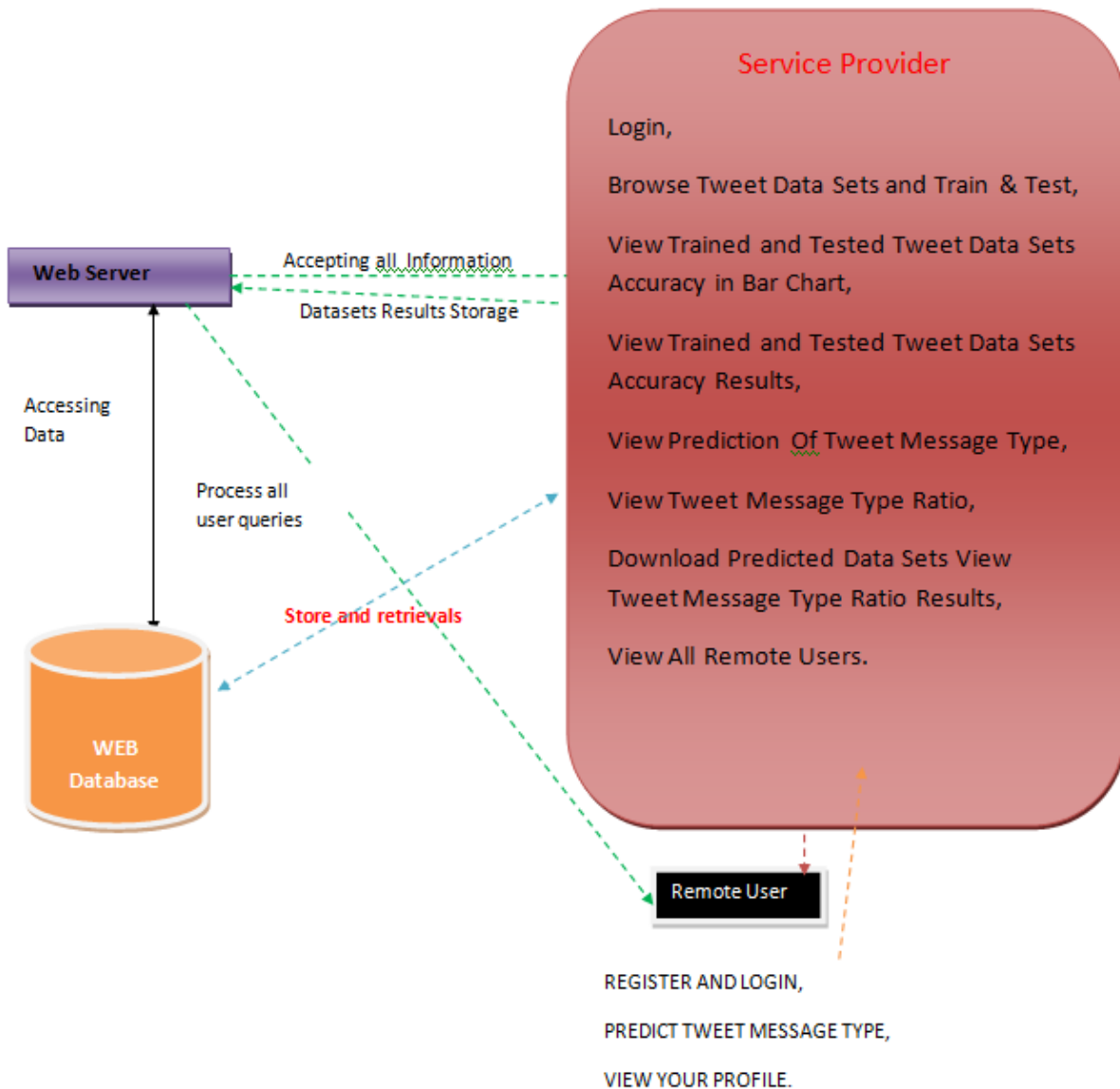
Many approaches employed machine learning methods to train a classifier to detect spammers. SMFSR [16] jointly modeled user activities' information and the social following information to learn a classifier. SSDM [17] incorporated users' text information and social following information into an efficient sparse supervised model for spammer detection. Mateen et al. [23] proposed a hybrid technique that utilizes user-based, content-based, and graph-based characteristics for spammer profiles detection. Gupta et al. [24] presented a policy for the detection of spammers on Twitter and used the popular techniques, i.e., Naive Bayes, clustering, and decision tree.

An important line of research in spam detection relies on analyzing the tweet content, as shown in [25] and [26] where suspicious use of hashtags or URLs is traced. The main objective in [26] is to study the semantics of short texts or messages in contrast with a set of Wikipedia text pages that are modeled and used as an aggregation of entities. The work presented in [25] stresses the need for efficient URL detection schemes utilizing different features such as lexical ones and dynamic behaviors.

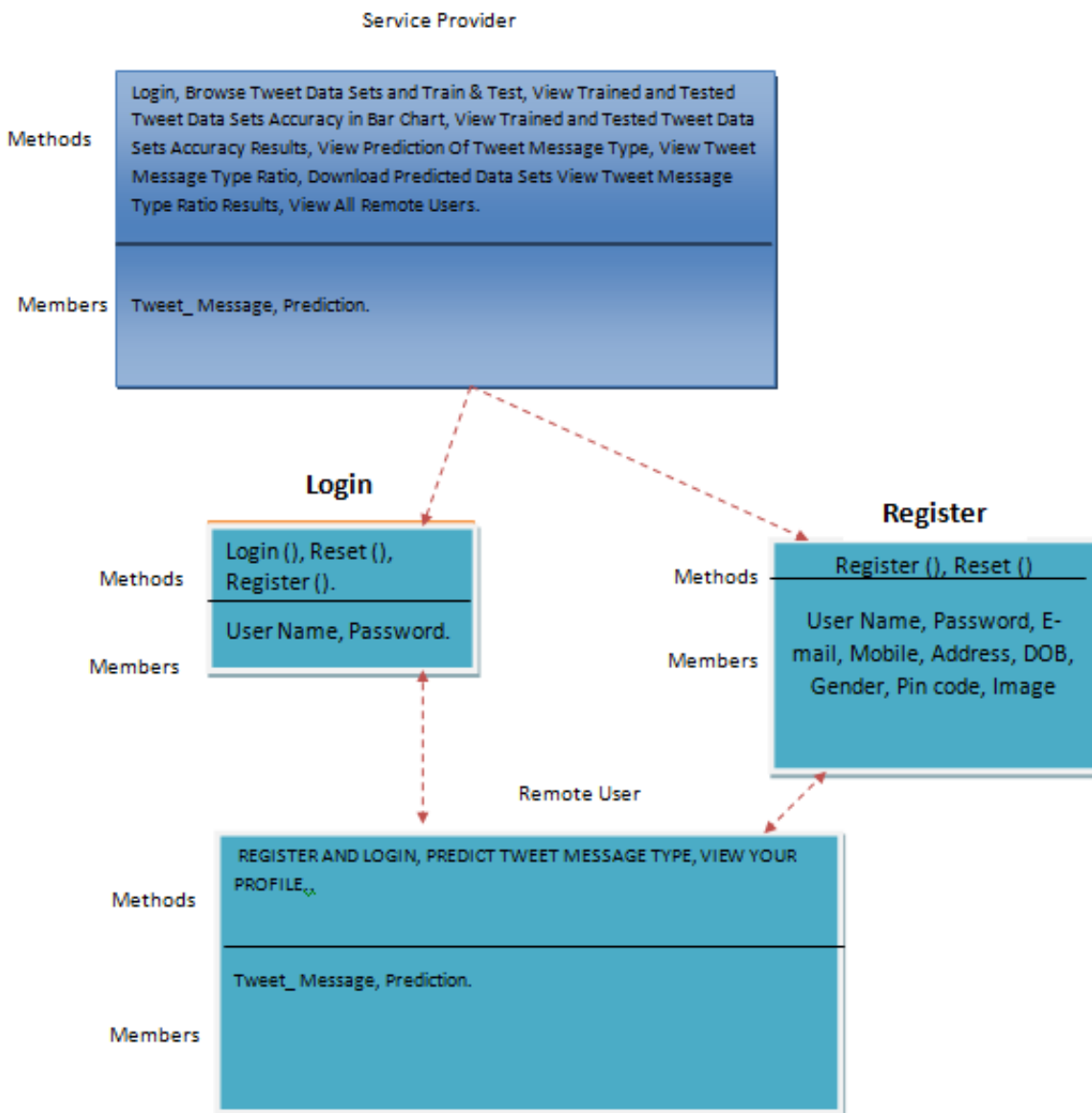
Other directions adopted in detecting Twitter spammers focus on discovering traits or patterns that best describe the spammer's behavioral profile. In such works like [27], the main contribution is to determine deceptive double characters for user profiles, which is done by analyzing nonverbal behavior variables as a function of time, such as follows and retweets. Also, Sumner et al. [28] follow a similar technique. Direct approaches to checking up the user's portfolio include, but are not limited to, the notion of having no profile photo/biography/personal tweets or a suspiciously high/low number of followers/followees. Examples of different profile-based behavior analysis activities are demonstrated in [29] and [30].

Different from discovering traits or patterns, some work considers social network information to identify spammers. Ghosh et al. [31] investigated link farming on Twitter and proposed a ranking scheme to deter spam. Yang et al. [32] proposed a criminal account inference algorithm by exploiting criminal accounts' social relationships. Cao et al. [33] presented the SybilRank algorithm relying on social graph properties to rank users. Cui et al. [34] proposed a Hybrid Factor Non-Negative Matrix Factorization method to incorporate the predictive factors for user-post specific social influence prediction.

Architecture Diagram



➤ **Class Diagram :**



3. SYSTEM STUDY

2.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the

feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system

configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfactory, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

5. CONCLUSION

In this paper, we propose a new framework by taking advantage of content and social interaction information for social spammer detection. Different from existing methods that utilize users' the following information, the proposed method CNMFSD integrates users' interaction information based on the trained classification model. In addition, we introduce a new strategy to induce latent features using CNMF in spammers and legitimate users space for improving the performance of detecting spammers. Experimental results on a real dataset show that CNMFSD obtains better detection performance compared with existing methods. In this work, we employ Convex-NMF to learn latent user features for legitimate users and spammers, respectively. Such a fine-grained learning strategy makes the proposed model obtain accurate latent user representations, which further helps the model to achieve better performance. Besides, introducing social interaction into this task can also improve prediction performance. Although the proposed model outperforms baselines, it also has some disadvantages. First, in the classifier

training stage, we do not consider the social interaction graph, which is trained solely based on the outputs from CNMF. Second, we use tf-idf to extract the user content matrices. However, a spammer always posts some normal tweets to imitate the behavior of legitimate users. Thus, it is essential to distinguish the importance of tweets when we extract the user content matrix. In future work, we will directly use raw tweets as the model input to learn user representations by distinguishing the importance of each tweet via deep learning techniques. After that, we plan to use graph neural networks to model social interactions among users.

REFERENCES

- [1] Aliaksandr Barushka and Petr Hajek. Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks. *Neural Computing and Applications*, 32(9):4239–4257, 2020.
- [2] Qiang Fu, Bo Feng, Dong Guo, and Qiang Li. Combating the evolving spammers in online social networks. *Computers & Security*, 72:60–73, 2018.
- [3] Zhijie Zhang, Rui Hou, and Jin Yang. Detection of social network spam based on improved extreme learning machine. *IEEE Access*, 8:112003– 112014, 2020.
- [4] Nexgate2013. 2013 state of social media spam. <http://nexgate.com/wpcontent/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-\ Research-Report.pdf>.
- [5] Dehai Liu, Benjin Mei, Jinchuan Chen, Zhiwu Lu, and Xiaoyong Du. Community based spammer detection in social networks. In *International Conference on Web-Age Information Management*, pages 554–558. Springer, 2015.
- [6] Faiza Masood, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani, and Mansour Zuair. Spammer detection and fake user identification on social networks. *IEEE Access*, 7:68140– 68152, 2019.
- [7] Sanjeev Rao, Anil Kumar Verma, and Tarunpreet Bhatia. A review on social spam detection: Challenges, open issues, and future directions. *Expert Systems with Applications*, 186:115742, 2021.
- [8] Chao Chen, Jun Zhang, Yi Xie, Yang Xiang, Wanlei Zhou, Mohammad Mehedi Hassan, Abdulhameed AlElaiwi, and Majed Alrubaian. A performance evaluation of machine learning-based streaming spam tweets detection. *IEEE Transactions on Computational social systems*, 2(3):65– 76, 2015.

- [9] Xianghan Zheng, Zhipeng Zeng, Zheyi Chen, Yuanlong Yu, and Chunming Rong. Detecting spammers on social networks. *Neurocomputing*, 159:27–34, 2015.
- [10] Chao Yang, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293, 2013.

APPLICATION OF MACHINE LEARNING IN THE FIELD OF MEDICAL CARE

Polireddy Pushpa Meghana (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

This paper aims to develop a tool for predicting accurate and timely traffic flow Information. Traffic Environment involves everything that can affect the traffic flowing on the road, whether it's traffic signals, accidents, rallies, even repairing of roads that can cause a jam. If we have prior information which is very near approximate about all the above and many more daily life situations which can affect traffic then, a driver or rider can make an informed decision. Also, it helps in the future of autonomous vehicles. In the current decades, traffic data have been generating exponentially, and we have moved towards the big data concepts for transportation. Available prediction methods for traffic flow use some traffic prediction models and are still unsatisfactory to handle real-world applications. This fact inspired us to work on the traffic flow forecast problem build on the traffic data and models. It is cumbersome to forecast the traffic flow accurately because the data available for the transportation system is insanely huge. In this work, we planned to use machine learning, genetic, soft computing, and deep learning algorithms to analyses the big-data for the transportation system with much-reduced complexity. Also, Image Processing algorithms are involved in traffic sign recognition, which eventually helps for the right training of autonomous vehicles.

2. INTRODUCTION

Machine learning (ML) is a science which aims to make machine capable of learning. Machine learning returned to the public's vision after the famous competition between Alpha Go of Google and the Go player Li Sedol, ending with the score 4:1 in 2015. And this event made machine learning more well known among people even among those who were not familiar to computer science and it has caused intense debate in related field. Actually, although machine learning is a

young branch of AI, it is not a new subject. ML is broadly defined as the application of certain computer algorithms to a set of data known to the event outcomes, and the ability to learn to training data and predict new data based on learning outcomes. Its core is induction and summary instead of deductive. Early in the medium of 1950s, Samuel, a computer scientist of United States, designed a chess program that could learn by itself through continuous play. This program shows people the ability of machine at the first time, meanwhile, the unpredictable potential of machine to learn came into people's sight. However, as the research continued, machine learning entered a period of cooling off. Until 1970s, it staged a comeback gradually. And during this period of continuous research and development, until today, machine learning has become an important subject including data mining, pattern recognition, natural language processing and so on. It has also become a core of AI.

In today's society, medical care problems have become a hot topic, and problems such as the unbalance and insufficient allocation of medical resources has become increasingly apparent. In this situation, the application of ML has become the unavoidable trend in the current development of medical care. As early as 1972, the scientists in the University of Leeds in the UK has been trying to use artificial intelligence (ANN) algorithms to judge abdominal pain. Now, more and more researchers are committed to the combination of ML and medical care. The methods of pathological diagnosis of tumours, lung cancer, etc. by ML has gradually entered the field of vision. Some companies, such as Alibaba, Amazon, and Baidu have established their own research team working for it. This introduction of ML in medical care has greatly saved medical resources and provided a new way for citizens to see a doctor and facilitate people's lives. At the same time, the demand of people also provides a new impetus for the research and development of ML, with promoting its continuous improvement.

3. LITERATURE SURVEY

1) Journal of Medical Imaging and Health Informatics ISSN

AUTHORS: Dr. Eddie Yin-Kwee NG , Singapore.

Journal of Medical Imaging and Health Informatics (JMIHI) is a medium to disseminate novel experimental and theoretical research results in the field of biomedicine, biology, clinical,

rehabilitation engineering, medical image processing, bio-computing, D2H2, and other health related areas. As an example, the Distributed Diagnosis and Home Healthcare (D2H2) aims to improve the quality of patient care and patient wellness by transforming the delivery of healthcare from a central, hospital-based system to one that is more distributed and home-based. Different medical imaging modalities used for extraction of information from MRI, CT, ultrasound, X-ray, thermal, molecular and fusion of its techniques is the focus of this journal.

2)Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images.

AUTHORS: Qin Yu, Tao Jiang, Aiyun Zhou, Lili Zhang, Cheng Zhang & Pan Xu

The objective of this study is to evaluate the diagnostic value of combination of artificial neural networks (ANN) and support vector machine (SVM)-based CAD systems in differentiating malignant from benign thyroid nodes with gray-scale ultrasound images. Two morphological and 65 texture features extracted from regions of interest in 610 2D-ultrasound thyroid node images from 543 patients (207 malignant, 403 benign) were used to develop the ANN and SVM models. Tenfold cross validation evaluated their performance; the best models showed accuracy of 99% for ANN and 100% for SVM. From 50 thyroid node ultrasound images from 45 prospectively enrolled patients, the ANN model showed sensitivity, specificity, positive and negative predictive values, Youden index, and accuracy of 88.24, 90.91, 83.33, 93.75, 79.14, and 90.00%, respectively, the SVM model 76.47, 90.91, 81.25, 88.24, 67.38, and 86.00%, respectively, and in combination 100.00, 87.88, 80.95, 100.00, 87.88, and 92.00%, respectively. Both ANN and SVM had high value in classifying thyroid nodes. In combination, the sensitivity increased but specificity decreased. This combination might provide a second opinion for radiologists dealing with difficult to diagnose thyroid node ultrasound images.

3)Liver segmentation from CT images using a sparse priori statistical shape model (SPSSM)

AUTHORS: Xuehu Wang ,Yongchang Zheng ,Lan Gan,Xuan Wang,Xinting Sang,Xiangfeng Kong,Jie Zhao

This study proposes a new liver segmentation method based on a sparse a priori statistical shape model (SP-SSM). First, mark points are selected in the liver a priori model and the original image. Then, the a priori shape and its mark points are used to obtain a dictionary for the liver boundary information. Second, the sparse coefficient is calculated based on the correspondence between mark points in the original image and those in the a priori model, and then the sparse statistical model is established by combining the sparse coefficients and the dictionary. Finally, the intensity energy and boundary energy models are built based on the intensity information and the specific boundary information of the original image. Then, the sparse matching constraint model is established based on the sparse coding theory. These models jointly drive the iterative deformation of the sparse statistical model to approximate and accurately extract the liver boundaries. This method can solve the problems of deformation model initialization and a priori method accuracy using the sparse dictionary. The SP-SSM can achieve a mean overlap error of 4.8% and a mean volume difference of 1.8%, whereas the average symmetric surface distance and the root mean square symmetric surface distance can reach 0.8 mm and 1.4 mm, respectively.

4) Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks

AUTHORS: Qin Yu, Tao Jiang, Aiyun Zhou, Lili Zhang, Cheng Zhang & Pan Xu

Cerebral microbleeds (CMBs) are small haemorrhages nearby blood vessels. They have been recognized as important diagnostic biomarkers for many cerebrovascular diseases and cognitive dysfunctions. In current clinical routine, CMBs are manually labelled by radiologists but this procedure is laborious, time-consuming, and error prone. In this paper, we propose a novel automatic method to detect CMBs from magnetic resonance (MR) images by exploiting the 3D convolutional neural network (CNN). Compared with previous methods that employed either low-level hand-crafted descriptors or 2D CNNs, our method can take full advantage of spatial contextual information in MR volumes to extract more representative high-level features for CMBs, and hence achieve a much better detection accuracy. To further improve the detection performance while reducing the computational cost, we propose a cascaded framework under 3D CNNs for the task of CMB detection. We first exploit a 3D fully convolutional network (FCN) strategy to retrieve the candidates with high probabilities of being CMBs, and then apply a

welltrained 3D CNN discrimination model to distinguish CMBs from hard mimics. Compared with traditional sliding window strategy, the proposed 3D FCN strategy can remove massive redundant computations and dramatically speed up the detection process. We constructed a large dataset with 320 volumetric MR scans and performed extensive experiments to validate the proposed method, which achieved a high sensitivity of 93.16% with an average number of 2.74 false positives per subject, outperforming previous methods using low-level descriptors or 2D CNNs by a significant margin. The proposed method, in principle, can be adapted to other biomarker detection tasks from volumetric medical data

5)Automatic Classification of Specific Melanocytic Lesions Using Artificial Intelligence

AUTHORS:Joanna Jaworek-Korjakowska 1 and Pawel Kleczek 1

Background. Given its propensity to metastasize, and lack of effective therapies for most patients with advanced disease, early detection of melanoma is a clinical imperative. Different computeraided diagnosis (CAD) systems have been proposed to increase the specificity and sensitivity of melanoma detection. Although such computer programs are developed for different diagnostic algorithms, to the best of our knowledge, a system to classify different melanocytic lesions has not been proposed yet. *Method.* In this research we present a new approach to the classification of melanocytic lesions. This work is focused not only on categorization of skin lesions as benign or malignant but also on specifying the exact type of a skin lesion including melanoma, Clark nevus, Spitz/Reed nevus, and blue nevus. The proposed automatic algorithm contains the following steps: image enhancement, lesion segmentation, feature extraction, and selection as well as classification. *Results.* The algorithm has been tested on 300 dermoscopic images and achieved accuracy of 92% indicating that the proposed approach classified most of the melanocytic lesions correctly. *Conclusions.* A proposed system can not only help to precisely diagnose the type of the skin mole but also decrease the amount of biopsies and reduce the morbidity related to skin lesion excision.

3. SYSTEM TEST

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel

threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

4. CONCLUSION

This article reviews the main methods of machine learning, and summarizes several representative applications after understanding the history of machine learning in the medical field and its current application. The typical ideas and algorithms are summarized. At the same time, the improvement method based on machine learning in the process of visiting is proposed. However, this does not mean that ML is perfect. Whether in terms of technology, ethic or law, it has certain problems. The solution of these problems requires technicians and legal personnel. Working together, and how to strike a balance between manpower and machine is also a problem that everyone of us must face.

5. REFERENCES

- [1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (references)
- [2] Jiang M, Zhang S, Huang J, et al. Scalable histopathological image analysis via supervised hashing with multiple features[J]. Medical Image Analysis, 2016, 34:3-12.
- [3] Joanna J K, Pawel K . Automatic Classification of Specific Melanocytic Lesions Using Artificial Intelligence[J]. BioMed Research International, 2016, 2016:1-17.

- [4] Lu-Cheng, Zhu, Yun-Liang, Ye, Wen-Hua, Luo, Meng, Su, Hang-Ping, Wei, Xue-Bang, Zhang, Juan, Wei, Chang-Lin, Zou. A model to discriminate malignant from benign thyroid nodules using artificial neural network. [J]. PloS one, 2013, 8(12): e82211.
- [5] Huang W C , Chang C P . Automatic Nasal Tumor Detection by grey prediction and Fuzzy C-Means clustering [C]// IEEE International Conference on Systems. IEEE, 2006. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.
- [6] Sarraf S , Tofighi G . Classification of Alzheimer's Disease using fMRI Data and Deep Learning Convolutional Neural Networks [J]. 2016.
- [7] Dou Q , Chen H , Yu L , et al. Automatic Detection of Cerebral Microbleeds from MR Images via 3D Convolutional Neural Networks [J]. IEEE Transactions on Medical Imaging, 2016: 1-1.
- [8] Pang-ning Tan, Michael Steinbach, Vipin Kumar, Introduction to data mining, Beijing: Posts & Telecom Press, 2011.
- [9] Xue-Hu WANG, Study Liver Segmentation Method from CT Images based on Deformation Optimization and Sparse Statistics [D]. Beijing Institute of Technology, 2015.
- [10] Yu Q , Jiang T , Zhou A , et al. Computer-aided diagnosis of malignant or benign thyroid nodes based on ultrasound images [J]. European Archives of Oto-Rhino-Laryngology, 2017, 274(7): 2891- 2897.
- [11] Fei Liu, Jun-Ran Zhang, Hao Yang. Advances in medical images recognition based on deep learning [J]. Chinese Journal of Biomedical Engineering, 2018.
- [12] Ke-Yang Zhao, Mu-Yue Yang, Jing-Yu Zhu, Ze-Qi Wang, Wei-Wei Shen. Machine learning AIDS in tumor diagnosis [J]. Tumor, 2018, 38(10): 987-991.
- [13] Bin Huang, Feng Liao, Yu-Feng Ye. Advances in machine learning in image analysis of nasopharyngeal carcinoma [J]. International Journal of Medical Radiology, 2019(1).
- [14] Li F , Tran L , Thung K H , et al. A Robust Deep Model for Improved Classification of AD/MCI Patients [J]. IEEE Journal of Biomedical and Health Informatics, 2015, 19(5): 1-1.



BULDING SEARCH ENGINE USING MACHINE LEARNING

Pragada Durga Sai Prasad (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

The web is the huge and most extravagant wellspring of data. To recover the information from the World Wide Web, Search Engines are commonly utilized. Search engines provide a simple interface for searching for user query and displaying results in the form of the web address of the relevant web page, but using traditional search engines has become very challenging to obtain suitable information. This paper proposed a search engine using Machine Learning technique that will give more relevant web pages at top for user queries.

1. INTRODUCTION

World Wide Web is actually a web of individual systems and servers which are connected with different technology and methods. Every site comprises the heaps of site pages that are being made and sent on the server. So if a user needs something, then he or she needs to type a keyword. Keyword is a set of words extracted from user search input. Search input given by a user may be syntactically incorrect. Here comes the actual need for search engines. Search engines provide you a simple interface to search user queries and display the results.

1) Web crawler Web crawlers help in collecting data about a website and the links

related to them. We are only using web crawlers for collecting data and information from WWW and storing it in our database.

2) Indexer Indexer which arranges each term on each web page and stores the subsequent list of terms in a tremendous repository.

3) Query Engine It is mainly used to reply to the user's keyword and show the effective outcome for their keyword. In the query engine, the Page ranking algorithm ranks the URL by using different algorithms in the query engine.

4) This paper utilizes Machine Learning Techniques to discover the utmost suitable web address for the given keyword. The output of the PageRank algorithm is given as input to the machine learning algorithm.



5)The section II discusses the related work in search engine and PageRank algorithm. In section III Objective is explained. Section IV deals with a proposed system which is based on machine learning technique and section V contains the conclusion.

2. LITERATURE SURVEY

1) Weighted page rank algorithm based on in-out weight of webpages

AUTHORS: Kalyani Desikan, B. Jaganathan.

In its classical formulation, the well known page rank algorithm ranks web pages only based on in-links between web pages. We propose a new in-out weight based page rank algorithm. In this paper, we have introduced a new weight matrix based on both the in-links and out-links between web pages to compute the page ranks. We have illustrated the working of our algorithm using a web graph. We notice that the page rank values of the web pages computed using the original page rank algorithm and our proposed algorithm are comparable. Moreover, our algorithm is found to be efficient with respect to the time taken to compute the page rank values.

2)Web Page Ranking Using Machine Learning Approach

AUTHORS:Junaid Khan, Arunima Jaiswal. One of the key components which ensures the acceptance of web search service is the web page ranker - a component which is said to have been the main contributing factor to the early successes of Google. It is

well established that a machine learning method such as the Graph Neural Network (GNN) is able to learn and estimate Google's page ranking algorithm. This paper shows that the GNN can successfully learn many other web page ranking methods e.g. TrustRank, HITS and OPIC. Experimental results show that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this observation lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

3)Review of features and machine learning techniques for web searching.

AUTHORS:[Neha Sharm](#) ,Narendra Kohli

As the amount of information is growing rapidly on world wide web, it has become very difficult to get relevant information using traditional search engines within a stipulated time. The main reasons for irrelevant search results are the lack of understanding of user's search intention or user's preferences, keyword based searching, short queries. In this paper, we will study different features that are used in information retrieval. We will also discuss various machine learning techniques that are helpful in deciding the relevance of web page to user. We have done classification on the basis of features. In the end we will compare different techniques and their pros and cons are also discussed.

3. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal



is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application. It is done after the completion of an individual unit before integration. This is a



structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements



document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or –

one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

5. CONCLUSION

Search engines are very useful for finding out more relevant URLs for given keywords. Due to this, user time is reduced for searching the relevant web page. For this, Accuracy is a very important factor. From the above observation, it can be concluded that XGBoost is better in terms of accuracy than SVM and ANN. Thus, Search engines built using XGBoost and PageRank algorithms will give better accuracy.

6. REFERENCES

- [1] Manika Dutta, K. L. Bansal, “A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)”, International Journal on Recent and Innovation Trends in Computing and Communication, 2016.
- [2] Gunjan H. Agre, Nikita V.Mahajan, “Keyword Focused Web Crawler”, International Conference on Electronic and Communication Systems, IEEE, 2015.
- [3] Tuhena Sen, Dev Kumar Chaudhary, “Contrastive Study of Simple PageRank, HITS and Weighted PageRank Algorithms:



Review”, International Conference on Cloud Computing, Data Science & Engineering, IEEE, 2017.

[4] Michael Chau, Hsinchun Chen, “A machine learning approach to web page filtering using content and structure analysis”, Decision Support Systems 44 (2008) 482–494, scienceDirect, 2008.

[5] Taruna Kumari, Ashlesha Gupta, Ashutosh Dixit, “Comparative Study of Page Rank and Weighted Page Rank Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering, February 2014.

[6] K. R. Srinath, “Page Ranking Algorithms – A Comparison”, International Research Journal of Engineering and Technology (IRJET), Dec 2017.

[7] S. Prabha, K. Duraiswamy, J. Indhumathi, “Comparative Analysis of Different Page Ranking Algorithms”, International Journal of Computer and Information Engineering, 2014.

[8] Dilip Kumar Sharma, A. K. Sharma, “A Comparative Analysis of Web Page Ranking Algorithms”, International Journal on Computer Science and Engineering, 2010.

[9] Vijay Chauhan, Arunima Jaiswal, Junaid Khalid Khan, “Web Page Ranking Using Machine Learning Approach”, International Conference on Advanced Computing Communication Technologies, 2015.

[10] Amanjot Kaur Sandhu, Tiewei s. Liu., “Wikipedia Search Engine: Interactive Information Retrieval Interface Design”, International Conference on Industrial and Information Systems, 2014.

[11] Neha Sharma, Rashi Agarwal, Narendra Kohli, “Review of features and machine learning techniques for web searching”, International Conference on Advanced Computing Communication Technologies, 2016.

[12] Sweah Liang Yong, Markus Hagenbuchner, Ah Chung Tsoi, “Ranking Web Pages using Machine Learning Approaches”, International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

MOVIE RECOMMENDATION SYSTEM USING SENTIMENT ANALYSIS FROM MICROBLOGGING DATA

Pydikondala V V S Rama Krishna (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract

Recommendation systems are important intelligent systems that play a vital role in providing selective information to users. Traditional approaches in recommendation systems include collaborative filtering and content-based filtering. However, these approaches have certain limitations like the necessity of prior user history and habits for performing the task of recommendation. In order to reduce the effect of such dependencies, this paper proposes a hybrid recommendation system which combines the collaborative filtering, content-based filtering with sentiment analysis of movie tweets. The movie tweets have been collected from microblogging websites to understand the current trends and user response of the movie. Experiments conducted on public database produce promising results.

Keywords Recommendation System _ Twitter _ Collaborative filtering _ Contentbased filtering _ Sentiment Analysis.

1. INTRODUCTION

A recommendation system, sometimes known as a recommendation engine, is a paradigm for information filtering that aims to anticipate user preferences and provide suggestions in accordance with these preferences. These technologies are now extensively employed in a variety of industries, including those that deal with utilities, books, music, movies, television, apparel, and restaurants. These systems gather data on a user's preferences and behaviour, which they then use to enhance their future recommendations. Movies are a fundamental aspect of life. There are many various kinds of movies, such as those meant for amusement, those meant for teaching, children's animation movies, horror movies, and action movies. Movies' genres, such as comedy, thriller, animation, action, etc., make it simple to distinguish between them. Another

approach to differentiate between movies is to look at their release year, language, director, etc. When watching movies online, there are plenty to choose from in our list of top picks. We may find our favourite movies among all of these various kinds of movies with the aid of movie recommendation systems, which saves us the stress of having to spend a lot of time looking for our preferred movies. As a result, it is essential that the system for suggesting movies to us is highly trustworthy and gives us recommendations for the films that are either most similar to or identical to our tastes. Recommendation systems are being used by a lot of businesses to improve customer engagement and the purchasing experience. The most significant advantages of recommendation systems are client happiness and income. A highly effective and crucial mechanism is the movie recommendation system. However, because of the limitations with a pure collaborative method, scalability concerns and poor suggestion quality also affect movie recommendation systems.

Objective: The goal of this project is to provide people reliable movie suggestions. The project's objective is to make movie recommendation systems better than pure techniques in terms of accuracy, quality, and scalability. By combining content-based filtering with collaborative filtering, a hybrid strategy is used to accomplish this. In social networking sites, recommendation systems are employed as information filtering tools to reduce data overload. Therefore, there is a lot of room for research in this area to enhance the quality, accuracy, and scalability of movie recommendation systems. A highly effective and crucial mechanism is the movie recommendation system. However, because of the limitations with a pure collaborative method, scalability concerns and poor suggestion quality also affect movie recommendation systems.

A Description Of The Project: In order to create a weighted similarity measure based on evolutionary algorithms and fuzzy k means clustering technique, the hybrid approach offered an integrative strategy. In comparison to the current system, the suggested movie recommendation system provides finer similarity metrics and quality. By using the clustered data points as an input dataset, this issue may be resolved. The suggested method aims to enhance the quality and scalability of the movie recommendation system. By combining Content-Based Filtering with Collaborative Filtering, we build a hybrid strategy that allows both methods to benefit from one another. We utilised the cosine similarity measure to quickly and accurately determine how

similar the various movies in the supplied dataset are to one another as well as to shorten the calculation time for the movie recommender engine.

2. LITERATURE SURVEY

2.1 Existing System

To suggest movies to viewers that they would find interesting, the current system incorporates updated k-means and k-nearest neighbour algorithms. The k-nearest neighbour algorithm's central principle is to measure the distance between each sample in the dataset and the unlabeled sample in the test data and training data spaces, and then make a decision based on the k-nearest neighbor's vote. However, since the k-nearest neighbour classification computation is so large, think about pre-processing the data using the k-means clustering technique. • The clustering algorithm K-means Data classification into several groups is a technique known as clustering, which has emerged as a key tool in data mining. The choice of the first clustering centre will have a big impact on the clustering outcomes. The first k clusters are set as null values, and the initial k data points are chosen at random from the dataset to serve as the initial cluster centre vector. Determine the separation between every data point and every cluster centre vector that was chosen at random. The goal function value is reduced and the allocated cluster centre is designated as being a member of the cluster according to the minimal distance concept. Calculate the cluster average as the new cluster centre by averaging all the data points in each cluster. Get K clusters and centres by repeating the method until the cluster centres do not change. The choice of the starting clustering centre and the k value have a significant impact on the performance of the k-means clustering method. Both the running duration and the quantity of runs may be decreased by choosing the k value using the elbow approach. The first clustering centres have an impact on the fundamental k-means method. • Determination of K value: To solve the issue of manually entering the k value in the conventional k-means clustering algorithm, a technique to automatically retrieve the k value is suggested. The number of clusters in the data are estimated using the elbow approach, and each category's initial cluster centre and k value are determined prior to clustering. When there is no specified k value in the k-means clustering algorithm, the optimal solution of k-means parameters is to minimize the cost function is the sum of the degree of distortion of each class, the degree of distortion of each class is equal

to the sum of the squared distances from each variable point to the center of class and the compactness of the members in the class is proportional to the degree of distortion of the class.

2.2 Proposed System

A kind of artificial neural network that takes inspiration from biological processes is called a convolutional neural network. Neural networks have the capacity to learn on their own and generate output that is independent of the input that is given to them. Comparatively speaking, CNNs employ very minimal pre-processing; as a result, the network automatically learns how to improve its filters (or kernels). There are input and output layers, as well as a number of hidden layers, in a typical CNN network. Convolutional layers often make up the hidden layers of a CNN. ReLU is a standard activation function, and further operations like pooling layers, fully linked layers, and normalising layers are often performed after it. Building components like convolutional layers, pooling layers, and fully linked layers are all part of the CNN architecture. One or more fully connected layers are followed by one or more repeats of a stack of multiple convolutional layers, a pooling layer, and so on.

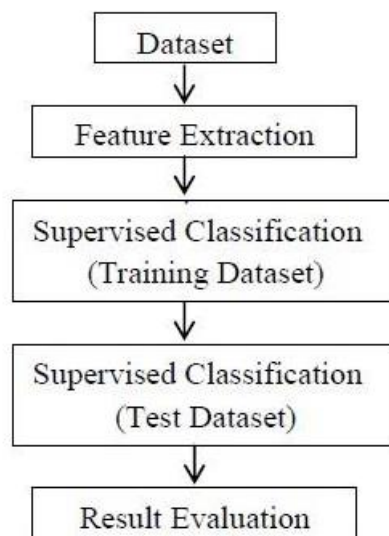


Fig 1: System Flow

3. SYSTEM ANALYSIS & DESIGN

3.1 Functional Requirements: Functional requirements are those that pertain to the technical functioning of the programme. It improves and describes the flow of components and their structural flow. Function statements classify and analyse large raw data sets while also learning from them. After that, the data sets are grouped into clusters and their spuriousness is examined. After the dataset has been cleaned, the machine learns the specified pattern for the same sample, performs several iterations, and then generates an output.

3.2 Performance Requirements: An application's output is used to evaluate its performance. An essential step in the study of a system is the definition of the requirements. It is only feasible to develop a system that will fit into the appropriate environment when the requisite specifications are adequately provided. Because they are the ones who will really utilise the system, the users of the current system should provide the required requirements. This is because it's necessary to be aware of the needs at the first

3.3 System Architecture: Movie Lens provided the dataset for the project we are proposing. But the format of this dataset is raw. The dataset is a compilation of stock market value data for various firms. Making processed data out of raw data is the first stage. Which is accomplished by feature extraction, since just a few of the many qualities included in the raw data gathered are required for the prediction. A reducing procedure is called feature extraction. A structural model provides information on a system's structure, behaviour, and viewpoints.

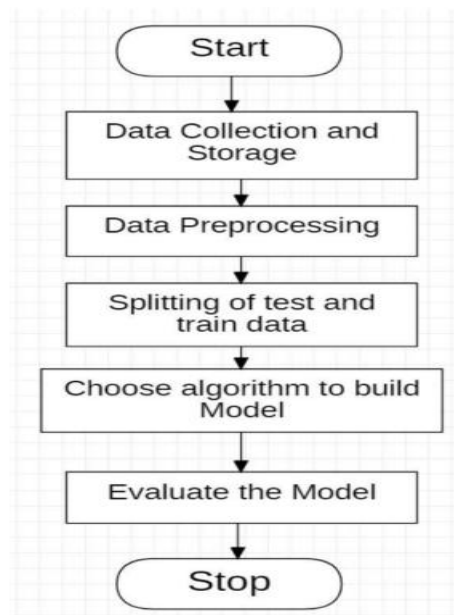


Fig 2 : System Architecture

3.4 Data Flow Diagram : Flow chart of the process consists of dataset split into training data and testing data. Training data gets feed into machine learning algorithms. A model is evaluated using algorithms and Testing data. The model is used to predict crop to be yield based on given parameters.

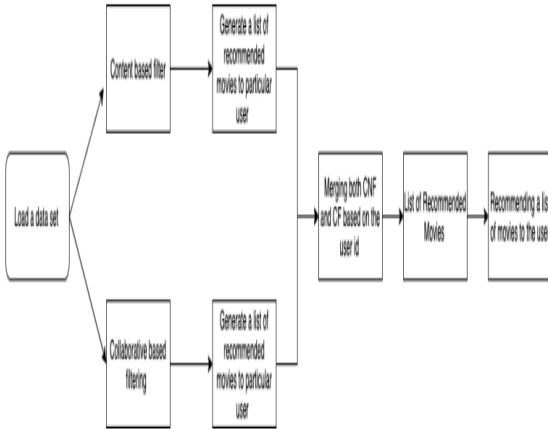


Fig 3: Data Flow Diagram

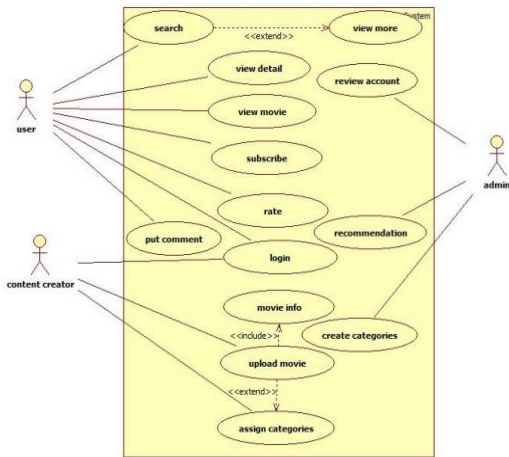


Fig 4 Use Case UML Diagrams

4. IMPLIMENTATION AND RESUTS

Methods / Algorithms Used These are the Machine Learning Algorithms implemented during the building of the project.

- RANDOM FOREST
- LOGISTIC REGRESSION
- ANN
- CNN

```
ratings.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105339 entries, 0 to 105338
Data columns (total 4 columns):
userId      105339 non-null int64
movieId     105339 non-null int64
rating      105339 non-null float64
timestamp   105339 non-null int64
dtypes: float64(1), int64(3)
memory usage: 3.2 MB

movies.shape

(10329, 3)

movies.describe()

   movieId
count  10329.000000
mean   31924.282893
std    37734.741149
min     1.000000
25%    3240.000000
50%    7088.000000
75%    59900.000000
max    149532.000000
```

Fig 5 : Loading data files

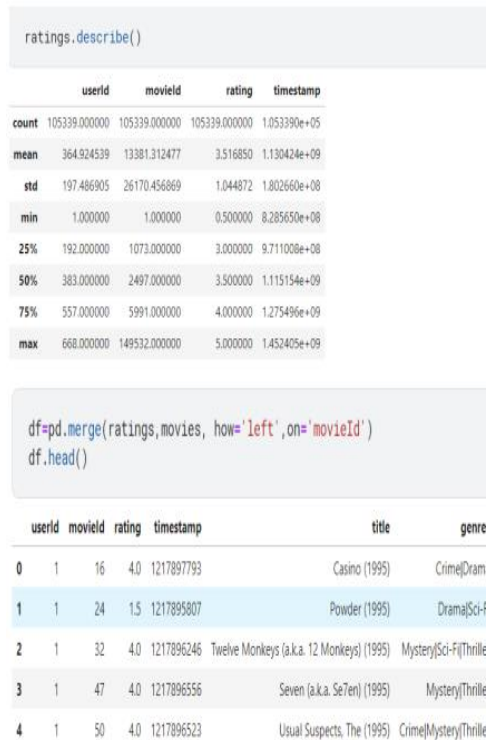


Fig 6: Combining data sets



Fig 7 : Ratings for movies

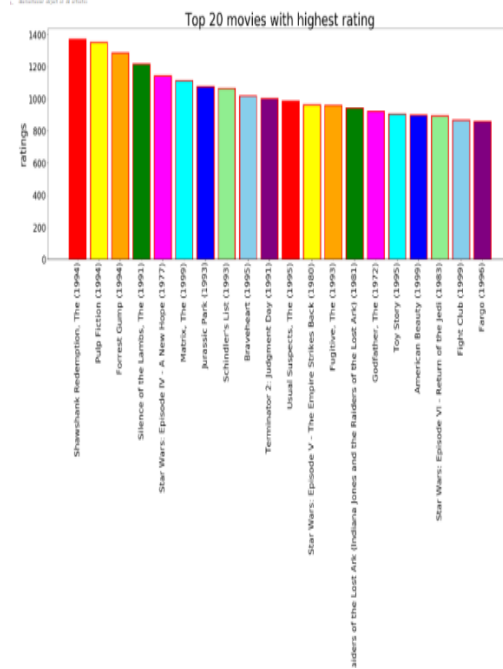


Fig 8: Top 20 movies based on ratings



Fig 9: Recommendations after giving input

5. CONCLUSION

In this project, a hybrid approach is presented by combining content-based filtering and collaborative filtering; using Singular Value Decomposition (SVD) as a classifier and Cosine Similarity as the proposed methodology. This approach aims to improve the accuracy, quality,

and scalability of movie recommendation systems. On three separate movie datasets, existing pure methodologies and the hybrid approach are put into practise, and the outcomes are compared. Comparative findings reveal that the suggested technique is superior than pure approaches in terms of accuracy, quality, and scalability of the movie recommendation system. The suggested solution takes less time to compute than the other two pure approaches.

6. FUTURE SCOPE

In the suggested technique, movie genres have been taken into account; however, in the future, we should also take user age into account since movie tastes vary with age. For instance, when we are young, we tend to favour animated films over other types of films. The memory needs of the suggested solution need to be improved in the future. Here, the suggested methodology has only been applied to several movie datasets. The performance may be calculated in the future, and it can also be used to the Film Affinity and Netflix databases.

7. REFERENCES

- [1] Hirdesh Shivhare, Anshul Gupta and Shalki Sharma (2015), “Recommender system using fuzzy c-means clustering and genetic algorithm based weighted similarity measure”, IEEE International Conference on Computer, Communication and Control
- [2] Manoj Kumar, D.K. Yadav, Ankur Singh and Vijay Kr. Gupta (2015), “A Movie Recommender System: MOVREC”, International Journal of Computer Applications (0975 – 8887) Volume 124 – No.3.
- [3] ryuri Kim, Ye Jeong Kwak, hyeonjeong Mo, Mucheol Kim, Seungmin Rho, Ka Lok Man, Woon Kian Chong (2015), “Trustworthy Movie Recommender System with Correct Assessment and Emotion Evaluation”, Proceedings of the International multiconference of Engineers and Computer Scientists Vol II.
- [4] Zan Wang, Xue Yu*, Nan Feng, Zhenhua Wang (2014), “An Improved Collaborative Movie Recommendation System using Computational Intelligence”, Journal of Visual Languages & Computing, Volume 25, Issue 6.

[5] Debadrita Roy, Arnab Kundu, (2013), “Design of Movie Recommendation System by Means of Collaborative Filtering”, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 4.

CROP RECOMMENDER SYSTEM USING MACHINE LEARNING APPROACH

Ramayanapu Venkata Someswara Rao (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

ABSTRACT

Agriculture and its allied sectors are undoubtedly the largest providers of livelihoods in rural India. The agriculture sector is also a significant contributor factor to the country's Gross Domestic Product (GDP). Blessing to the country is the overwhelming size of the agricultural sector. However, regrettable is the yield per hectare of crops in comparison to international standards. This is one of the possible causes for a higher suicide rate among marginal farmers in India. This paper proposes a viable and user-friendly yield prediction system for the farmers. The proposed system provides connectivity to farmers via a mobile application. GPS helps to identify the user location. The user provides the area & soil type as input.

Machine learning algorithms allow choosing the most profitable crop list or predicting the crop yield for a user-selected crop. To predict the crop yield, selected Machine Learning algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), Multivariate Linear Regression (MLR), and K-Nearest Neighbour (KNN) are used. Among them, the Random Forest showed the best results with 95% accuracy. Additionally, the system also suggests the best time to use the fertilizers to boost up the yield.

1. INTRODUCTION

Agriculture has an extensive history in India. Recently, India is ranked second in the farm output worldwide [15]. Agriculture-related industries such as forestry and fisheries contributed for 16.6% of 2009 GDP and around 50% of the total workforce. Agriculture's monetary contribution to India's GDP is decreasing [1]. The crop yield is the significant factor contributing in agricultural monetary. The crop yield depends on multiple factors such as climatic, geographic, organic, and financial elements [6]. It is difficult for farmers to decide

when and which crops to plant because of fluctuating market prices [7]. Citing to Wikipedia figures India's suicide rate ranges from 1.4-1.8% per 100,000 populations, over the last 10 years [15]. Farmers are unaware of which crop to grow, and what is the right time and place to start due to uncertainty in climatic conditions. The usage of various fertilizers is also uncertain due to changes in seasonal climatic conditions and basic assets such as soil, water, and air. In this scenario, the crop yield rate is steadily declining [2]. The solution to the problem is to provide a smart user-friendly recommender system to the farmers.

The crop yield prediction is a significant problem in the agriculture sector [3]. Every farmer tries to know crop yield and whether it meets their expectations [4], thereby evaluating the previous experience of the farmer on the specific crop predict the yield [3]. Agriculture yields rely primarily on weather conditions, pests, and preparation of harvesting operations. Accurate information on crop history is critical for making decisions on agriculture risk management [5].

In this paper, we have proposed a model that addresses these issues. The novelty of the proposed system is to guide the farmers to maximize the crop yield as well as suggest the most profitable crop for the specific region. The proposed model provides crop selection based on economic and environmental conditions, and benefit to maximize the crop yield that will subsequently help to meet the increasing demand for the country's food supplies [8]. The proposed model predicts the crop yield by studying factors such as rainfall, temperature, area, season, soil type etc. The system also helps to determine the best time to use fertilizers. The existing system which recommends crop yield is either hardware based being costly to maintain, or not easily accessible. The proposed system suggests a mobile-based application that precisely predicts the most profitable crop by predicting the crop yield. The use of GPS helps to identify the user location. The user provides an area under cultivation and soil type as inputs. According to the requirement, the model predicts the crop yield for a specific crop. The model also recommends the most profitable crop and suggests the right time to use the fertilizers

The major contributions of the paper are enlisted below

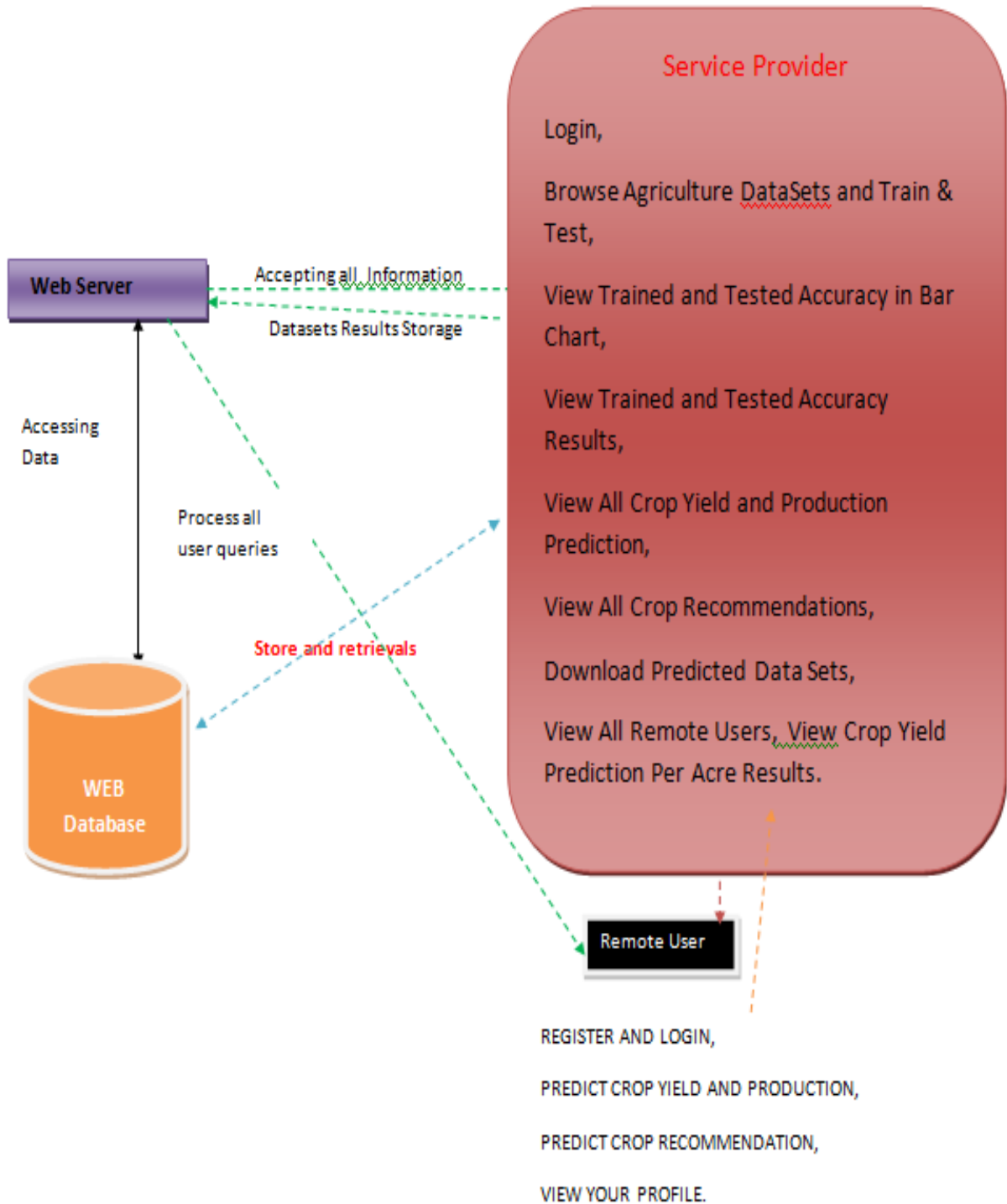
1. Prediction of the crop yield for specific regions by executing various Machine Learning algorithms, with a comparison of error rate and accuracy.
2. A user-friendly mobile application to recommend the most profitable crop.
3. A GPS based location identifier to retrieve the rainfall estimation at the given area.

4. A recommender system to suggest the right time for using fertilizers. The organization of the rest of the paper is as follows. Section II discusses the background work of researchers in the field of agriculture and yield prediction. Section III presents the proposed model for yield prediction and recommends which crop for cultivation. The model also suggests the best suitable time for the use of fertilizers. Section IV discusses the results and Section V concludes the paper.

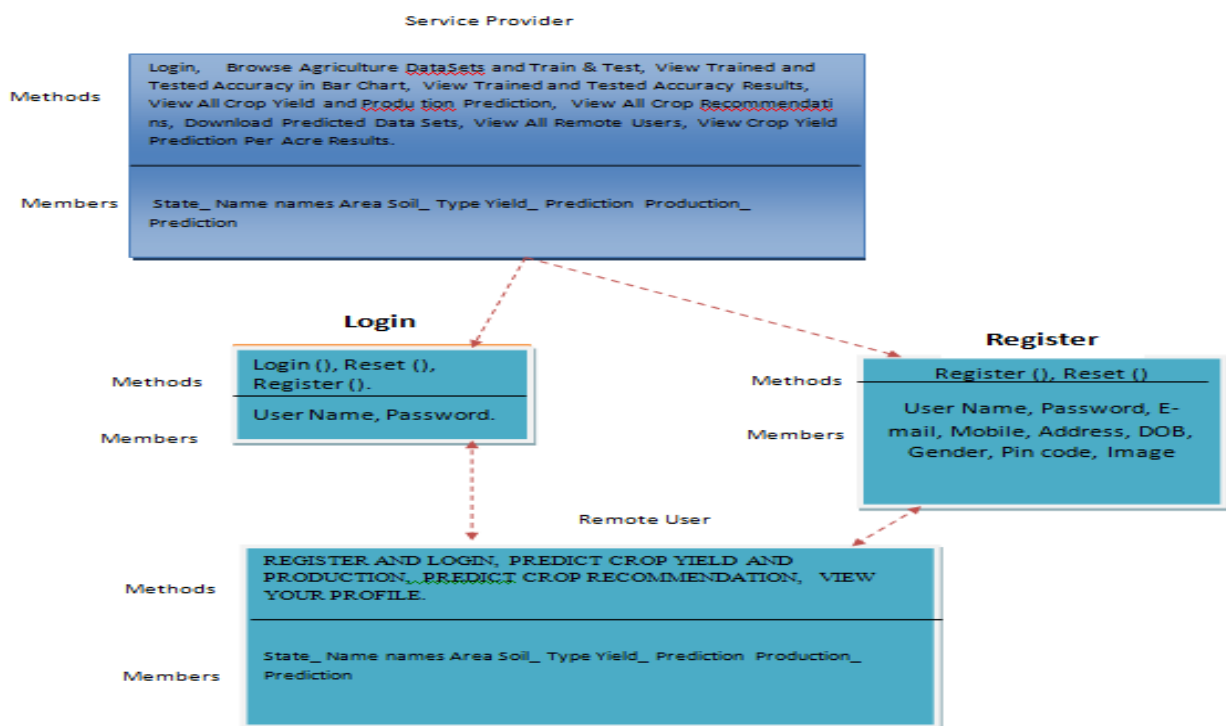
2.EXISTING SYSTEM

- ❖ One of the early works developed a dedicated website to assess the impact of weather parameters on crop production in the identified districts of Madhya Pradesh [10]. The districts were selected on the basis of the region covered by the crop. Based on these criteria, the first five top districts with a maximum crop area were chosen. The basis of the crops selected for the study was on prevailing crops in the selected districts. The crops picked included maize, soybean, wheat and paddy, for which the yield for a continuous period of 20 years of knowledge, were tabulated. The accuracy of the established model ranged from 76% to 90% for the chosen crops with an average accuracy of 82%.
- ❖ Another important work checks the soil quality and predicts the crop yield along with a suitable recommendation of fertilizers [11]. The Ph value and the location from the user were inputs used in this model. AnAPI was used to predict the weather, temperature for the current place. The system used both supervised as well as unsupervised ML algorithms and compares the results of the two.
- ❖ A classifier that uses a greedy strategy to predict the crop yield was proposed in [12]. A decision tree classifier that uses an attribute has been shown to yield better results. An ensemble model proposed suggests integrating the effects of different models, which has been shown to be typically better than the individual models. Random forests ensemble classification uses multiple decision tree models to predict the crop yield. The data are split up into two sets, such as training data and test data, with a ratio of 67% and 33%, with which the mean and standard deviation are calculated. This work also incorporates the clustering of similar crops to get the most accurate results.

Architecture Diagram



➤ **Class Diagram :**



3. CONCLUSION

This paper highlighted the limitations of current systems and their practical usage on yield prediction. Then walks through a viable yield prediction system to the farmers, a proposed system provides connectivity to farmers via a mobile application. The mobile application includes multiple features that users can leverage for the selection of a crop. The inbuilt predictor system helps the farmers to predict the yield of a given crop. The inbuilt recommender system allows a user exploration of the possible crops and their yield to take more educated decisions. For yield to accuracy, various machine learning algorithms such as Random Forest, ANN, SVM, MLR, and KNN were implemented and tested on the given datasets from the Maharashtra and Karnataka states. The various algorithms are compared with their accuracy. The results obtained indicate that Random Forest Regression is the best among the set of standard algorithms used on the given datasets with an accuracy of 95%. The proposed model also explored the timing of applying fertilizers and recommends appropriate duration.

The future work will be focused on updating the datasets from time to time to produce accurate predictions, and the processes can be automated. Another functionality to be implemented is to provide the correct type of fertilizer for the 1070 given crop and location. To implement this thorough study of available fertilizers and their relationship with soil and climate needs to be done. An analysis of available statistical data needs to be done.

4. REFERENCES

- [1] Umamaheswari S, Sreeram S, Kritika N, Prasanth DJ, “BIoT: Blockchain-based IoT for Agriculture”, 11th International Conference on Advanced Computing (ICoAC), 2019 Dec 18 (pp. 324-327). IEEE.
- [2] Jain A. “Analysis of growth and instability in the area, production, yield, and price of rice in India”, Journal of Social Change and Development, 2018;2:46-66
- [3] Manjula E, Djodiltachoumy S, “A model for prediction of crop yield” International Journal of Computational Intelligence and Informatics, 2017 Mar;6(4):2349-6363.
- [4] Sagar BM, Cauvery NK., “Agriculture Data Analytics in Crop Yield Estimation: A Critical Review”, Indonesian Journal of Electrical Engineering and Computer Science, 2018 Dec;12(3):1087-93.
- [5] Wolfert S, Ge L, Verdouw C, Bogaardt MJ, “Big data in smart farming– a review. Agricultural Systems”, 2017 May 1;153:69-80.
- [6] Jones JW, Antle JM, Basso B, Boote KJ, Conant RT, Foster I, Godfray HC, Herrero M, Howitt RE, Janssen S, Keating BA, “Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. Agricultural systems”, 2017 Jul 1;155:269-88.
- [7] Johnson LK, Bloom JD, Dunning RD, Gunter CC, Boyette MD, Creamer NG, “Farmer harvest decisions and vegetable loss in primary production. Agricultural Systems”, 2019 Nov 1;176:102672.
- [8] Kumar R, Singh MP, Kumar P, Singh JP, “Crop Selection Method to maximize crop yield rate using a machine learning technique”, International conference on smart technologies and management for computing, communication, controls, energy, and materials (ICSTM), 2015 May 6 (pp. 138-145). IEEE.
- [9] Sriram Rakshith.K, Dr.Deepak.G, Rajesh M, Sudharshan K S, Vasanth S, Harish Kumar N, “A Survey on Crop Prediction using Machine Learning Approach”, In International Journal for

Research in Applied Science & Engineering Technology (IJRASET), April 2019, pp(3231-3234)

[10] Veenadhari S, Misra B, Singh CD, “Machine learning approach for forecasting crop yield based on climatic parameters”, In 2014 International Conference on Computer Communication and Informatics, 2014 Jan 3 (pp. 1-5). IEEE.

[11] Ghadge R, Kulkarni J, More P, Nene S, Priya RL, “Prediction of crop yield using machine learning”, Int. Res. J. Eng. Technol. (IRJET), 2018 Feb;5.

[12] Priya P, Muthaiah U, Balamurugan M, “Predicting yield of the crop using machine learning algorithm”, International Journal of Engineering Sciences & Research Technology, 2018 Apr;7(1):1-7.

[13] S. Pavani, Augusta Sophy Beulet P., “Heuristic Prediction of Crop Yield Using Machine Learning Technique”, International Journal of Engineering and Advanced Technology (IJEAT), December 2019, pp (135-138)

[14] <https://web.dev/progressive-web-apps/>

OBJECT DETECTION AND RECOGNITION FRAMEWORK FOR THE VISUALLY IMPAIRED

Ravuri Harini (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called ‘fake news’ is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Textblob, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in a Bayesian machine learning system as a key feature to estimate the likelihood that a news article is fake. The resultant process precision is 63.333% effective at assessing the likelihood that an article with quotes is fake. This process is called influence mining and this novel technique is presented as a method that can be used to enable fake news and even propaganda detection. In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system.

1. INTRODUCTION

Intentionally deceptive content presented under the guise of legitimate journalism (or ‘fake news,’ as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake

news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

2. LITERATURE SURVEY

1) When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism

AUTHORS: M. Balmas

This research assesses possible associations between viewing fake news (i.e., political satire) and attitudes of inefficacy, alienation, and cynicism toward political candidates. Using survey data collected during the 2006 Israeli election campaign, the study provides evidence for an indirect positive effect of fake news viewing in fostering the feelings of inefficacy, alienation, and cynicism, through the mediator variable of perceived realism of fake news. Within this process, hard news viewing serves as a moderator of the association between viewing fake news and their perceived realism. It was also demonstrated that perceived realism of fake news is stronger among individuals with high exposure to fake news and low exposure to hard news than among those with high exposure to both fake and hard news. Overall, this study contributes to the scientific knowledge regarding the influence of the interaction between various types of media use on political effects.

2) Miley, CNN and The Onion

AUTHORS: D. Berkowitz and D. A. Schwartz

Following a twerk-heavy performance by Miley Cyrus on the Video Music Awards program, CNN featured the story on the top of its website. The Onion—a fake-news organization—then ran a satirical column purporting to be by CNN's Web editor explaining this decision. Through textual analysis, this paper demonstrates how a Fifth Estate comprised of bloggers, columnists and fake-news organizations worked to relocate mainstream journalism back to within its professional boundaries.

3) The Impact of Real News about “Fake News” ’ : Intertextual Processes and Political Satire

AUTHORS: P. R. Brewer, D. G. Young, and M. Morreale

This study builds on research about political humor, press metacoverage, and intertextuality to examine the effects of news coverage about political satire on audience members. The analysis uses experimental data to test whether news coverage of Stephen Colbert's Super PAC influenced knowledge and opinion regarding *Citizens United*, as well as political trust and internal political efficacy. It also tests whether such effects depended on previous exposure to *The Colbert Report* (Colbert's satirical television show) and traditional news. Results indicate that exposure to news coverage of satire can influence knowledge, opinion, and political trust. Additionally, regular satire viewers may experience stronger effects on opinion, as well as increased internal efficacy, when consuming news coverage about issues previously highlighted in satire programming.

4) Stopping Fake News

AUTHORS: M. Haigh, T. Haigh, and N. I. Kozak

Social media is acting as a double-edged sword for universe in a way of consuming news. On one side, its ease of access, popularity and low cost distribution channel lead people to gain news from social media. On other side, it is also acting as a source of spread of 'fake news'. The extensive spread of fake news on social media, websites are impacting society negatively. This makes extremely important to combat the spread of fake news and to aware the society. In this paper, we offer a review which lists out the sources of fake news, its types, generation, motivation and examples. Also, some approaches are suggested to spot and stop fake news spread.

5) With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity"

AUTHORS: R. Marchi

This article examines the news behaviors and attitudes of teenagers, an understudied demographic in the research on youth and news media. Based on interviews with 61 racially diverse high school students, it discusses how adolescents become informed about current events and why they prefer certain news formats to others. The results reveal changing ways news information is being accessed, new attitudes about what it means to be informed, and a youth preference for opinionated rather than objective news. This does not indicate that young people disregard the basic ideals of professional journalism but, rather, that they desire more authentic renderings of them.

6) Social Media and Fake News in the 2016 Election

AUTHORS: H. Allcott and M. Gentzkow

Following the 2016 US presidential election, many have expressed concern about the effects of false stories ("fake news"), circulated largely through social media. We discuss the economics of fake news and present new data on its consumption prior to the election. Drawing on web browsing data, archives of fact-checking websites, and results from a new online survey, we find: 1) social media was an important but not dominant source of election news, with 14 percent of Americans calling social media their "most important" source; 2) of the known false news stories that appeared in the three months before the election, those favoring Trump were shared a total of 30 million times on Facebook, while those favoring Clinton were shared 8 million times; 3) the average American adult saw on the order of one or perhaps several fake news stories in the months around the election, with just over half of those who recalled seeing them believing them; and 4) people are much more likely to believe stories that favor their preferred candidate, especially if they have ideologically segregated social media networks.

7) The spread of fake news by social bots.

AUTHORS: C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer

The massive spread of fake news has been identified as a major global risk and has been alleged to influence elections and threaten democracies. Communication, cognitive, social, and computer scientists are engaged in efforts to study the complex causes for the viral diffusion of digital misinformation and to develop solutions, while search and social media platforms are beginning to deploy countermeasures. However, to date, these efforts have been mainly informed by anecdotal evidence rather than systematic data. Here we analyze 14 million messages spreading 400 thousand claims on Twitter during and following the 2016 U.S. presidential campaign and election. We find evidence that social bots play a key role in the spread of fake news. Accounts that actively spread misinformation are significantly more likely to be bots. Automated accounts are particularly active in the early spreading phases of viral claims, and tend to target influential users. Humans are vulnerable to this manipulation, retweeting bots who post false news. Successful sources of false and biased claims are heavily supported by social bots. These results suggests that curbing social bots may be an effective strategy for mitigating the spread of online misinformation.

8) Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy

AUTHORS: A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi

In today's world, online social media plays a vital role during real world events, especially crisis events. There are both positive and negative effects of social media coverage of events, it can be used by authorities for effective disaster management or by malicious entities to spread rumors and fake news. The aim of this paper, is to highlight the role of Twitter, during Hurricane Sandy (2012) to spread fake images about the disaster. We identified 10,350 unique tweets containing fake images that were circulated on Twitter, during Hurricane Sandy. We performed a characterization analysis, to understand the temporal, social reputation and influence patterns for the spread of fake images. Eighty six percent of tweets spreading the fake images were retweets, hence very few were original tweets. Our results showed that top thirty users out of 10,215 users (0.3%) resulted in 90% of the retweets of fake images; also network links such as follower relationships of Twitter, contributed very less (only 11%) to the spread of these fake photos URLs. Next, we used classification models, to distinguish fake images from real images of Hurricane Sandy. Best results were obtained from Decision Tree classifier, we got 97% accuracy in predicting fake images from real. Also, tweet based features were very effective in distinguishing fake images tweets from real, while the performance of user based features was very poor. Our results, showed that, automated techniques can be used in identifying real images from fake images posted on Twitter.

9) The Fake News Spreading Plague: Was it Preventable

AUTHORS: E. Mustafaraj and P. T. Metaxas

In 2010, a paper entitled "From Obscurity to Prominence in Minutes: Political Speech and Real-time search" won the Best Paper Prize of the Web Science 2010 Conference. Among its findings were the discovery and documentation of what was termed a "Twitter-bomb", an organized effort to spread misinformation about the democratic candidate Martha Coakley through anonymous Twitter accounts. In this paper, after summarizing the details of that event, we outline the recipe of how social networks are used to spread misinformation. One of the most important steps in such a recipe is the "infiltration" of a community of users who are already engaged in conversations about a topic, to use them as organic spreaders of misinformation in their extended subnetworks. Then, we take this misinformation spreading recipe and indicate how it was successfully used to spread fake news during the 2016 U.S. Presidential Election. The main differences between the scenarios are the use of Facebook instead of Twitter, and the respective motivations (in 2010: political influence; in 2016: financial benefit through online advertising). After situating these events in the broader context of exploiting the Web, we seize this opportunity to address limitations of the reach of

research findings and to start a conversation about how communities of researchers can increase their impact on real-world societal issues.

10) Fake News Mitigation via Point Process Based Intervention.

AUTHORS: M. Farajtabar et al.

We propose the first multistage intervention framework that tackles fake news in social networks by combining reinforcement learning with a point process network activity model. The spread of fake news and mitigation events within the network is modeled by a multivariate Hawkes process with additional exogenous control terms. By choosing a feature representation of states, defining mitigation actions and constructing reward functions to measure the effectiveness of mitigation activities, we map the problem of fake news mitigation into the reinforcement learning framework. We develop a policy iteration method unique to the multivariate networked point process, with the goal of optimizing the actions for maximal total reward under budget constraints. Our method shows promising performance in real-time intervention experiments on a Twitter network to mitigate a surrogate fake news campaign, and outperforms alternatives on synthetic datasets.

3. SYSTEM ANALYSIS

EXISTING SYSTEM:

Up to now, most of the research on PDS has focused on how to enforce user privacy preferences and how to secure data when stored into the PDS. In contrast, the key issue of helping users to specify their privacy preferences on PDS data has not been so far deeply investigated. This is a fundamental issue since average PDS users are not skilled enough to understand how to translate their privacy requirements into a set of privacy preferences. As several studies have shown, average users might have difficulties in properly setting potentially complex privacy preferences.

DISADVANTAGES OF EXISTING SYSTEM:

Personal data we are digitally producing are scattered in different online systems managed by different providers (e.g., online social media, hospitals, banks, airlines, etc). In this way, on the one hand users are losing control on their data, whose protection is under the responsibility of the data provider, and, on the other, they cannot fully exploit their data, since each provider keeps a separate view of them.

PROPOSED SYSTEM:

Personal Data Storage (PDS) has inaugurated a substantial change to the way people can store and control their personal data, by moving from a service-centric to a user-centric model. PDSs enable individuals to collect into a single logical vault personal information they are producing. Such data can then be connected and exploited by proper analytical tools, as well as shared with third parties under the control of end users.

4. INPUT DESIGN AND OUTPUT DESIGN**INPUT DESIGN**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

5. CONCLUSION

This paper presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

6. REFERENCES

- [1] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 9–18.
- [2] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "Gps estimation for places of interest from social users' uploaded photos," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2058–2071, 2013.
- [3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," IEEE Transactions on Multimedia, vol. 17, no. 6, pp. 907–918, 2015.
- [4] J. Sang, T. Mei, and C. Sun, J.T.and Xu, "Probabilistic sequential pois recommendation via check-in data," in Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2012.
- [5] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," ACM Transactions on the Web, vol. 5, no. 1, p. 5, 2011.

[6] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in Proceedings of 29th International Conference on AAAI. AAAI, 2015.

FEATURE EXTRACTION AND ANALYSIS OF NATURAL LANGUAGE PROCESSING FOR DEEP LEARNING ENGLISH LANGUAGE

Ravuri Seshu (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

NLP (Natural Language Processing) is a technology that enables computers to understand human languages. Deep-level grammatical and semantic analysis usually uses words as the basic unit, and word segmentation is usually the primary task of NLP. In order to solve the practical problem of huge structural differences between different data modalities in a multi-modal environment and traditional machine learning methods cannot be directly applied, this paper introduces the feature extraction method of deep learning and applies the ideas of deep learning to multi-modal feature extraction. This paper proposes a multi-modal neural network. For each mode, there is a multilayer sub-neural network with an independent structure corresponding to it. It is used to convert the features in different modes to the same-modal features. In terms of word segmentation processing, in view of the problems that existing word segmentation methods can hardly guarantee long-term dependency of text semantics and long training prediction time, a hybrid network English word segmentation processing method is proposed. This method applies BI-GRU (Bidirectional Gated Recurrent Unit) to English word segmentation, and uses the CRF (Conditional Random Field) model to annotate sentences in sequence, effectively solving the long-distance dependency of text semantics, shortening network training and predicted time. Experiments show that the processing effect of this method on word segmentation is similar to that of BI-LSTM-CRF (Bidirectional- Long Short Term Memory-Conditional Random Field) model, but the average predicted processing speed is 1.94 times that of BI-LSTM-CRF, effectively improving the efficiency of word segmentation processing.

1. INTRODUCTION

With the rapid development of Internet information technology and the continuous advancement of science and technology, a large amount of data of various types and structures have been accumulated in the real life and scientific research fields. In the real world, for the observation target of the same semantic conceptual ontology, multiple observation methods can often be used to obtain data information from multiple different observation channels, and these data from different information channels describe the same concept. Each of these kinds of information data can be called a different modal, or different observation perspectives. Different information



modalities together constitute multi-modal data for the same problem. NLP (Natural Language Processing) is one of the key technologies for realizing human-computer interaction and artificial intelligence [1]–[3]. It is listed as the three major elements of artificial intelligence research with voice processing and image processing [4], [5]. In the early days of NLP research, the main focus was on the analysis of language structure, technology-driven machine translation, and language recognition [6]–[9]. The current focus is on how NLP can be used in the real world. The corresponding research areas include dialogue systems and social media data. However, the training of deep frames is a difficult task, and traditional shallow proven methods that have proven effective cannot be transplanted into deep learning to ensure their effectiveness [10]–[13]. Another realistic problem is that there is no necessary connection between increasing the layer structure and obtaining better feature representations. For example, in a neural network, the more hidden layers, the less impact the first layer in the backpropagation algorithm. When using the gradient descent algorithm, it will also fall into the local optimum and lose the effect of continued transmission.

Related scholars have proposed a word segmentation algorithm based on supervised machine learning. This method implements a word-based word segmentation system. The main innovation is to use the maximum entropy model as a tokenizer to automatically label characters. This method has the highest recall rate of 72.9% in the AS2003 closed test experiment [14]–[17]. In the method of English word segmentation based on the dictionary and rules, it mainly focuses on the word segmentation algorithm and dictionary structure [18], [19]. The advantages of dictionary-based and rule-based methods are simple, easy to implement, and suitable dictionaries can be formulated according to special scenarios. In addition, in systems that require real-time performance, dictionary-based and rule-based methods are often more suitable because of their high efficiency [20], [21]. The disadvantages are: there is a problem of word segmentation ambiguity; there is no universal standard for word division, so the quality of the dictionary cannot be clearly defined. The dictionary has a great impact on the segmentation result [22]–[25]. With the advent of the era of big data, data has become more and more in natural language processing problems [26], [27]. Improving these labeling problems to support parallel computing and being able to perform parallel learning on large-scale training data has also become a research hotspot [28]–[32]. Parallel learning is currently supported, including maximum entropy models and conditional random field models. Some researchers have proposed the technical route of “understand first and then segmentation” [33]–[35]. The idea of understanding the segmentation first is to solve the lack of global information in the traditional matching segmentation, while the statistical method lacks the structural information of the sentence [36]–[39]. Relevant scholars use deep learning to perform sequence labeling in the NLP field [40], [41]. It can also add a sequence labeling model to combine with the output of the previous neural network to extract the best labeling sequence through the Viterbi algorithm. Related scholars have proposed an open domain question answering system based on relationship matching [42], [43]. The problem analysis problem based on relation matching is



solved through the associated data in the question answering system. The fragments in the question match the binary relationship in the triples and are automatically collected using the relational text pattern. Existing models do not take into account the importance of different modalities for the current learning task, but only focus on how to effectively use multiple modalities for feature extraction at the same time. Moreover, the selection of modals and the filtering of harmful modals are not involved, and this issue is also an important issue addressed in this paper. In terms of word segmentation processing, in view of the problems that existing word segmentation methods can hardly guarantee long-term dependency of text semantics and long training prediction time, a hybrid network English word segmentation processing method is proposed. Experimental results show that this method improves the efficiency of natural language processing.

In terms of English word segmentation, since traditional machine learning methods cannot solve the long-distance dependencies of texts, it is difficult to analyze the information contained in the problem as a whole and grasp the user's true intention. In order to solve the above problems and save the relevance of the forward and reverse information of the text, this paper uses BI-GRU (Bidirectional Gated Recurrent Unit) neural network and combines the CRF (Conditional Random Field) model to solve the problem of sequence labeling at the sentence level analysis, based on BI-GRU-CRF (Bidirectional-Gated Recurrent Unit- Conditional Random Field) hybrid network English word segmentation processing method. Specifically, the technical contributions of this article are summarized as follows:

Firstly, a multimodal fusion feature extraction method is proposed. The problem of heterogeneity of multi-modal data is solved through the feature transformation of deep neural networks.

Secondly, in view of the problems that traditional neural network models cannot capture the long-distance dependencies of text and the long cost of training and prediction of LSTM (Long Short Term Memory) neural networks, a word segmentation processing method based on BI-GRU-CRF hybrid network is proposed.

Thirdly, the proposed method is tested from two aspects, accuracy and timeliness. According to these two sets of experiments, the proposed hybrid network word segmentation processing method has good performance in English word segmentation processing.

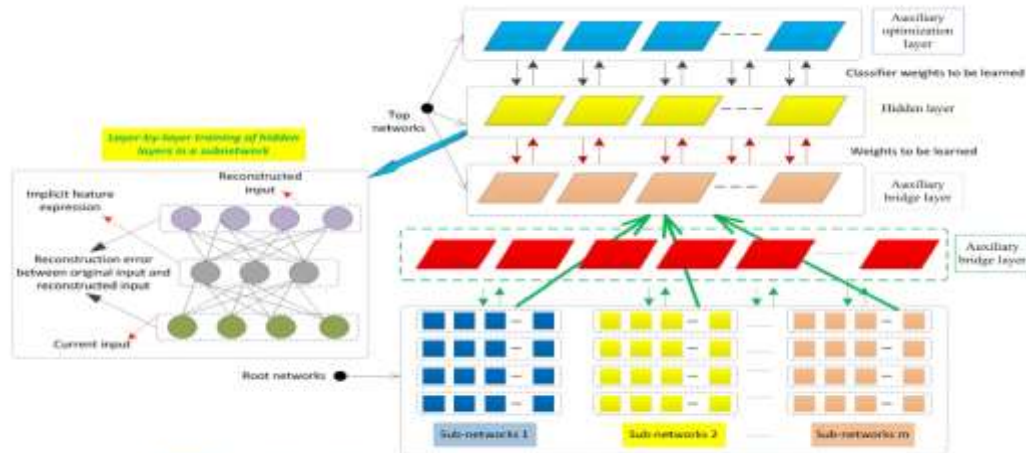


FIGURE 1 Schematic diagram of the overall structure of a semi-supervised multimodal neural network

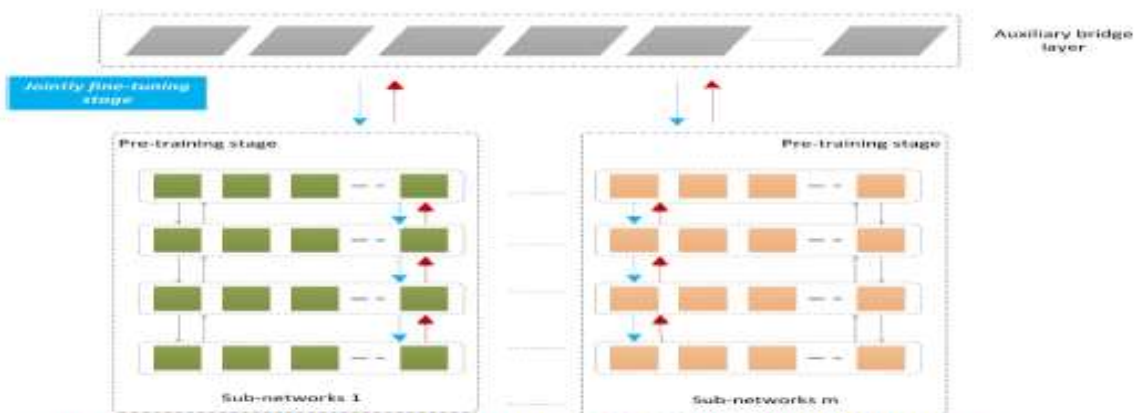


FIGURE 2 Schematic diagram of the root network structure

3. SYSTEM DESIGN

Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language

In this paper author is using Natural Language Processing with deep learning to detect English words segmentation and then evaluating performance of two deep learning neural networks called BI-GRU (Bidirectional Gated Recurrent UNIT) and BI-LSTM (Bidirectional Long Short Term Memory) and from both algorithm BI-GRU is taking less execution and giving less LOSS compare to BI-LSTM. Neural network model which give less LOSS can be consider as best model.

Word segmentation is identifying meaningful information form give data for example if we got data as 'commentsunderquestioning' then segmented output will be 'comments under questioning' and to get this output using neural network we will train neural network with all possible words and their ID's and whenever we gave such input then neural network will predict

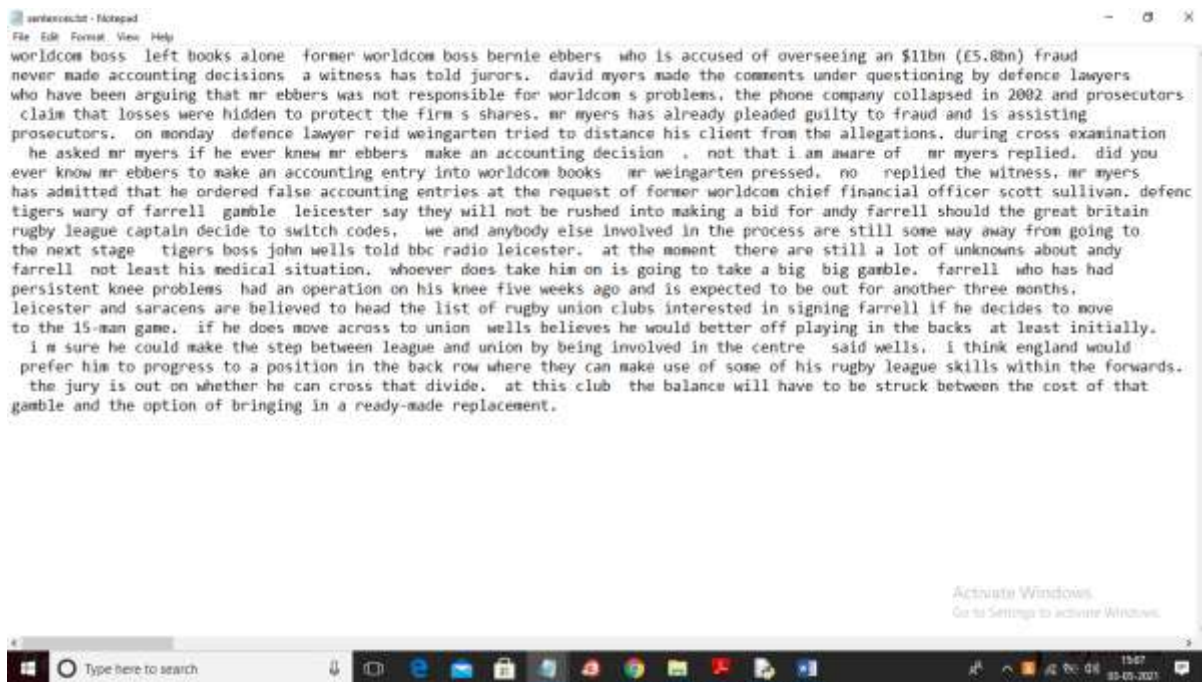


ID's of that word and then convert that ID into word and then look that word in vocabulary and if word found in vocabulary then segmented word will be identified.

About algorithms and other details you can read from paper and to implement this paper author has used WIKI dataset and this dataset contains lots of sentences and to train all those sentences may take days of time so I took few sentences which consists of 3000 words and then train both LSTM and GRU and then calculate LOSS of each model.

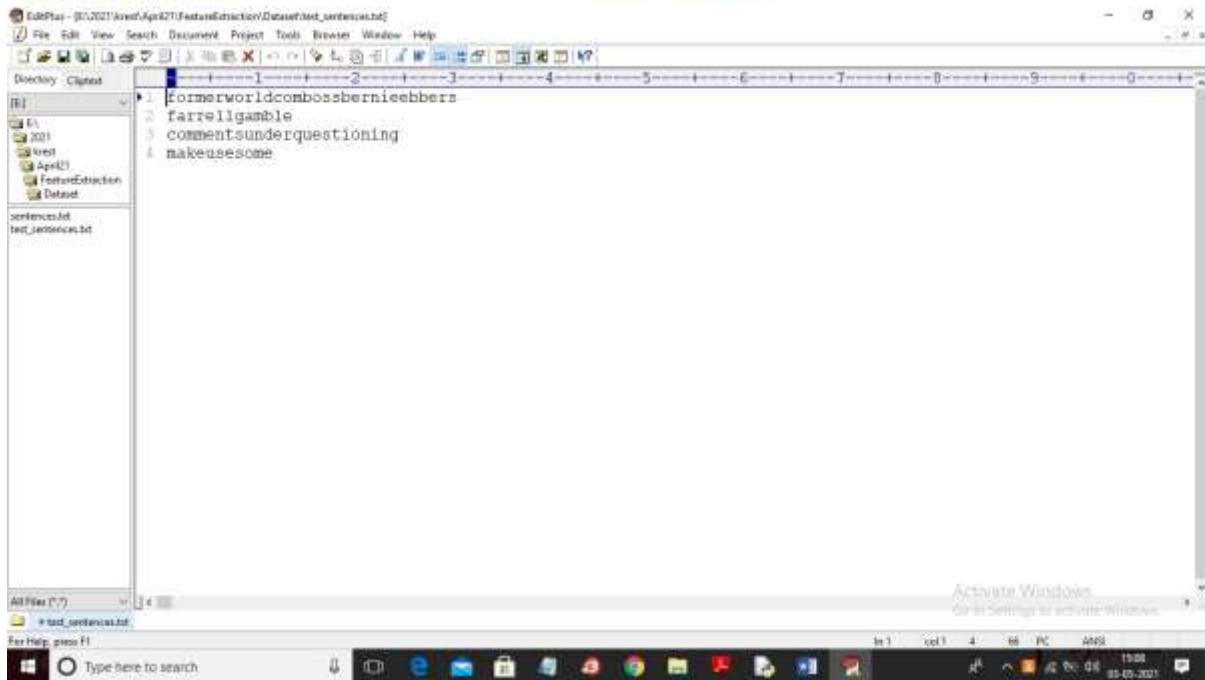
Reading all words from dataset and then preprocessing them is called as Features Extraction and then converting this features into vector is called as Natural Language Processing (NLP). Generating vector from features is referring as assigning unique ID to each word.

Below screen shots showing dataset sentences used to train above algorithms



We are using above sentences to train models and you can give any word from above sentences to get segmented output.

Below is the test words used to get segmented output



In above test data we cannot get any meaningful information so by applying GRU or LSTM we can get segmented words from above data.

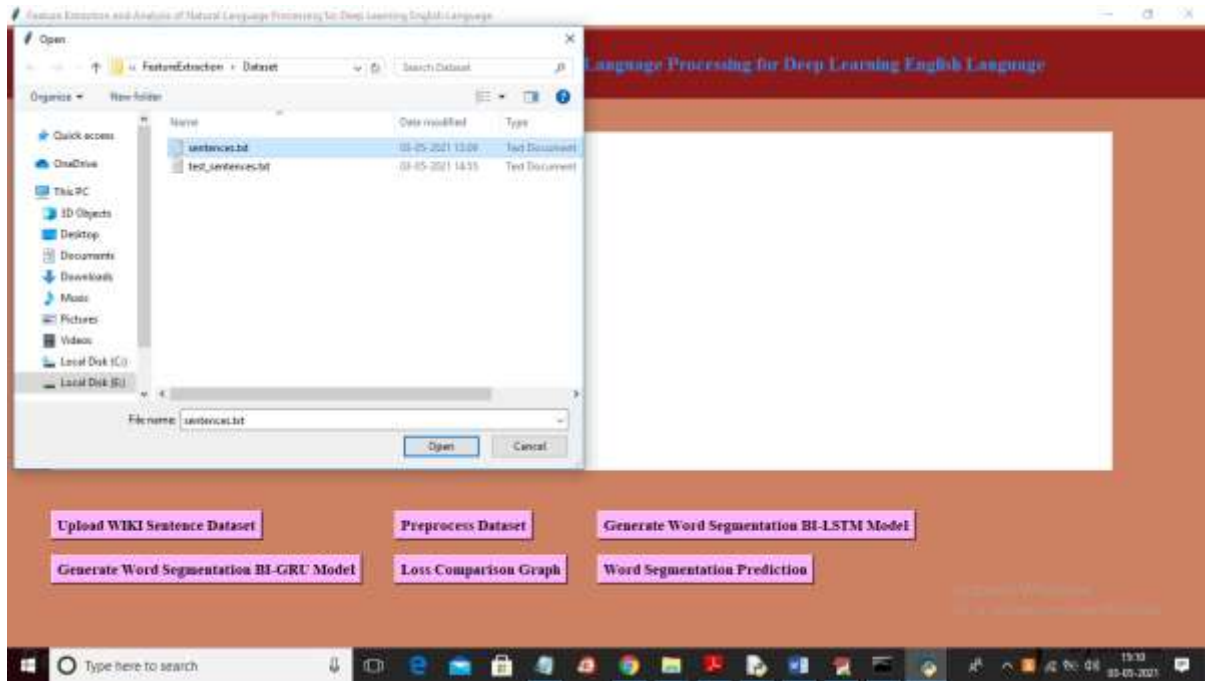
SCREEN SHOTS

To run project double click on 'run.bat' file to get below screen

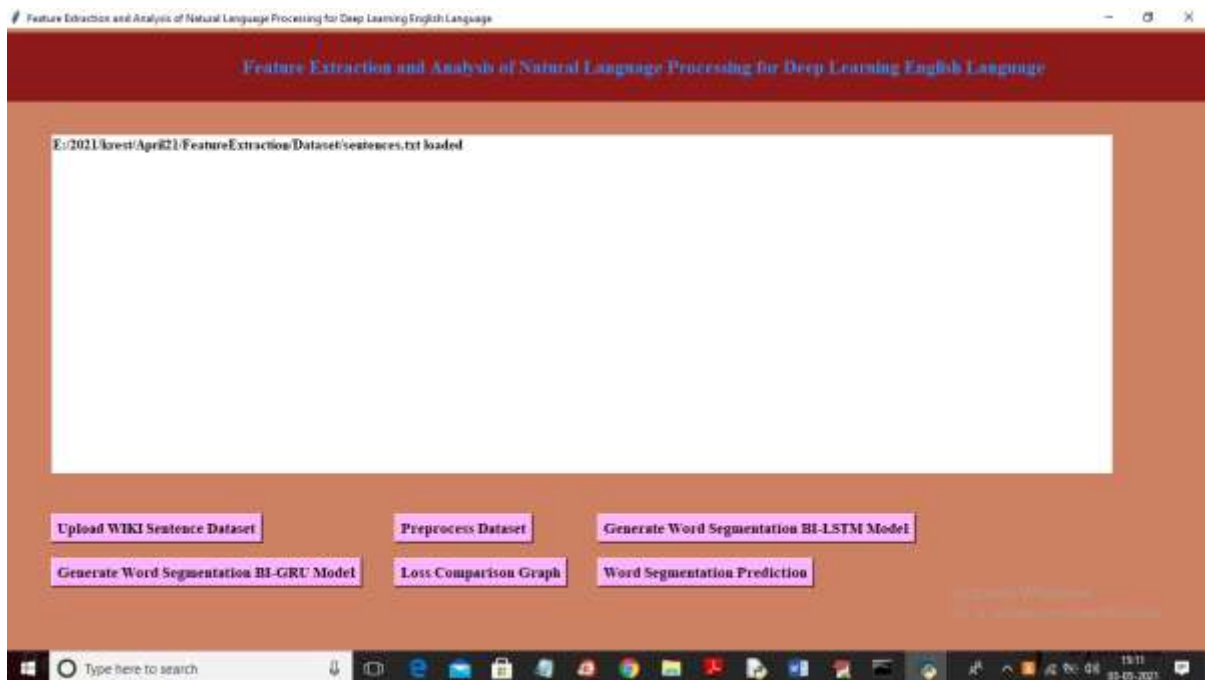




In above screen click on 'Upload WIKI Sentence Dataset' button to upload sentences dataset



In above screen selecting and uploading 'sentences.txt' file and then click on 'Open' button to load dataset and to get below screen

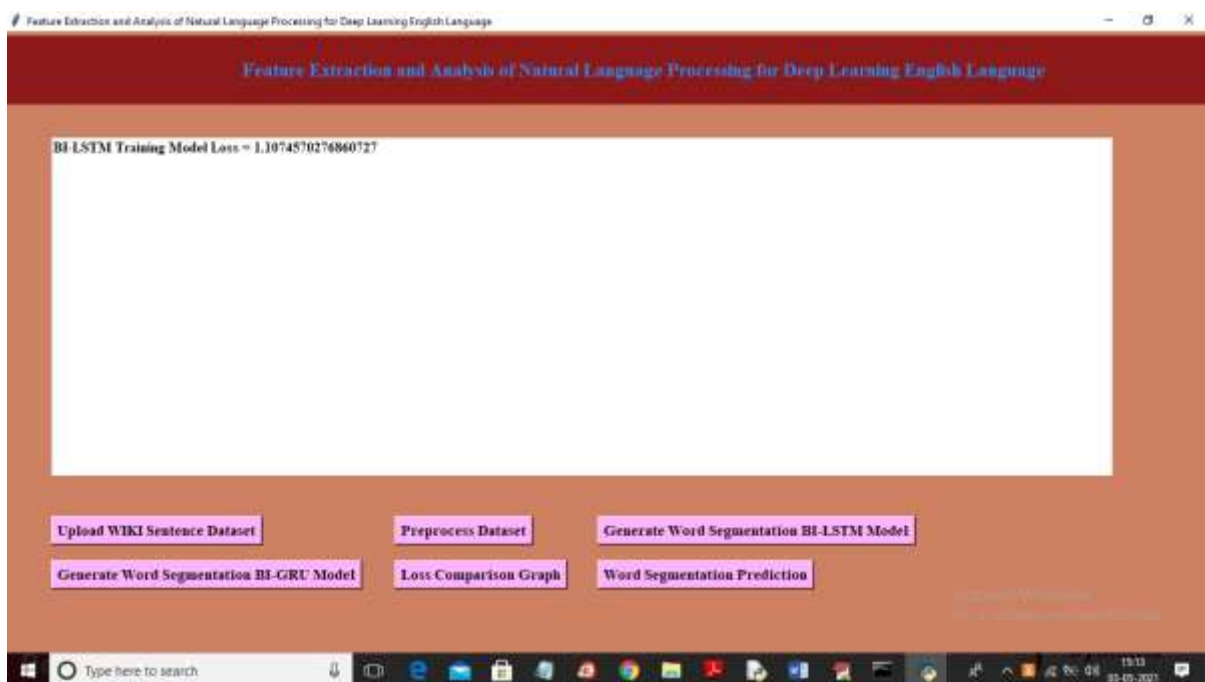




In above screen dataset loaded and now click on 'Preprocess Dataset' button to read and process dataset such as features extraction and generating vector

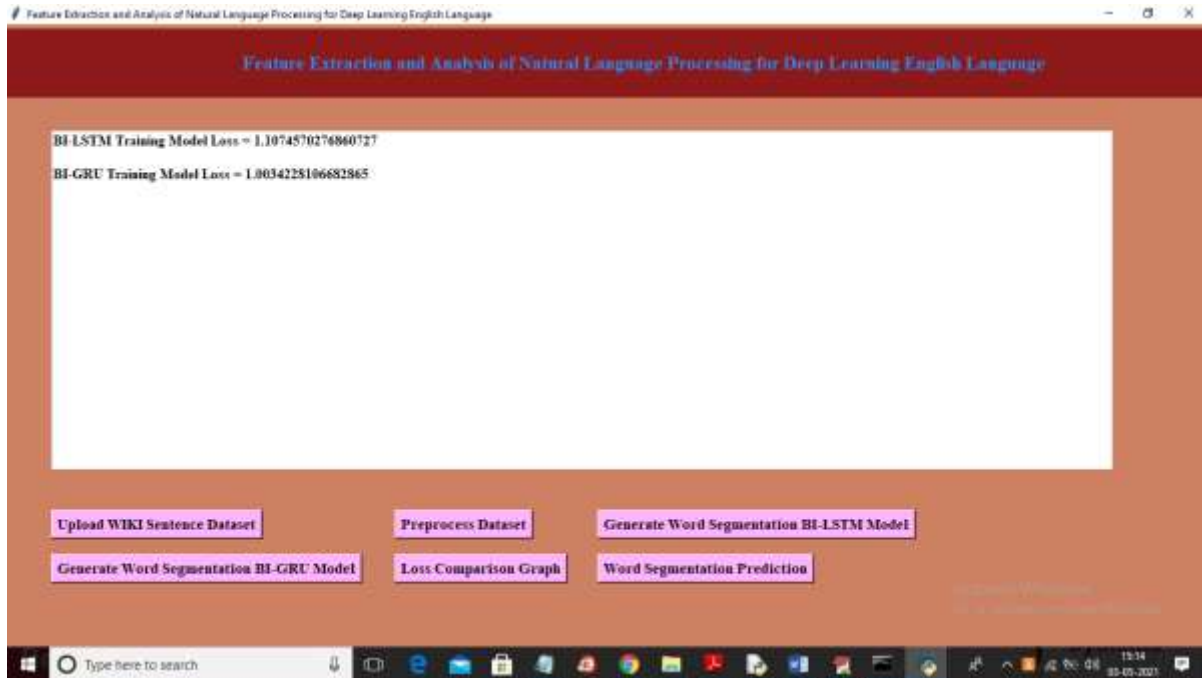


In above screen total features or characters extracted from dataset is 3185 and the vocabulary or total unique words found in dataset is 39. Now click on 'Generate Word Segmentation BI-LSTM Model' button to build LSTM model on above dataset and then calculate loss

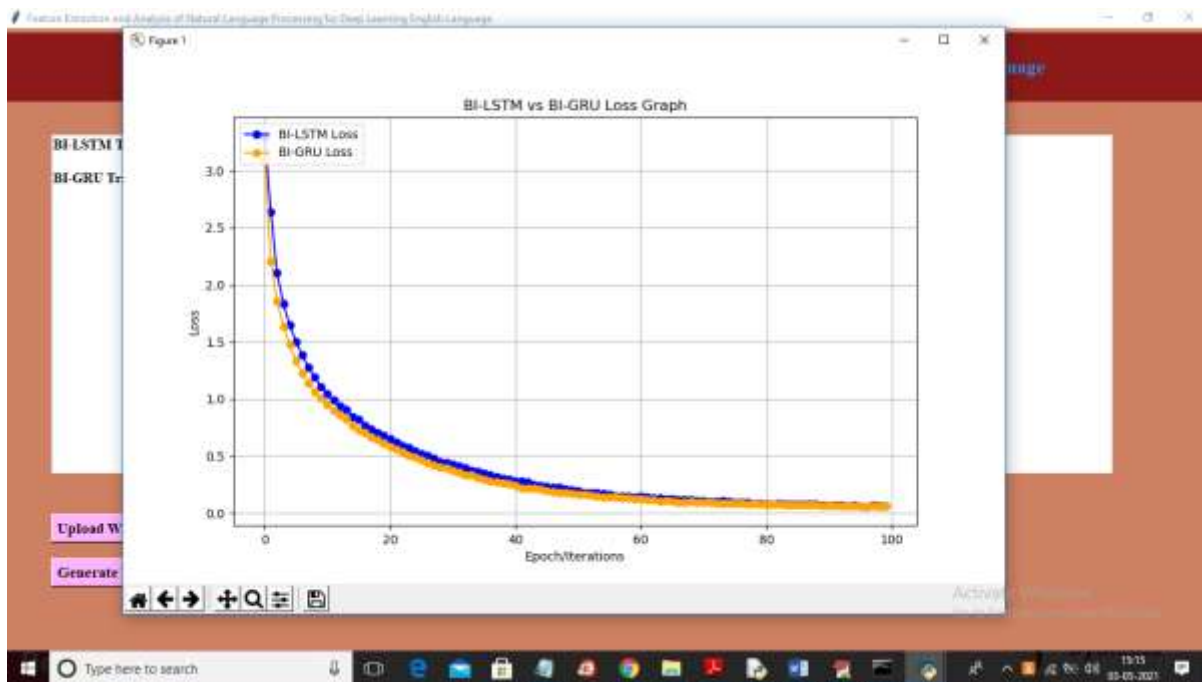




In above screen out of 100% LSTM loss reduce to 1.10% and now click on ‘Generate Word Segmentation BI-GRU Model’ button to build GRU model

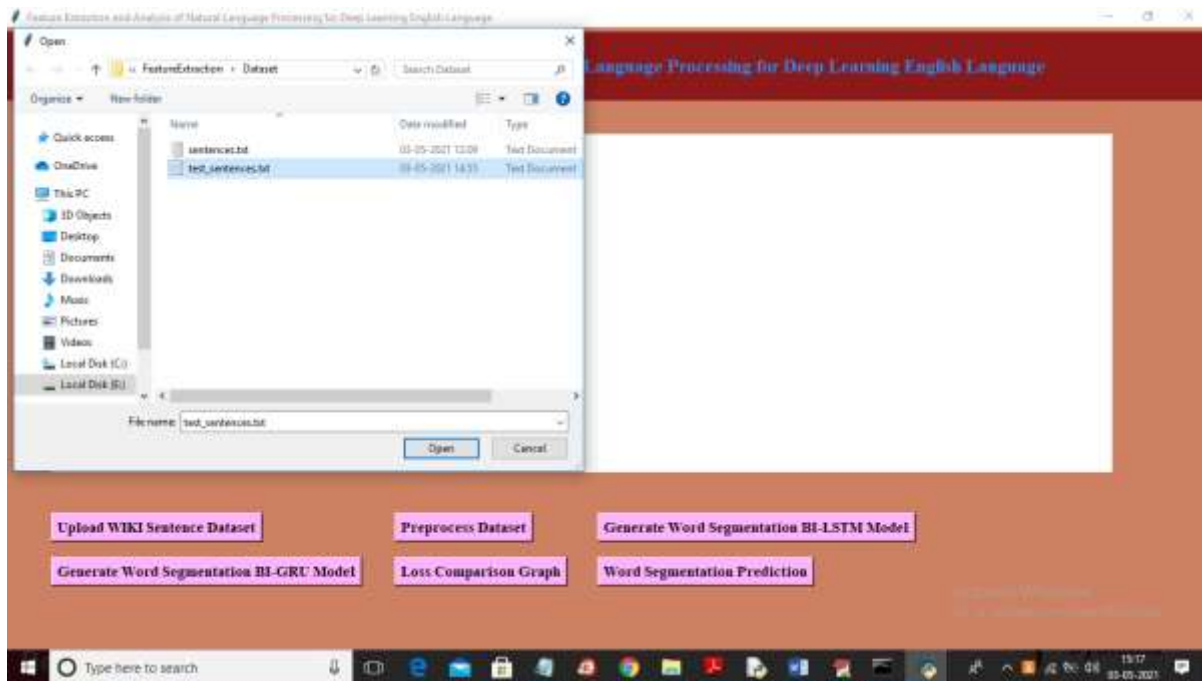


In above screen GRU loss reduce to 1.00% so GRU is better than LSTM and now click on ‘Loss Comparison Graph’ to get below graph of both LSTM and GRU loss

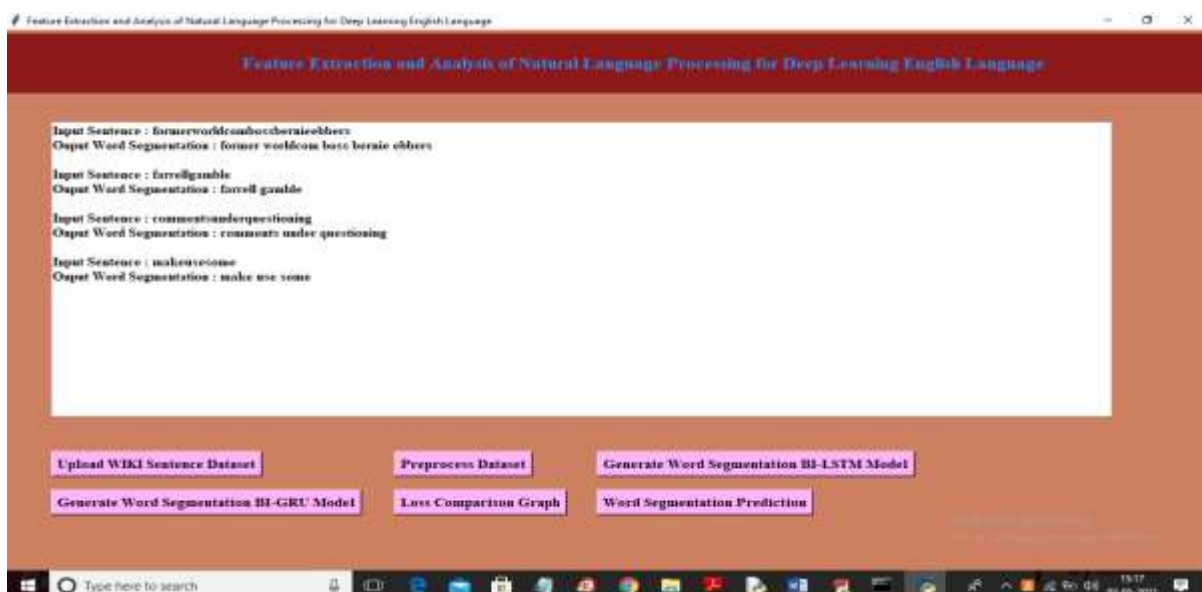




In above graph x-axis represents Epoch/Iterations and y-axis represents LOSS and blue line represents LSTM and orange line represents GRU and for both algorithms we too 100 EPCOH and both algorithms loss is reduce per increasing EPCOH but GRU loss is little less compare to LSTM so GRU is better and now click on 'Word Segmentation Prediction' button to upload test file and then will get segmented word



In above screen selecting and uploading 'test_sentences.txt' and then click on 'Open' button to get below result





In above screen we can see Input Sentence and then predicted segmented output from GRU and from above result we can see we extracted meaningful information from give data.

Note: I used few sentences to train both algorithms so application will perform segmentation on dataset words only. Actually to train large dataset application taking hours of time

4. CONCLUSION

This paper proposes a multimodal shared feature expression extraction algorithm based on deep neural network, gives the entire model structure of the algorithm, and details the design of the model structure and the model training method. In order to verify the effectiveness of the proposed model, a series of comparative experiments were carried out. The experimental results show that the proposed multimodal fusion feature extraction model can effectively extract low-dimensional fusion features from the original multiple high-dimensional data. The obtained fusion feature expression has a strong discriminative ability while possessing a lower feature dimension, thereby proving the validity of the proposed model. In terms of English word segmentation, this article has studied LSTM and GRU in depth. After analysis and research, both networks can solve the problem of traditional word segmentation in the long-range dependency relationship of text. However, due to the complexity of its structure, LSTM consumes a lot of time in the process of training and predicting the data set. The GRU is a simplified version of the LSTM. It has a simple structure and consumes less time in training and prediction. Based on the two-way network's ability to better capture the contextual relationship between semantics, this paper combines BI-GRU and CRF models, and proposes a hybrid neural network word segmentation processing method. The experimental results show that the model proposed in this paper is better than most previous models in terms of accuracy, and in terms of timeliness, the method proposed in this paper is 1.62 times faster than the BI-LSTM-CRF network word segmentation method in training speed. The average speed is 1.94 times that of the word segmentation method based on BI-LSTM-CRF network. Based on these two sets of data, the hybrid network word segmentation method proposed in this paper has good performance in English word segmentation. In future work, we can consider analyzing the impact of different feature extraction methods and feature selection methods on the model, thereby further enhancing the learning ability of the model. The proposed method treats different features obtained by different extraction methods in each kind of raw data as an independent mode, and does not learn directly on the raw data. How to get the multi-modal fusion low-dimensional features directly from the original multi-modal data needs further research.



5. REFERENCES

1.

V. K. Ha, J.-C. Ren and X.-Y. Xu, "Deep learning based single image super-resolution: A survey", *Int. J. Autom. Comput.*, vol. 16, pp. 413-426, 2019.

Show in Context [CrossRef](#) [Google Scholar](#)



2.

F. Meng, P. Chen, L. Wu and X. Wang, "Automatic modulation classification: A deep learning enabled approach", *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10760-10772, Nov. 2018.

Show in Context [View Article](#)

[Google Scholar](#)

3.

Q. Xia, S. Li, A. M. Hao and Q. P. Zhao, "Deep learning for digital geometry processing and analysis: A review", *J. Comput. Res. Develop.*, vol. 56, no. 1, pp. 155-182, 2019.

Show in Context [Google Scholar](#)



4.

Y. Chen, Y. Zhang, S. Maharjan, M. Alam and T. Wu, "Deep learning for secure mobile edge computing in cyber-physical transportation systems", *IEEE Netw.*, vol. 33, no. 4, pp. 36-41, Jul. 2019.

Show in Context [View Article](#)

[Google Scholar](#)

5.

K. A. Weber, A. C. Smith, M. Wasielewski, K. Egtesad, P. A. Upadhyayula, M. Wintermark, et al., "Deep learning convolutional neural networks for the automatic quantification of muscle fat infiltration following whiplash injury", *Sci. Rep.*, vol. 9, no. 1, pp. 7973, May 2019.

Show in Context [CrossRef](#) [Google Scholar](#)



6.

O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved", *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514-2525, Nov. 2018.

Show in Context [View Article](#)

[Google Scholar](#)

7.

M. Abdughani, J. Ren, L. Wu, J.-M. Yang and J. Zhao, "Supervised deep learning in high energy phenomenology: A mini review", *Commun. Theor. Phys.*, vol. 71, no. 8, pp. 955, Aug. 2019.

Show in Context [CrossRef](#) [Google Scholar](#)



8.

R. Ranjan, S. Sankaranarayanan, A. Bansal, N. Bodla, J.-C. Chen, V. M. Patel, et al., "Deep learning for understanding faces: Machines may be just as good or better than humans", *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 66-83, Jan. 2018.

Show in Context [View Article](#)

[Google Scholar](#)



IJARST

International Journal For Advanced Research In Science & Technology

A peer reviewed international journal

ISSN: 2457-0362

www.ijarst.in

9.

Y. Tian and X. Liu, "A deep adaptive learning method for rolling bearing fault diagnosis using immunity", *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 750-762, Dec. 2019.

Show in Context [View Article](#)

[Google Scholar](#)

10.

C. Wu, R. Zeng, J. Pan, C. C. L. Wang and Y.-J. Liu, "Plant phenotyping by deep-learning-based planner for multi-robots", *IEEE Robot. Autom. Lett.*, vol. 4, no. 4, pp. 3113-3120, Oct. 2019.

Show in Context [View Article](#)

[Google Scholar](#)

MACHINE LEARNING FOR WEB VULNERABILITY DETECTION

Ravuri Venkata Arjun Naidu (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr.I.R.Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

In this project, we propose a methodology to leverage Machine Learning (ML) for the detection of web application vulnerabilities. Web applications are particularly challenging to analyse, due to their diversity and the widespread adoption of custom programming practices. ML is thus very helpful for web application security: it can take advantage of manually labeled data to bring the human understanding of the web application semantics into automated analysis tools. We use our methodology in the design of Mitch, the first ML solution for the black-box detection of CrossSite Request Forgery (CSRF) vulnerabilities. Mitch allowed us to identify 35 new CSRFs on 20 major websites and 3 new CSRFs on production software.

1. INTRODUCTION

Web applications are the most common interface to security sensitive data and functionality available nowadays. They are routinely used to file tax incomes, access the results of medical screenings, perform financial transactions, and share opinions with our circle of friends, just to mention a few popular use cases. On the downside, this means that web applications are appealing targets to malicious users (attackers) who are determined to force economic losses, unduly access confidential data or create embarrassment to their victims. Securing web applications is well known to be hard.

There are several reasons for this, ranging from the heterogeneity and complexity of the web platform to the adoption of undisciplined scripting languages offering dubious security guarantees and not amenable for static analysis. In such a setting, black-box vulnerability detection methods are particularly popular. As opposed to white-box techniques which require access to the web application source code, black-box methods operate at the level of HTTP traffic, i.e., HTTP

requests and responses. Though this limited perspective might miss important insights, it has the key advantage of offering a language-agnostic vulnerability detection approach, which abstracts from the complexity of scripting languages and offers a uniform interface to the widest possible range of web applications. This sounds appealing, yet previous work showed that such an analysis is far from trivial. One of the main challenges there is how to expose to automated tools a critical ingredient of effective vulnerability detection, i.e., an understanding of the web application semantics. Example: Cross-Site Request Forgery (CSRF) Cross-Site Request Forgery (CSRF) is a well-known web attack that forces a user into submitting unwanted, attacker controlled HTTP requests towards a vulnerable web application in which she is currently authenticated. The key concept of CSRF is that the malicious requests are routed to the web application through the user's browser, hence they might be indistinguishable from intended benign requests which were actually authorized by the user.

A typical CSRF attack works as follows:

- 1) Alice logs into an honest yet vulnerable web application, e.g., her preferred social network. Session authentication is implemented through a session cookie that is automatically attached by the browser to any subsequent request towards the web application;
- 2) Alice opens another tab and visits an unrelated website, e.g., a newspaper website, which returns a web page including malicious advertisement;
- 3) The malicious advertisement sends a cross-site request to the social network using HTML or JavaScript, e.g., asking to "like" a given political party.

Since the request includes Alice's cookies, it is processed in her authentication context at the social network. This way, the malicious advertisement can force Alice into putting a "like" to the desired political party, which might skew the result of online surveys.

Notice that CSRF does not require the attacker to intercept or modify user's requests and responses: it suffices that the Preventing CSRF

To prevent CSRF, web developers have to implement explicit protection mechanisms. If adding extra user interaction does not affect usability too much, it is possible to force re-authentication or use one-time passwords / CAPTCHAs to prevent cross-site requests going through unnoticed. In many cases, however, automated prevention is preferred: the recently introduced SameSite cookie attribute can be used to prevent cookie attachment on cross-site requests, which solves the root cause of CSRF and is highly recommended for new web applications. Unfortunately, this defense

is not yet widespread and existing web applications typically filter out cross-site request by using any of the following techniques:

- 1) checking the value of standard HTTP request headers such as Referrer and Origin, indicating the page originating the request;
- 2) checking the presence of custom HTTP request headers like X-Requested-With, which cannot be set from a cross-site position;
- 3) checking the presence of unpredictable anti-CSRF tokens, set by the server into sensitive forms.

A recent paper discusses the pros and cons of these different solutions. However, all three options suffer from the same limitation: they require a careful and fine-grained placement of security checks. For example, tokens should be attached to all and only the security-sensitive HTTP requests, so as to ensure complete protection without harming the user experience.

Using a token to protect a “like” button is useful to prevent the attack discussed above, yet having a token on the social network homepage is undesirable, because it might lead to rejecting legitimate cross-site requests, e.g., from clicks on the results of a search engine indexing the social network. In the end, finding the “optimal” placement of anti-CSRF defenses is typically a daunting task for web developers. Modern web application development frameworks provide Automated support for this, yet CSRF vulnerabilities are still routinely found even in top-ranked websites. This motivates the need for effective CSRF detection tools. But how can we provide automated tool support for CSRF detection if we have no mechanized way to detect which HTTP requests are actually security-sensitive. are passed - No splits.

This work presents the most current and comprehensive understanding of a not very well understood web vulnerability known as the CSRF (Cross-Site Request Forgery) and provides specific solutions to identify and defend CSRF vulnerabilities. The immediate benefits of this work include tangible and pragmatic application framework for use by individuals, organizations and developers, either as consumers or providers of web services. This work responds directly to the challenges of keeping pace with the evolving cyber technologies and vulnerabilities that increasingly expose businesses towards privacy and identity theft specific attacks, where the traditional anti-virus and anti-spyware approaches fail. The urgency to come up with appropriate detection and defense mechanism against the lethal CSRF attacks is indicated due to expanding

cloud based technologies, HTML5, Semantic Web, and various emerging security frameworks comprised of inchoate vestigial of “Big Data” that demand exceedingly evolved defense mechanisms. A methodical approach is used to investigate CSRF attacks and remedies are proposed by introducing a novel distinctive set of algorithms that use intelligent assumptions to detect and defend CSRF. In this work, design details of a CSRF Detection Model (CDM), implantation and experimentation results of CDM are elaborated to detect, predict and provide solutions for CSRF attacks on contemporary Web Applications and Web Services environment. Additionally, CDM based recommendations for users and providers of cyber security products and services are presented. Cross-Site Request Forgery (CSRF) attack causes actions on a web application without the knowledge of the user in an authenticated browser session. CSRF attacks specifically target state-changing requests like transferring funds, changing email address, and so forth. If the victim is an administrative account, CSRF can compromise the entire web application. CSRF, also known as the Sleeping Giant, was considered to be one of the top 5 web vulnerabilities only 4 years ago. Even so, at least 270 incidents of CSRF attacks have been reported as of 2016. Not much has improved in terms of new CSRF solutions since the CSRF problem appeared in the horizon in 2010. Cross-Site Reference Forgery (CSRF) and Cross-Site Scripting (XSS) vulnerabilities have received much attention recently. An XSS attack, one of the top 3 current cyber security challenges, occurs when an attacker injects malicious code (typically JavaScript), including a CSRF attack code, into a site for the purpose of targeting users of the site, e.g., sites that allow posting comments. According to the Open Web Application Security Project (OWASP), an open web community dedicated to address cyber security challenges, CSRF is one of the top eight cyber security vulnerabilities in the world, today. While CSRF attacks are simple to create and exploit, amazingly, they are difficult to identify and mitigate.

A search for “Cross Site Scripting” (which differs from CSRF) on the ACM Digital Library returned 117 papers, while a search for “CSRF” returned only four papers. A search for “XSS” on Safari Books Online (a collection of over 5000 books on technology) showed the term appeared in 96 books, while “CSRF OR XSRF” appeared in only 13 books. Very few CSRF solutions are developed and implemented. Even so, while current solutions still lack common applicability all the pieces for large scale massive CSRF attacks are already in place [53]. This state of the current relentless CSRF attacks and meager defenses dynamics is the primary motivation for undertaking this study.

2. LITERATURE SURVEY

1) **Surviving The Web: A Journey Into Web Session Security** AUTHORS: **Stefano Calzavara, Riccardo Focardi, Marco Squarcina, and Mauro Tempesta**

The Web is the primary access point to on-line data and applications. It is extremely complex and variegated, as it integrates a multitude of dynamic contents by different parties to deliver the greatest possible user experience. This heterogeneity makes it very hard to effectively enforce security, since putting in place novel security mechanisms typically prevents existing websites from working correctly or negatively affects the user experience, which is generally regarded as unacceptable, given the massive user base of the Web. However, this continuous quest for usability and backward compatibility had a subtle effect on web security research: designers of new defensive mechanisms have been extremely cautious and the large majority of their proposals consists of very local patches against very specific attacks. This piecemeal evolution hindered a deep understanding of many subtle vulnerabilities and problems, as testified by the proliferation of different threat models against which different proposals have been evaluated, occasionally with quite diverse underlying assumptions. It is easy to get lost among the multitude of proposed solutions and almost impossible to understand the relative benefits and drawbacks of each single proposal without a full picture of the existing literature. In this work, we take the delicate task of performing a systematic overview of a large class of common attacks targeting the current Web and the corresponding security solutions proposed so far. We focus on attacks against web sessions, i.e., attacks which target honest web browser users establishing an authenticated session with a trusted web application. This kind of attacks exploits the intrinsic complexity of the Web by tampering, e.g., with dynamic contents, client-side storage or crossdomain links, so as to corrupt the browser activity and/or network communication. Our choice is motivated by the fact that attacks against web sessions cover a very relevant subset of serious web security incidents and many different defenses, operating at different levels, have been proposed to prevent these attacks.

We consider typical attacks against web sessions and we systematise them based on: (i) their attacker model and (ii) the security properties they break. This first classification is useful to understand precisely which intended security properties of a web session can be violated by a certain attack and how. We then survey existing security solutions and mechanisms that prevent or mitigate the different attacks and we evaluate each proposal with respect to the security guarantees it provides. When security is guaranteed only under certain assumptions, we make these assumptions explicit. For each security solution, we also evaluate its impact on both compatibility and usability, as well as its ease of deployment. These are important criteria to judge the practicality of a certain solution and they are useful to understand to which extent each solution, in its current state, may be amenable for a large-scale adoption on the Web. Moreover, since there are several proposals in the literature which aim at providing robust safeguards against multiple attacks, we also provide an overview of them. For each of these proposals, we discuss which attacks it prevents with respect to the attacker model considered in its original design and we assess its adequacy according to the criteria described above.

2) Large-Scale Analysis & Detection Of Authentication Cross-Site Request Forgeries

AUTHORS: Avinash Sudhodanan, Roberto Carbone, Luca Compagna, Nicolas Dolgin, Alessandro Armando, and Umberto Morelli

Cross-Site Request Forgery (CSRF) attacks are one of the critical threats to web applications. In this paper, we focus on CSRF attacks targeting web sites' authentication and identity management functionalities. We will refer to them collectively as Authentication CSRF (AuthCSRF in short). We started by collecting several Auth-CSRF attacks reported in the literature, then analyzed their underlying strategies and identified 7 security testing strategies that can help a manual tester uncover vulnerabilities enabling Auth-CSRF. In order to check the effectiveness of our testing strategies and to estimate the incidence of Auth-CSRF, we conducted an experimental analysis considering 300 web sites belonging to 3 different rank ranges of the Alexa global top 1500. The results of our experiments are alarming: out of the 300 web sites we considered, 133 qualified for conducting our experiments and 90 of these suffered from at least one vulnerability enabling Auth-CSRF (i.e. 68%). We further generalized our testing strategies, enhanced them with the knowledge we acquired during our experiments and implemented them as an extension (namely CSRF-checker) to the open-source penetration testing tool OWASP ZAP. With the help of CSRFchecker, we tested 132 additional web sites (again from the Alexa global top 1500) and identified 95 vulnerable ones (i.e. 72%). Our findings include serious vulnerabilities among the web sites of Microsoft, Google, eBay etc. Finally, we responsibly disclosed our findings to the affected vendors.

3) State Of The Art: Automated Black-Box Web Application Vulnerability Testing

AUTHORS : Jason Bau, Elie Bursztein, Divij Gupta, and John C. Mitchell

Black-box web application vulnerability scanners are automated tools that probe web applications for security vulnerabilities. In order to assess the current state of the art, we obtained access to eight leading tools and carried out a study of: (i) the class of vulnerabilities tested by these scanners, (ii) their effectiveness against target vulnerabilities, and (iii) the relevance of the target vulnerabilities to vulnerabilities found in the wild. To conduct our study we used a custom web application vulnerable to known and projected vulnerabilities, and previous versions of widely used web applications containing known vulnerabilities. Our results show the promise and effectiveness of automated tools, as a group, and also some limitations. In particular, "stored" forms of Cross Site Scripting (XSS) and SQL Injection (SQLI) vulnerabilities are not currently found by many tools. Because our goal is to assess the potential of future research, not to evaluate specific

vendors, we do not report comparative data or make any recommendations about purchase of specific tools.

4) Why johnny can't pentest: An analysis of black-box web vulnerability scanners

AUTHORS : Adam Doup'e, Marco Cova, and Giovanni Vigna

Black-box web vulnerability scanners are a class of tools that can be used to identify security issues in web applications. These tools are often marketed as “point-and-click pentesting” tools that automatically evaluate the security of web applications with little or no human support. These tools access a web application in the same way users do, and, therefore, have the advantage of being independent of the particular technology used to implement the web application. However, these tools need to be able to access and test the application's various components, which are often hidden behind forms, JavaScript-generated links, and Flash applications. This paper presents an evaluation of eleven black-box web vulnerability scanners, both commercial and open-source. The evaluation composes different types of vulnerabilities with different challenges to the crawling capabilities of the tools. These tests are integrated in a realistic web application. The results of the evaluation show that crawling is a task that is as critical and challenging to the overall ability to detect vulnerabilities as the vulnerability detection techniques themselves, and that many classes of vulnerabilities are completely overlooked by these tools, and thus research is required to improve the automated detection of these flaws.

5) Mitch: A Machine Learning Approach To The Blackbox Detection Of Csrp Vulnerabilities

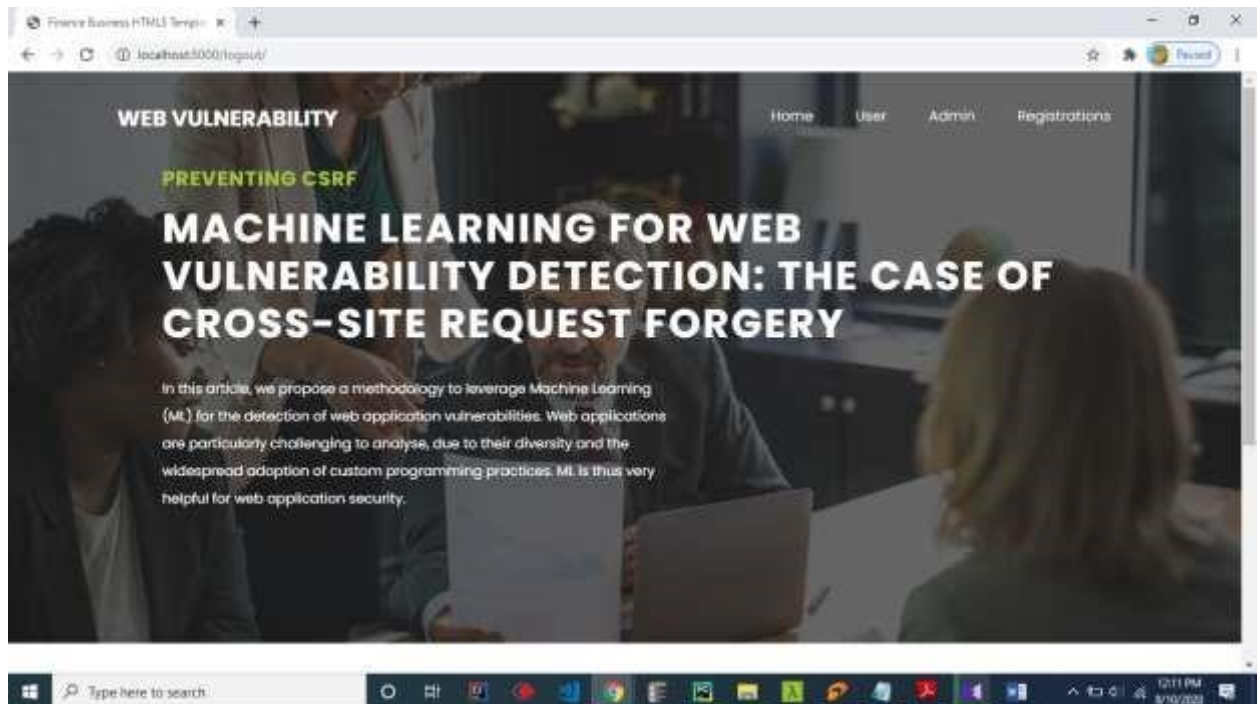
AUTHORS: Stefano Calzavara, Mauro Conti, Riccardo Focardi, Alvisè Rabitti, and Gabriele Tolomei

Cross-Site Request Forgery (CSRF) is one of the oldest and simplest attacks on the Web, yet it is still effective on many websites and it can lead to severe consequences, such as economic losses and account takeovers. Unfortunately, tools and techniques proposed so far to identify CSRF vulnerabilities either need manual reviewing by human experts or assume the availability of the source code of the web application. In this paper we present Mitch, the first machine learning solution for the black-box detection of CSRF vulnerabilities. At the core of Mitch there is an automated detector of sensitive HTTP requests, i.e., requests which require protection against CSRF for security reasons. We trained the detector using supervised learning techniques on a dataset of 5,828 HTTP requests collected on popular websites, which we make available to other

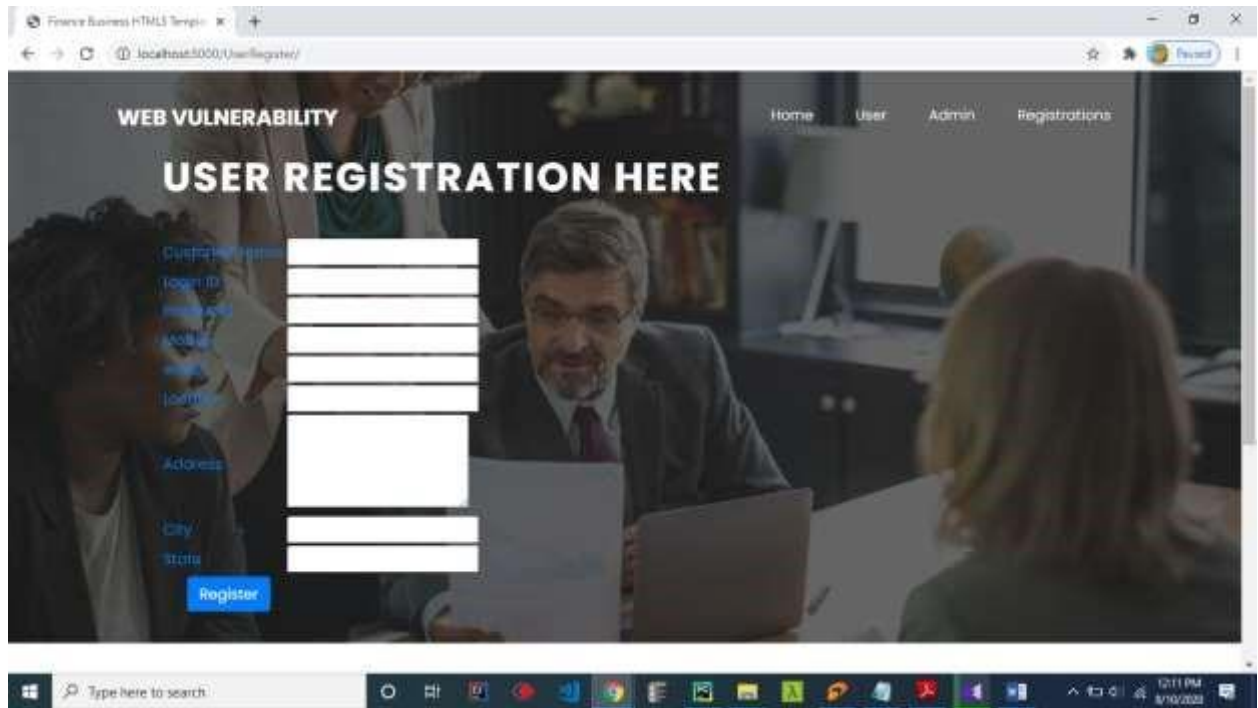
security researchers. Our solution outperforms existing detection heuristics proposed in the literature, allowing us to identify 35 new CSRF vulnerabilities on 20 major websites and 3 previously undetected CSRF vulnerabilities on production software already analyzed using a state-of-the-art tool.

3. SCREEN SHOTS

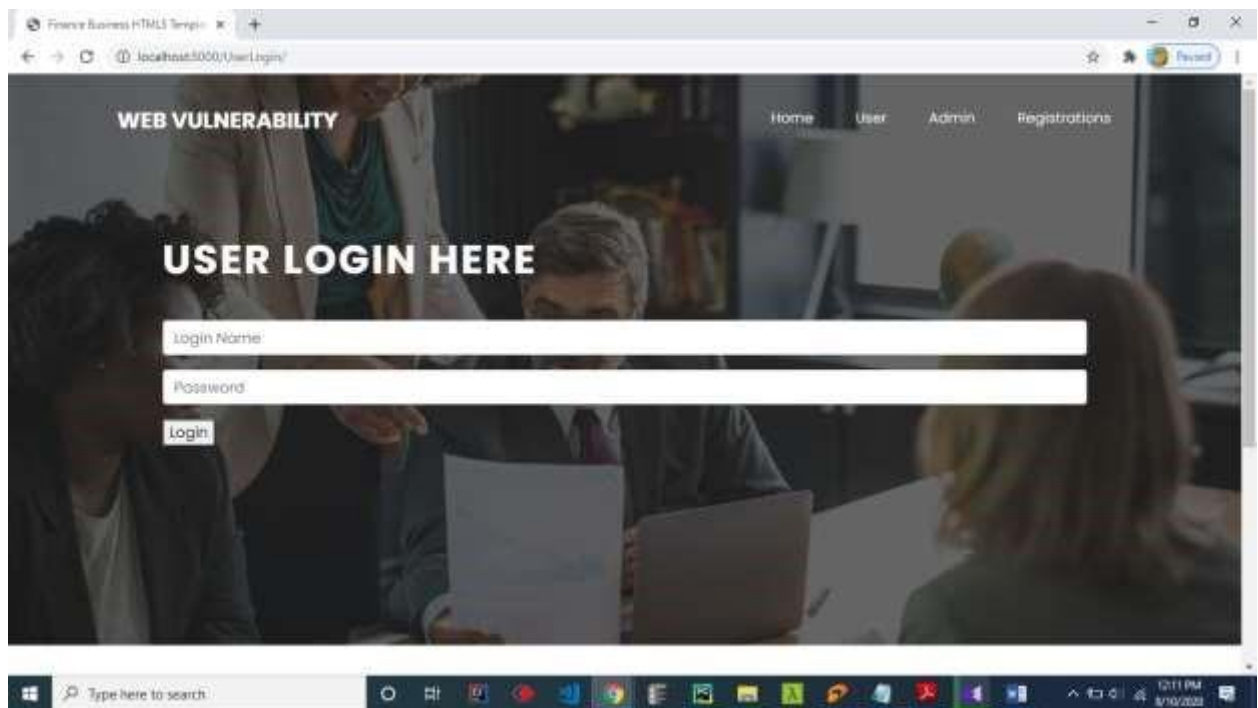
Home page:



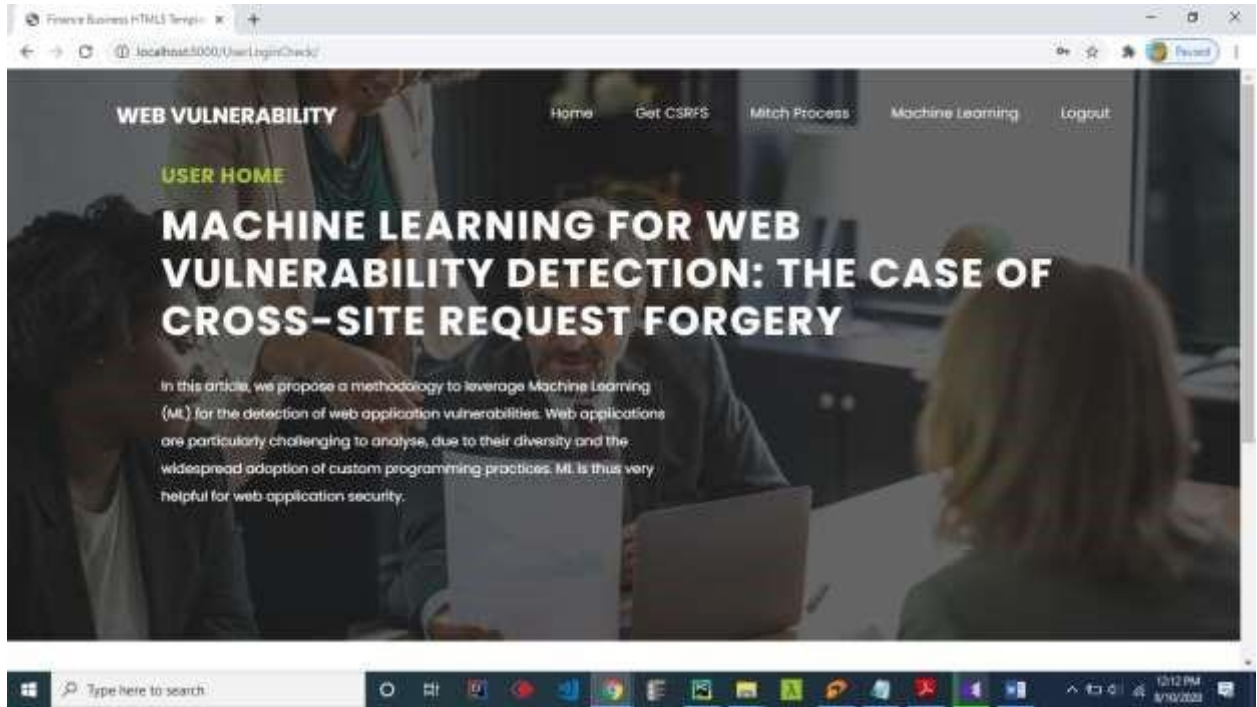
User Registration Form



User Login Form:



User Home:



Getting website csrfs:



Scanning urls:

request forgeries. In 2017 IEEE European Symposium on Security and Privacy, EuroS&P 2017, Paris, France, April 26-28, 2017, pages 350–365, 2017.

[3] Stefano Calzavara, Alvise Rabitti, Alessio Ragazzo, and Michele Bugliesi. Testing for integrity flaws in web sessions. In Computer Security - 24rd European Symposium on Research in Computer Security, ESORICS 2019, Luxembourg, Luxembourg, September 23-27, 2019, pages 606–624, 2019.

[4] OWASP. OWASP Testing Guide. https://www.owasp.org/index.php/OWASP_Testing_Guide_v4_Table_of_Contents, 2016.

[5] Jason Bau, Elie Bursztein, Divij Gupta, and John C. Mitchell. State of the art: Automated black-box web application vulnerability testing. In 31st IEEE Symposium on Security and Privacy, S&P 2010, 16-19 May 2010, Berkeley/Oakland, California, USA, pages 332–345, 2010.

[6] Adam Doup'è, Marco Cova, and Giovanni Vigna. Why johnny can't pentest: An analysis of black-box web vulnerability scanners. In Detection of Intrusions and Malware, and Vulnerability Assessment, 7th International Conference, DIMVA 2010, Bonn, Germany, July 8-9, 2010. Proceedings, pages 111–131, 2010.

[7] Adam Barth, Collin Jackson, and John C. Mitchell. Robust defenses for cross-site request forgery. In Proceedings of the 2008 ACM Conference on Computer and Communications Security, CCS 2008, Alexandria, Virginia, USA, October 27-31, 2008, pages 75–88, 2008. [8] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. The MIT Press, 2012.

[9] Michael W. Kattan, Dennis A. Adams, and Michael S. Parks. A comparison of machine learning with human judgment. *Journal of Management Information Systems*, 9(4):37–57, March 1993

[10] D. A. Ferrucci. Introduction to “This is Watson”. *IBM Journal of Research and Development*, 56(3):235–249, May 2012.

[11] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, Jan 2016.

- [12] Michele Bugliesi, Stefano Calzavara, Riccardo Focardi, and Wilayat Khan. Cookiext: Patching the browser against session hijacking attacks. *Journal of Computer Security*, 23(4):509–537, 2015.
- [13] Stefano Calzavara, Gabriele Tolomei, Andrea Casini, Michele Bugliesi, and Salvatore Orlando. A supervised learning approach to protect client authentication on the web. *TWEB*, 9(3):15:1–15:30, 2015.
- [14] Stefano Calzavara, Mauro Conti, Riccardo Focardi, Alvisè Rabitti, and Gabriele Tolomei. Mitch: A machine learning approach to the blackbox detection of CSRF vulnerabilities. In *IEEE European Symposium on Security and Privacy, EuroS&P 2019, Stockholm, Sweden, June 17-19, 2019*, pages 528–543, 2019.
- [15] Giancarlo Pellegrino, Martin Johns, Simon Koch, Michael Backes, and Christian Rossow. Deemon: Detecting CSRF with dynamic analysis and property graphs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, Dallas, TX, USA, October 30 - November 03, 2017*, pages 1757–1771, 2017.

DESIGNING CYBER INSURANCE POLICIES: THE ROLE OF PRE-SCREENING AND SECURITY INTERDEPENDENCE

Revu Balaji (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT:

Cyber insurance is a viable method for cyber risk transfer. However, it has been shown that depending on the features of the underlying environment, it may or may not improve the state of network security. In this paper, we consider a single profit-maximizing insurer (principal) with voluntarily participating insureds/clients (agents). We are particularly interested in two distinct features of cybersecurity and their impact on the contract design problem. The first is the interdependent nature of cybersecurity, whereby one entity's state of security depends not only on its own investment and effort, but also the efforts of others' in the same eco-system (i.e. externalities). The second is the fact that recent advances in Internet measurement combined with machine learning techniques now allow us to perform accurate quantitative assessments of security posture at a firm level. This can be used as a tool to perform an initial security audit, or prescreening, of a prospective client to better enable premium discrimination and the design of customized policies. We show that security interdependency leads to a "profit opportunity" for the insurer, created by the inefficient effort levels exerted by interdependent agents who do not account for the risk externalities when insurance is not available; this is in addition to risk transfer that an insurer typically profits from. Security pre-screening then allows the insurer to take advantage of this additional profit opportunity by designing the appropriate contracts which incentivize agents to increase their effort levels, allowing the insurer to "sell commitment" to interdependent agents, in addition to insuring their risks. We identify conditions under which this type of contracts leads to not only increased profit for the principal, but also an improved state of network security.

1. INTRODUCTION

The Existing works consider competitive insurance markets under compulsory insurance, and analyze the effect of insurance on agents' security expenditures. The authors of consider a competitive market with homogeneous agents, and show that insurance often deteriorates the state of network security as compared to the no-insurance scenario. The existing studies a network of heterogeneous agents and show that the introduction of insurance cannot improve the state of network security. Study the impact of the degree of agents' interdependence, and show that agents' investments decreases as the degree of interdependence increases. Study a competitive market under the assumption of voluntary participation by agents, with and without moral hazard. In the absence of moral hazard, the insurer can observe agents' investments in security, and hence premium discriminates based on the observed investments. They show that such a market can provide incentives for agents to increase their investments in self-protection. However, they show that under moral hazard, the market will not provide an incentive for improving agents' investments. The impact of insurance on the state of network security in the presence of a monopolistic welfare maximizing insurer has been studied in existing system. In these models, as the insurer's goal is to maximize social welfare, assuming compulsory insurance, agents are incentivized through premium discrimination, i.e., agents with higher investments in security pay lower premiums. As a result, these studies show that insurance can lead to improvement of network security. An insurance market with a monopolistic profit maximizing insurer, under the assumption of voluntary participation, has been studied in existing work, which shows that in the presence of moral hazard, insurance cannot improve network security as compared to the no-insurance scenario.

2. EXISTING SYSTEM

The Existing works consider competitive insurance markets under compulsory insurance, and analyze the effect of insurance on agents' security expenditures. The authors of consider a competitive market with homogeneous agents, and show that insurance often deteriorates the state of network security as compared to the no-insurance scenario. The existing studies a network of heterogeneous agents and show that the introduction of insurance cannot improve the state of network security. Study the impact of the degree of agents' interdependence, and show

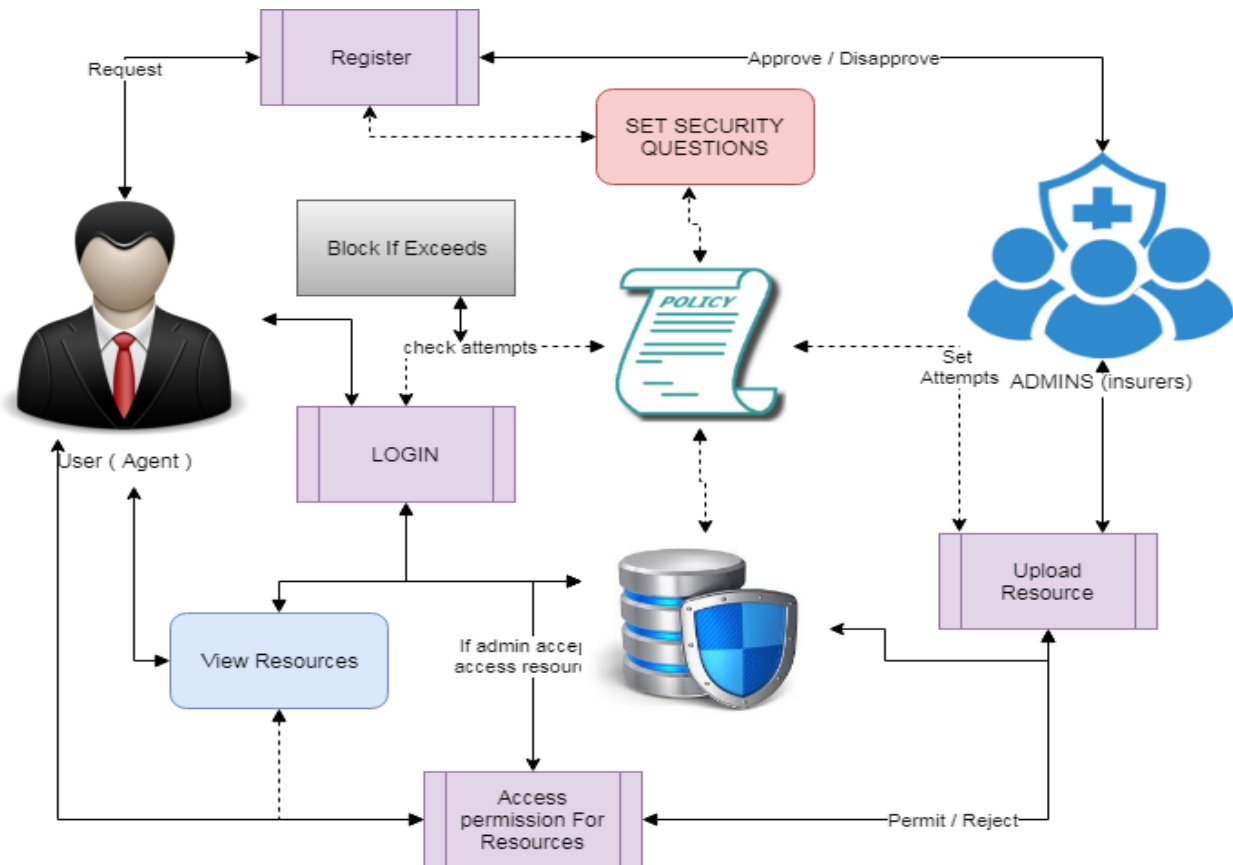
that agents' investments decreases as the degree of interdependence increases. Study a competitive market under the assumption of voluntary participation by agents, with and without moral hazard. In the absence of moral hazard, the insurer can observe agents' investments in security, and hence premium discriminates based on the observed investments. They show that such a market can provide incentives for agents to increase their investments in self-protection. However, they show that under moral hazard, the market will not provide an incentive for improving agents' investments. The impact of insurance on the state of network security in the presence of a monopolistic welfare maximizing insurer has been studied in existing system. In these models, as the insurer's goal is to maximize social welfare, assuming compulsory insurance, agents are incentivized through premium discrimination, i.e., agents with higher investments in security pay lower premiums. As a result, these studies show that insurance can lead to improvement of network security. An insurance market with a monopolistic profit maximizing insurer, under the assumption of voluntary participation, has been studied in existing work, which shows that in the presence of moral hazard, insurance cannot improve network security as compared to the no-insurance scenario.

3. PROPOSED SYSTEM

In this paper, we are interested in analyzing the possibility of using cyber-insurance as an incentive for improving network security. We adopt two model assumptions which we believe better capture the current state of cyber insurance markets but differ from the majority of the existing literature; we shall assume a profit maximizing cyber insurer, and voluntary participation, i.e., agents may opt out of purchasing a contract. Under this model, we focus on two features of cyber-insurance: (i) availability of risk assessment for mitigating moral hazard, and (ii) the interdependent nature of security. The first feature is due to the fact that recent advances in Internet measurements combined with machine learning techniques now allow us to perform accurate, quantitative security posture assessments at a firm level. This can be used as a tool to perform an initial security audit, or pre-screening, of a prospective client to mitigate moral hazard by premium discrimination and the design of customized policies. The second distinct feature, the interdependent nature of security, refers to the observation that the security standing of an entity often depends not only on its own effort towards implementing security metrics, but also on the efforts of other entities interacting with it within the eco-system. Such

interdependency is crucial for the insurer’s contract design problem, as the insurer will need to offer coverage to each insured for both its losses due to direct breaches, as well as indirect losses caused by breaches of other entities.

ARCHITECTURE:



4. CONCLUSION

We studied the problem of designing cyber insurance contracts by a single profit-maximizing insurer, for both risk-neutral and risk-averse agents. While the introduction of insurance worsens network security in a network of independent agents, we showed that the result could be different in a network of interdependent agents. Specifically, we showed that security interdependency leads to a profit opportunity for the insurer, created by the inefficient effort levels exerted by free-riding agents when insurance is not available but interdependency is present; this is in addition to risk transfer that an insurer typically profits from. We showed that security prescreening then allows the insurer to take advantage of this additional profit opportunity by

designing the right contracts to incentivize the agents to increase their effort levels and essentially selling commitment to interdependent agents. We show under what conditions this type of contracts leads to not only increased profit for the principal and utility for the agents, but also improved state of network security.

5. REFERENCES

- [1]. DipankarDasgupta. Immunity-based intrusion detection system: A general frame-work. In Proceedings of the 22nd National Information Systems Security Confer-ence (NISSC). Arlington, Virginia, USA, 1999.
- [2]. Jonatan Gomez and DipankarDasgupta. Evolving fuzzy classi_ers for intrusion detection. In Proceedings of the 2002 IEEE Workshop on Information Assurance, West Point, NY, USA, 2002.
- [3]. Steven A. Hofmeyr, Stephanie Forrest, and Anil Somayaji. Intrusion detection using sequences of system calls. *Journal of Computer Security*, 6(3):151{180, August 1998.
- [4]. Peter Mell Karen Scarfone. Guide to intrusion detection and prevention systems (idps). National Institute of Standards and Technology, NIST SP - 800-94, 2007.
- [5]. Jungwon Kim, Peter J. Bentley, Uwe Aickelin, Julie Greensmith, Gianni Tedesco, and Jamie Twycross. Immune system approaches to intrusion detection { a review. *Natural Computing*, 6(4):413{466, December 2007.
- [6] A. Shabtai, E. Menahem and Y. Elovici. F-Sign: automatic, function-based signature generation for malware, systems, man, and cybernetics, Part C: applications and reviews. *Transactions on IEEE*, 41, 494–508, 2011.
- [7] D. Kong, J. Gong, S. Zhu, P. Liu and H. Xi. SAS: semantics aware signature generation for polymorphic worm detection. *International Journal of Information Security*, 50, 1–19, 2011.
- [8] M. Sharma and D. Toshniwal. Pre-clustering algorithm for anomaly detection and clustering that uses variable size buckets. *RecentAdvances in Information Technology*, 515–519, 2012.

Agriculture land classification based on phenological information from dense satellite images

CHENDANSREE SAIDU
PG Scholar, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

DR.I.R.KRISHNAM RAJU*
Professor & Principal, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

Abstract—

Agricultural GHG Emissions Decarbonisation includes mapping and classification, which are among the most challenging tasks in the agricultural domain. Accurate prediction of agricultural land type in developing countries and cooperation with local communities can prevent famine, enhance food security, and contribute to when it comes to natural prediction; most agricultural land use forecasting relies on locally-sensed data, such as rainfall measurements and farmer surveys conducted by field visits. Locally-sensed data provide detailed information but are expensive to collect, noisy, and extremely hard to scale. However, remotely sensed and satellite imagery data, as a cheap and globally-accessible resource, can work in conjunction with modern Atmospheric scientist David Leininger's experiments examined the advantages of using satellite imagery to develop maps to inform farmers on their agricultural capabilities and develop optimal strategies for selecting the most productive land locations. Remote sensing and satellite imagery data have the potential to permit ground-based forecast models to be utilized despite El Niño's effects. Better information aids in identifying where great crop yields are located. Deep learning takes the massive search spaces into consideration, using computer techniques to find the most productive locations for farmers. This offers a way to bypass deficiencies in traditional radar information, which is limited.

*Index Terms:*Machine learning algorithms,Forestry,Credit cards,Control systems,Electronic commerce, Random forests.

I. Introduction

Agriculture has been at the very root of society since the beginning of human civilization. The farming revolution nearly 12,000 years ago, paved the way for modern-day colonization, allowing people to move towards a more stable, reliable and civilized lifestyle. Optimal survival conditions led to a surge in the human population and gave people the freedom to explore and develop innovations.

Agricultural Land Mapping describes how much of a global or regional area is surrounded by agricultural resources and human ventures, such as forests, agriculture, or other land types. With the recent advancement of satellites and remote sensing data providing constant access to such inputs, researchers have started focusing on land cover to better understand the features of Earth's surface.

Land cover information is presently utilized for various applications, some examples are forecasting weather, renewable energy, water control and supply, environment analysis, agriculture and its monitoring. Furthermore, it can be beneficial for disease control and disaster management. Likewise, the authorities of land management mainly use land cover to perceive and inspect the use of land, which is why this research can be useful for the society. Also, the human population is continuing to expand at unprecedented rates, which has brought the issues of food security and the impact of climate change to the forefront.

Given these real-world applications of agricultural land mapping, it is no wonder that a huge amount of research has been performed to produce accurate land cover datasets in many geographical locations and varying scales. The recent advancements in satellite imaging techniques have made it easier for data scientists to create more efficient and accurate prediction models. There has been a corresponding surge in the application of newer, more advanced and complex machine learning models to the domain.

Geospatial technologies like Remote Sensing acquire samples of electromagnetic radiation emanated and reflected by the earth's terrestrial, atmospheric and marine ecosystems. This sampling allows them to survey and spot physical attributes of a region without the need for any physical exposure. These techniques are actuated through the use of satellite-based sensors, which can be active or passive depending on their operational requirements.

Remote Sensing has made it possible for researchers to predict land type, crop yield and other meteorological and geological activities. These were previously

unpredictable as a result of inaccessibility, the risk to human life and feasibility related complications in the manual collection of data from such locations. Several machine learning models have been implemented for land use classification research but very few of them have used remote sensing data. Remote sensing is the future of the global climate analytics spectrum and that ideology has formed the crux of our paper.

This paper presents an ensemble framework consisting of multiple deep learning techniques to classify land utility and cover type. The model is realised on satellite image data acquired from the IKONOS Dataset and we calculate Precision Score, Recall and F-Score metrics and Support values to quantify the efficiency of the models.

II. Related Work

Land cover classification techniques for satellite imagery have been created and validated in many remote sensing researches. One such study proposes a unique, fully automatic and cheap land cover classification (ALCC) method. This approach does not need knowledge of the land or the assignment of training classes beforehand. The ALCC technique is founded on unsupervised grouping algorithms, that is carried out over the six Landsat-8 30m spatial resolution bands and spectral indices rasters. The main limitation of this model is the predetermined number of samples. Another paper introduces a research on improved use of polarization signatures for optimal land classification in mixed sample situations. A decision tree is made based on the class boundaries optimal to provide land cover classification. Although this method works relatively well for mixed class scenarios, the accuracy is only around 75% which is not impressive. A novel method called multimodal bilinear fusion network (MBFNet) was

introduced which merged the SAR and optical features for land use classification. In MBFNet, the fusion features extracted have strong discrimination to advancing classifying land-cover. However, this research went group land types according to the type of crops that can be planted. A research aimed on using artificial intelligence along with CNN to propose a new approach for land cover mapping. First a CNN model is trained with a broad range of images to get the land cover model. The model is then directly feeded satellite images which are split into pictures that are identical in size with the training ones. The results of the model are not satisfactory. This is because of the fact that it mixes up forested areas and water. Another framework was based on Spatial-Spectral Schroedinger Eigen maps (SSSE) for automatic land cover classification optimized using Cuckoo-Search (CS) method. Support-Vector Machine (SVM) was used for the final map generation after clustering and dimensionality reduction. The greatest drawback of this way of classification is that an increased classification accuracy is obtained at the cost of computational efficiency. A supervised technique for classification of land cover needs prior information of the terrain and training classes to classify the satellite imagery. Several people have studied and researched a supervised approach for this problem, like the Maximum-Likelihood Classifier (MLC). The limitation of this approach is the requirement for operator intervention, which slows down the processing chain. In remote sensing, a framework for quick and precise monitoring and classification of various land cover, various spectral indices have been used. Spectral indices were also applied to ascertain areas where certain crop lands are prevalent. The drawback is that this method saturates at high class content making it hard to differentiate relatively large plant cover

from very large plant cover. The data from the PALSAR-2 dataset was used to classify land in a project based on the use of a probability distribution function (PDF). The best PDF function is chosen by using separability index criteria and Chi-Squared test. This classification approach has good accuracy. However, these three PDFs (log-normal, weibull and normal) do not provide a good dissociated position result for splitting tall plant growth and modern samples. Most of the present land cover categorizing techniques only use single-modal remote sensing pictures, for instance, this one approach using optical images has faced the spectral confusion issue which lowers the accuracy of classification. This author [discovered that CNN performs classification with higher accuracy than Random Forest with Landsat image dataset. In this research CNN was implemented utilizing the Keras open-source library. It is necessary to find an optimal framework that works well with any given data since there are many ways to build the CNN architecture. However, classification using CNN for land cover faces a lot of limitations, such as the requirement for huge training image datasets. Multi-spectral Light Detection & Ranging, LiDAR, proposed by Suoyan Pan et. al. can yield point clouds advancing from several channels with variable wavelengths. Airborne LiDAR data produces relatively thorough and consistent spatial and spectral geometric data, which contributes to the classification of land cover and land use. This is classified into six different cover types – building, tree, water, grass, soil and road. The CNN model used the arguments are split into non-hyper-parameters and hyper-parameters; the training process aims to find the most suitable combination of model parameters. The performance of this CNN was better than the tradition CNN models eg. AlexNet and the deep Boltzmann machine. proposes a deep CNN architecture which has an input layer, 5

convolutional layers and 5 successive fully connected layers which automatically classifies outdoor laser mobile survey information. The architecture uses the spatial pyramid (SP) concept during voxelization of MLS samples to overcome the problem of several points within a voxel assigned to high point density. This model tested on 5 variable combinations of classes consisting of tree, non-tree and electric pole classes from the dataset. paper tests and evaluates a new approach in the classification of multispectral airborne lidar points. A 2-step method is employed to classify over 5 million points. These are the return points that are then classified using their three-channel intensities and height data. The SVM classifier performs exceptionally well, but the rule-based classification of multi-return points was not as successful because the recorded intensities are not reliable. has put forth a method to train transferable deep models, which allows the use of land-cover classification by using unlabeled multi-source remote sensing data and creating an amalgam of sorting land-use that concurrently extracts accurate class and edge data. A scale sequence joint deep learning method is proposed. This innovative joint deep learning (JDL) method involves MLP and Object-based CNN that replaces the previous paradigm of scale selection, by predicting land use using an object-based CNN and predicting land cover via Multilayer Perceptron (MLP). This clearly models the relationship between the predicted LU and LC variables as a joint distribution.

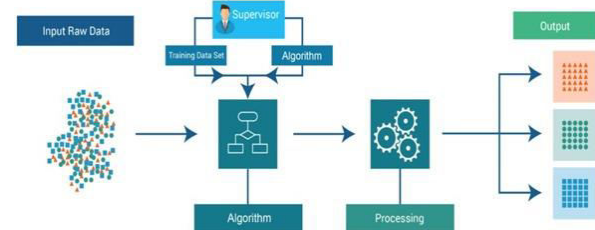
III. PROBLEM STATEMENT

A fog computing node can be any network device with storage, computing and networking capabilities (routers, switches, video surveillance cameras, servers, etc.) . Security and privacy issues predominate in fog computing infrastructures. The security and privacy issues can be mitigated by the mentioned counter technologies like proper authentication, access control, secure

channels, intrusion detection, trust management. While all these techniques are in place, fog computing can be considered a trusted device that users can rely on for processing, storing and managing data. This paper introduces fog computing as a trusted device. Based on fog computing, the present research proposes a secure cloud storage scheme.

Architecture

Supervised Learning



IV. EXISTINGSYSTEM

Land cover classification techniques for satellite imagery have been created and validated in many remote sensing researches. One such study [1] proposes a unique, fully automatic and cheap land cover classification (ALCC) method. This approach does not need knowledge of the land or the assignment of training classes beforehand. The ALCC technique is founded on unsupervised grouping algorithms, that is carried out over the six Landsat-8 30m spatial resolution bands and spectral indices rasters.

DISADVANTAGES:

- The main limitation of this model is the predetermined number of samples.

V. PROPOSED SYSTEM

The agricultural problem we aim to address is the classification and mapping of the agricultural land for a specific region of interest related to a series of satellite and remotely sensed pictures taken before the

crops harvest. In particular, we attempt to predict the agricultural land based on the unit area in a given geological location. This author [10] discovered that CNN performs classification with higher accuracy than Random Forest with Landsat image dataset. In this research CNN was implemented utilizing the Keras open-source library. It is necessary to find an optimal framework that works well with any given data since there are many ways to build the CNN architecture.

ADVANTAGES

- Classification using CNN for land cover faces a lot of limitations, such as the requirement for huge training image datasets.
- The CNN model used the arguments are split into non-hyper-parameters and hyper-parameters; the training process aims to find the most suitable combination of model parameters. The performance of this CNN was better than the tradition ALCC models

VI. MODULES

To implement this project we have designed 3 modules and each module contains two algorithms with their accuracy comparison graph.

1. Upload Dataset
2. Preprocessing
3. CNN Model

Upload Dataset

Upload dataset and run all other modules. To implement this project we have used LAND

satellite images which contain images of FOREST, AGRICULTURE LAND, URBAN AREA and Range LAND.

Preprocessing

Owing to the inadequacy of appropriate data for training, the unmediated application of deep learning models is an unviable task. The use of multi-spectral images also rules out employing conventional computer vision techniques for pre-training. We employ the image pillow library to prepare raw images before they are feeded into the main model. Pillow is a fork of the PIL, short for Python Imaging Library. The Pillow is a library that offers many standard techniques for manipulating pictures.

CNN Model: First, we start by training the foundational CNN model. The 3 main layers (Convolutional, Pooling and Fully-Connected) are created and we integrate the inception module as a dimensionality reduction technique and to allow the CNN model to propagate backwards. The results generated and the weightage values are stored in a separate file. Then we define the ResNet 50, ResNet 50V2 and ResNet152V2 neural networks to combine multiple perspectives across levels. The three models and the results generated by them are also stored in the same file. These deep learning models are trained multiple times to enhance their accuracy and we store the results from each iteration so that we can visualize the progressive improvement in accuracy. We plot the accuracy history for the resnet models. We also calculate Precision Score, Recall and F-Score metrics and Support values to quantify the efficiency of the models.

VII. PROCESS MODEL

The agricultural problem we aim to address is how to use a satellite and remotely-sensed

image to classify and map agricultural land based on different geological characteristics. While we need help mapping this land to make sure that it is associated with the right locations, we also need to understand what this agricultural land looks like to where it is being used to serve some or all of these needs. However, most satellite data, and, therefore, most data about agricultural land are unlabeled. Over the last 30 years, collection on remote sensing had largely existed, but these images have not yet been mapped in their relevant geological locations. The problem is that we have often 5 years left before the harvest.

We employ a dimensionality reduction technique. In particular, global average pooling (GAP), we make use of deep learning architectures namely Inception CNN, ResNet and VGG to achieve significant

The goal is to draw upon a framework through which to identify and project the succession of plant growth stages of a crop into the underlying structure. Factors identified with crop growth will naturally fall within these structure pictures.

When labelled data is lacking, it is necessary to use training sets associated with nearest neighbours. The deep learning models will make it easier to optimize and enhance accuracy.

PREPROCESSING

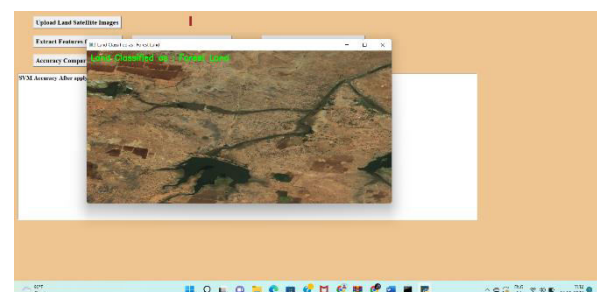
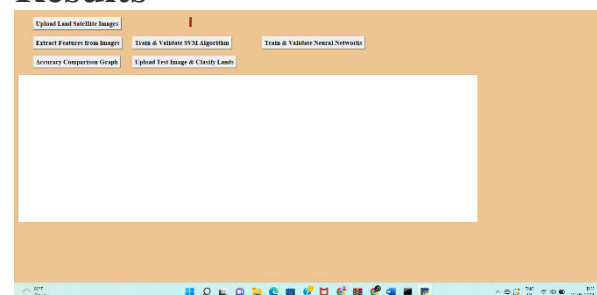
Computer vision often struggles with raw data. Since an abundance of data is usually the problem, we print the digital images on the color filter paper in 5-muh Kelvin and take them with the built-in camera. We use the software Pillow to prepare the digital photos before they are fed to Deep Learning models. Pillow is different from Picture

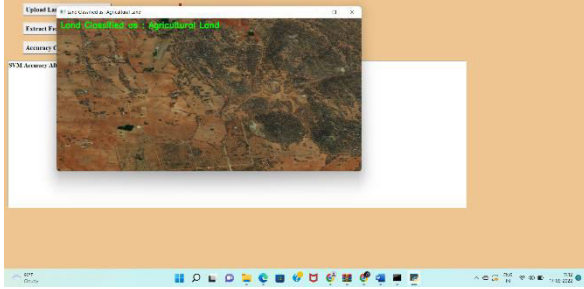
Image Library. It was developed as a pure Python (PIL) library that offers many common operations. The Pillow (a fork of the PIL) allows us to represent digitalized images.

PROCESSING RAW IMAGES

A computerized vegetation mask is applied to raw satellite images and pixel values with informative values are clipped in accordance with requirements for the algorithm. Permutation invariance allows the inference that only distinct pixel types from the mask map to a reduced form of image, eliminating the possibility of information/function loss when high-dimensional images are quantized into pixel amount data matrices. Icons or networks that visualize these measures several pixels/nodes to determine autostrate duplicates are laid in the noise/inappropriate data structure with analytical algorithmic rules.

Results





VII. CONCLUSION

The aim of this research is to develop a machine-learning model for the classification of land utility and patch type namely cover type within satellite images. The proposed system uses deep learning techniques namely convolutional neural network (CNN), residual network (ResNet) and VGG and is implemented in the Inception module that is actuated satellite images of land of area obtained from the IKONOS dataset which are categorized into eighteen categories that represent six classes of croplands. Evaluation has proven the proposed system superior to the previous approaches.

VIII. REFERENCES

1. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ (2020) A new coronavirus associated with human respiratory disease in china. *Nature* 44(59):265–269
2. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) <https://www.medscape.com/viewarticle/924596>
3. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, Qiu Y, Wang J, Liu Y, Wei Y, Xia J, Yu T, Zhang X, Zhang L (2020) Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 395(10223):507–513

4. World health organization: <https://www.who.int/new-room/g-adetail/q-a-coronaviruses#:text=symptoms>. Accessed 10 Apr 2020
5. Wikipedia coronavirus Pandemic data: https://en.m.wikipedia.org/wiki/Template:2019%E2%80%9320_coronavirus_pandemic_data. Accessed 10 Apr 2020
6. Khanday, A.M.U.D., Amin, A., Manzoor, I., & Bashir, R., “Face Recognition Techniques: A Critical Review” 2018
7. Kumar A, Dabas V, Hooda P (2018) Text classification algorithms for mining unstructured data: a SWOT analysis. *Int J Inf Technol.* <https://doi.org/10.1007/s41870-017-0072-1>
8. Verma P, Khanday AMUD, Rabani ST, Mir MH, Jamwal S (2019) Twitter Sentiment Analysis on Indian Government Project using R. *Int J Recent Tech Eng.* <https://doi.org/10.35940/ijrte.C6612.098319>
9. Chakraborti S, Choudhary A, Singh A et al (2018) A machine learning based method to detect epilepsy. *Int J Inf Technol* 10:257–263. <https://doi.org/10.1007/s41870-018-0088-1>
10. Sarwar A, Ali M, Manhas J et al (2018) Diagnosis of diabetes type-II using hybrid machine learning based ensemble model. *Int J Inf Technol.* <https://doi.org/10.1007/s41870-018-0270-5>
11. Bullock J, Luccioni A, Pham KH, Lam CSN, Luengo-Oroz M (2020) Mapping the landscape of artificial intelligence applications against COVID-19. <https://arxiv.org/abs/2003.11336v1>
12. Wang L, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 Cases from chest radiography images. <https://arxiv.org/abs/2003.09871>
13. Yan L, Zhang H-T, Xiao Y, Wang M, Sun C, Liang J, Li S, Zhang M, Guo Y, Xiao Y, Tang X, Cao H, Tan X, Huang N, Amd A, Luo BJ, Cao Z, Xu H, Yuan Y (2020)

Prediction of criticality in patients with severe covid-19 Infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. medRxiv.

<https://doi.org/10.1101/2020.02.27.20028027>

14. Jiang X, Coffee M, Bari A, Wang J, Jiang X, Huang J, Shi J, Dai J, Cai J, Zhang T, Wu Z, He G, Huang Y (2020) Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Compu Mater Contin* 63(1):537–551

15. Description of Logistic Regression Algorithm. <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>. Accessed 15 May 2019

16. Description of Multinomial Naïve Bayes Algorithm <https://www.3pillarglobal.com/insights/document-classification-using-multinomial-naive-bayes-classifier>. Accessed 15 May 2019

17. Khanday AMUD, Khan QR, Rabani ST. SVM-BPI: support vector machine based propaganda identification. *SN Appl. Sci.* (accepted)

18. Description of Decision Tree Algorithm: https://dataspirant.com/2017/01/30/how_decision_tree_algorithm_works/. Accessed 10 July 2019

19. Description of Boosting Algorithm: <https://towardsdatascience.com/boosting>. Accessed 10 July 2019

20. Description of Adaboost Algorithm: <https://towardsdatascience.com/boosting-algorithm-adaboost-b673719ee60c>. Accessed 10 July 2019



CLASSIFYING FAKE NEWS ARTICLES USING NATURAL LANGUAGE PROCESSING TO IDENTIFY IN-ARTICLE ATTRIBUTION AS A SUPERVISED LEARNING ESTIMATOR

Sannidhiraju N V Satya Lakshmi Sujitha (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Intentionally deceptive content presented under the guise of legitimate journalism is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most so-called ‘fake news’ is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as making excessive use of unsubstantiated hyperbole and non-attributed quoted content. In this paper, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Textblob, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in a Bayesian machine learning system as a key feature to estimate the likelihood that a news article is fake. The resultant process precision is 63.333% effective at assessing the likelihood that an article with quotes is fake. This process is called influence mining and this novel technique is presented as a method that can be used to enable fake news and even propaganda detection. In this paper, the research process, technical analysis, technical linguistics work, and classifier performance and results are presented. The paper concludes with a discussion of how the current system will evolve into an influence mining system.

1. INTRODUCTION

Intentionally deceptive content presented under the guise of legitimate journalism (or ‘fake news,’ as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

2. LITERATURE SURVEY

1) When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism

AUTHORS: M. Balmas

This research assesses possible associations between viewing fake news (i.e., political satire) and attitudes of inefficacy, alienation, and cynicism toward political candidates. Using survey



data collected during the 2006 Israeli election campaign, the study provides evidence for an indirect positive effect of fake news viewing in fostering the feelings of inefficacy, alienation, and cynicism, through the mediator variable of perceived realism of fake news. Within this process, hard news viewing serves as a moderator of the association between viewing fake news and their perceived realism. It was also demonstrated that perceived realism of fake news is stronger among individuals with high exposure to fake news and low exposure to hard news than among those with high exposure to both fake and hard news. Overall, this study contributes to the scientific knowledge regarding the influence of the interaction between various types of media use on political effects.

2) Miley, CNN and The Onion

AUTHORS: D. Berkowitz and D. A. Schwartz

Following a twerk-heavy performance by Miley Cyrus on the Video Music Awards program, CNN featured the story on the top of its website. The Onion—a fake-news organization—then ran a satirical column purporting to be by CNN's Web editor explaining this decision. Through textual analysis, this paper demonstrates how a Fifth Estate comprised of bloggers, columnists and fake-news organizations worked to relocate mainstream journalism back to within its professional boundaries.

3) The Impact of Real News about “Fake News” ’ : Intertextual Processes and Political Satire

AUTHORS: P. R. Brewer, D. G. Young, and M. Morreale

This study builds on research about political humor, press metacoverage, and intertextuality to examine the effects of news coverage about political satire on audience members. The analysis uses experimental data to test whether news coverage of Stephen Colbert's Super PAC influenced knowledge and opinion regarding *Citizens United*, as well as political trust and internal political efficacy. It also tests whether such effects depended on previous exposure to *The Colbert Report* (Colbert's satirical television show) and traditional news. Results indicate that exposure to news coverage of satire can influence knowledge, opinion, and political trust. Additionally, regular satire viewers may experience stronger effects on opinion, as well as increased internal efficacy, when consuming news coverage about issues previously highlighted in satire programming.

4) Stopping Fake News

AUTHORS: M. Haigh, T. Haigh, and N. I. Kozak

Social media is acting as a double-edged sword for universe in a way of consuming news. On one side, its ease of access, popularity and low cost distribution channel lead people to gain news from social media. On other side, it is also acting as a source of spread of 'fake news'. The extensive spread of fake news on social media, websites are impacting society negatively. This makes extremely important to combat the spread of fake news and to aware the society. In this paper, we offer a review which lists out the sources of fake news, its types, generation, motivation and examples. Also, some approaches are suggested to spot and stop fake news spread.



5) With Facebook, Blogs, and Fake News, Teens Reject Journalistic "Objectivity"

AUTHORS: R. Marchi

This article examines the news behaviors and attitudes of teenagers, an understudied demographic in the research on youth and news media. Based on interviews with 61 racially diverse high school students, it discusses how adolescents become informed about current events and why they prefer certain news formats to others. The results reveal changing ways news information is being accessed, new attitudes about what it means to be informed, and a youth preference for opinionated rather than objective news. This does not indicate that young people disregard the basic ideals of professional journalism but, rather, that they desire more authentic renderings of them.

3. IMPLEMENTATION

MODULES:

- ❖ Social Media Mining System Construction
- ❖ User Topical Package Model Mining
- ❖ Route Package Mining
- ❖ Travel sequence recommendation

MODULES DESCRIPTION:

Social Media Mining System Construction

- ❖ In the first module we develop the system for the evaluation of our proposed model and thus make the system construction module with social media mining system.
- ❖ Our topic package space is the extension of textual descriptions of topics such as ODP. We use the topical package space to measure the similarity of the user topical model package (user package) and the route topical model package (route package). In our paper, we construct the topical package space by the combination of two social media: travelogues and community-contribute photos. To construct topical package space, travelogues are used to mine representative tags, distribution of cost and visiting time of each topic, while community-contributed photos are used to mine distribution of visiting time of each topic.
- ❖ The reasons for using the combination of social media are (1) travelogues are more comprehensive to describe a location than the tags with the photos which are with so many noises; (2) it is difficult to mine a user's consumption capability and the cost of POIs directly by the photos or the tags with the photos; (3) to season, although both media could offer correct visiting season information of POIs, the number of photos of a POI is far larger than the number of travelogues. (4) the time difference between where the user lives and the "data taken" of community contributed photos of where he or she visits make the taken time inaccurate.

User Topical Package Model Mining

- ❖ User topical package model (user package) is learnt from mapping the tags of user's photos to topical package space. It contains user topical interest distribution



(U), user consumption capability (U), preferred travel time distribution (U) and preferred travel season distribution .

- ❖ In this module, we introduce how to extract the user package, which contains user topical interest distribution, user consumption capability distribution, preferred travel time distribution and preferred travel season distribution.
- ❖ First we introduce user's topical interest mining from mapping user's tags to the topical package space. Then, we introduce how to get topical space mapping method.
- ❖ We map the textual description (tags) of user's community photos to the topical package space to present the user's travel preference of different topics, which is defined as user topical interest distribution. We assume that if a user's tags appear frequently in one topic and less in others, the user has a higher interest towards this topic.
- ❖ We use the cost distributions of the all the topics and distribution of use's topical interest to present a user's consumption capability. If a user usually takes part in luxurious activities like Golf and Spas, his consumption capability is very likely to be. If a user usually takes part in some cheap things, his consumption capability is likely to be low, and we tend not to recommend him luxurious topics.

Route Package Mining

- ❖ Route topical package model (route package) is learnt from mapping the travelogues related to the POIs on the route to topical package space. It contains route topical interest, route's cost distribution, route's time distribution and season distribution.
- ❖ To save the online computing time, we mine travel routes and the attribute of the routes offline. After mining POIs, to construct travel routes, we analyze the spatio-temporal structure of the POIs among travelers' records.
- ❖ We construct the spatio-temporal structure of the POIs according to the "data taken". POI with the earlier timestamp is defined as the "in". POI with a later timestamp, on the contrary, is defined as "out". Then we count the times of "in" and "out" from POI to others by the records of all the users after filtering. A greedy algorithm is then applied to find the time sequence of these POIs. Thus, we finish famous routes mining and obtain famous routes of each city.

Travel sequence recommendation

- ❖ After mining user package and route package, in this module, we develop our travel routes recommendation module. It contains two main steps: (1) routes ranking according to the similarity between user package and routes packages, and (2) route optimizing according to similar social users' records.
- ❖ After POI and route ranking module, we get a set of ranked routes. Here, we further describe the optimization of top ranked routes according to social similar users' travel records. Firstly, we introduce how to mine social similar users and their travel records. Then we introduce how to optimize the roads by social users' travel records.

3 techniques will be used to calculate score

- 1) Source: any person who is writing news will give his name or a person name on which he writing articles



- 2) CUE: using this we will extract VERBS or VERBS phrases, if news is real then it will have verb types of words
- 3) Quotes: all articles will be on some topics and person will describe that topic name under quotes. So we will look for quotes in articles to determine fake or real news.

Document examples

"When "Mitt Romney" was governor of Massachusetts, we didnt just slow the rate of growth of our government, we actually cut it."

In above sentence quotes are there and it's talking about 'Mitt Romney' and it's contains some verbs such as 'was, didn't, slow, cut'. By analysing above 3 features from articles we can come to the conclusion whether news is FAKE or REAL.

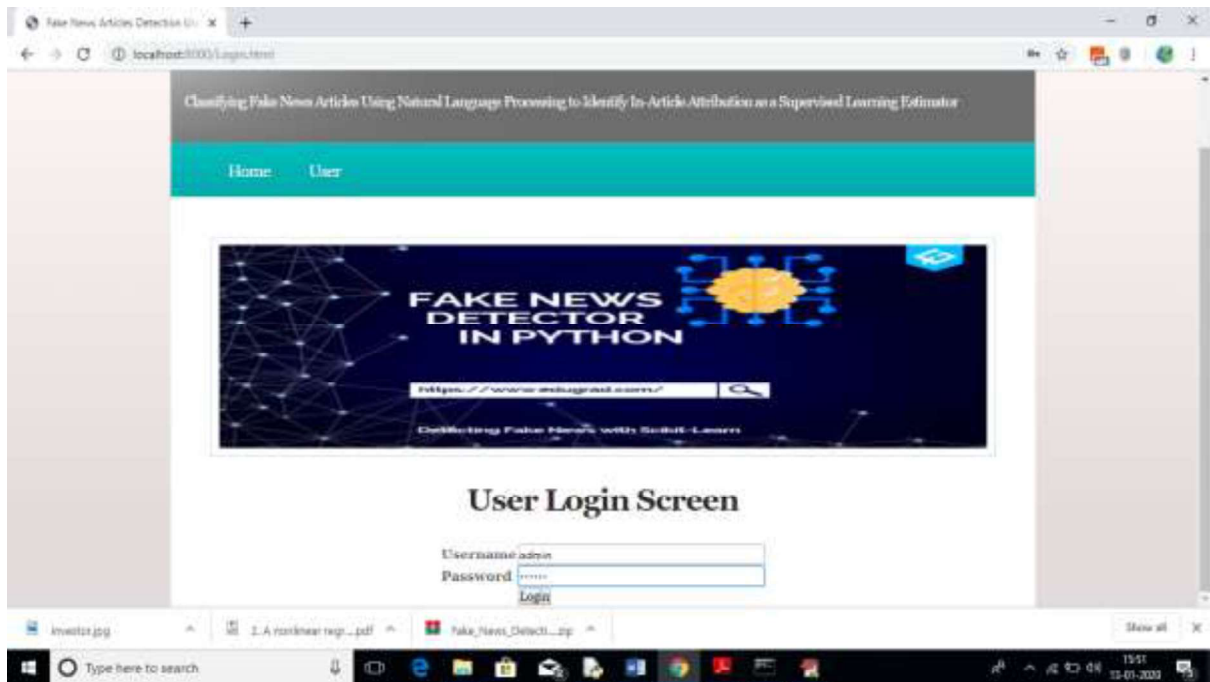
All FAKE peoples will not write such statements in their articles so we can detect by applying this techniques.

To implement this project we are using 'News' dataset and then by applying above technique we can detect whether this news are fake or real. This dataset I kept inside dataset folder. Upload this dataset when you are running application.

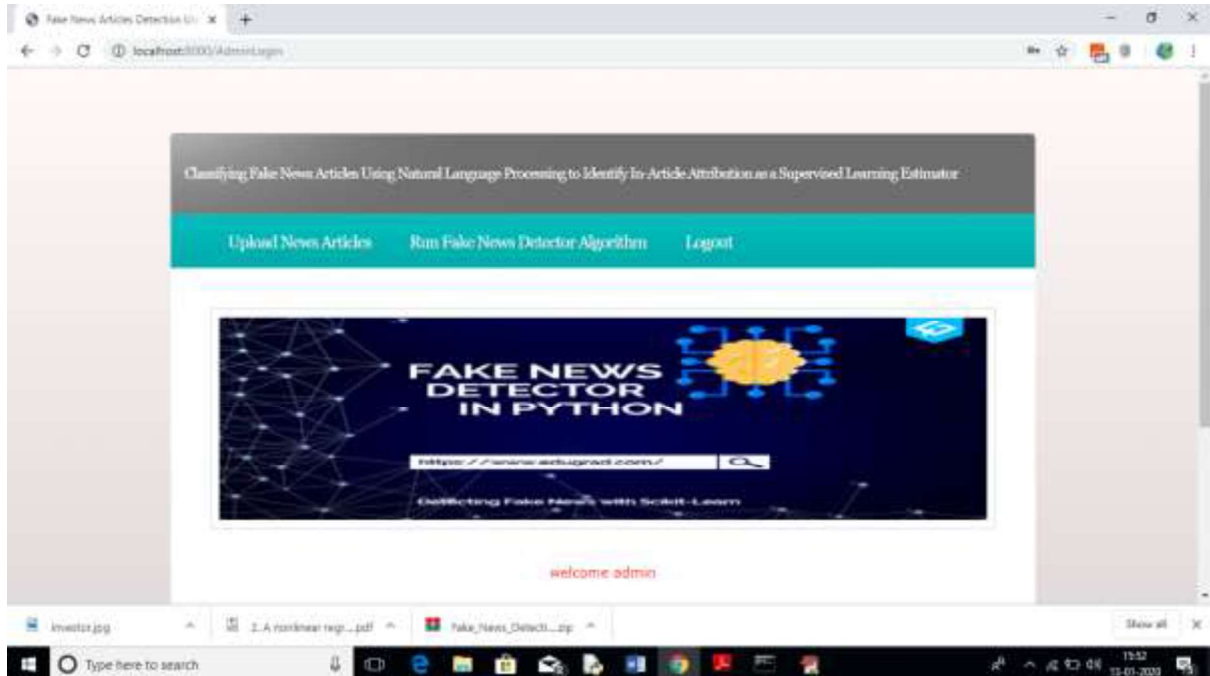
To run this project deploy 'FakeNews' folder on 'django' python web server and then start server and run in any web browser. After running code in web browser will get below page. Screen shots



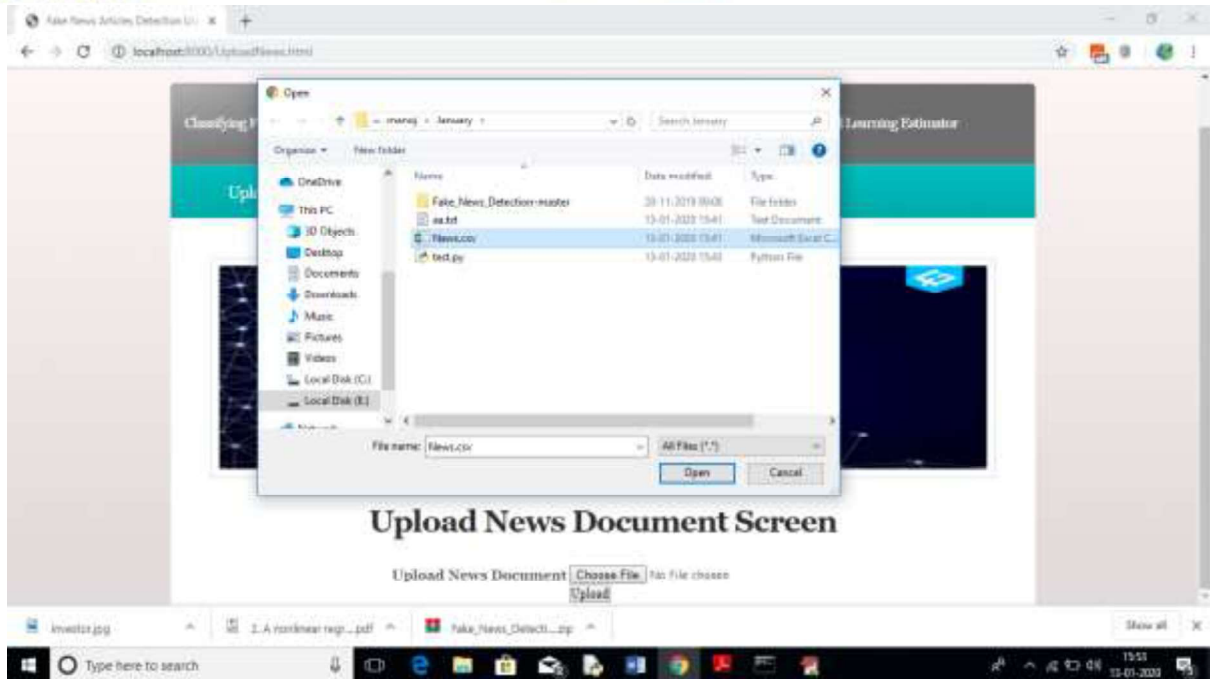
In above screen click on 'User' link to get below screen



In above screen enter username and password as 'admin' and then click on 'Login' button to get below screen



In above screen click on 'Upload News Articles' link to upload news document



Upload News Document Screen

In above screen I am uploading 'News.csv' file which contains 150 news paragraphs. After uploading news will get below screen



Upload News Document Screen

In above screen news file uploaded successfully, now click on 'Run Fake News Detector Algorithm' link to calculate Fake News Detection algorithm score and based on score and naïve bayes algorithm we will get result.



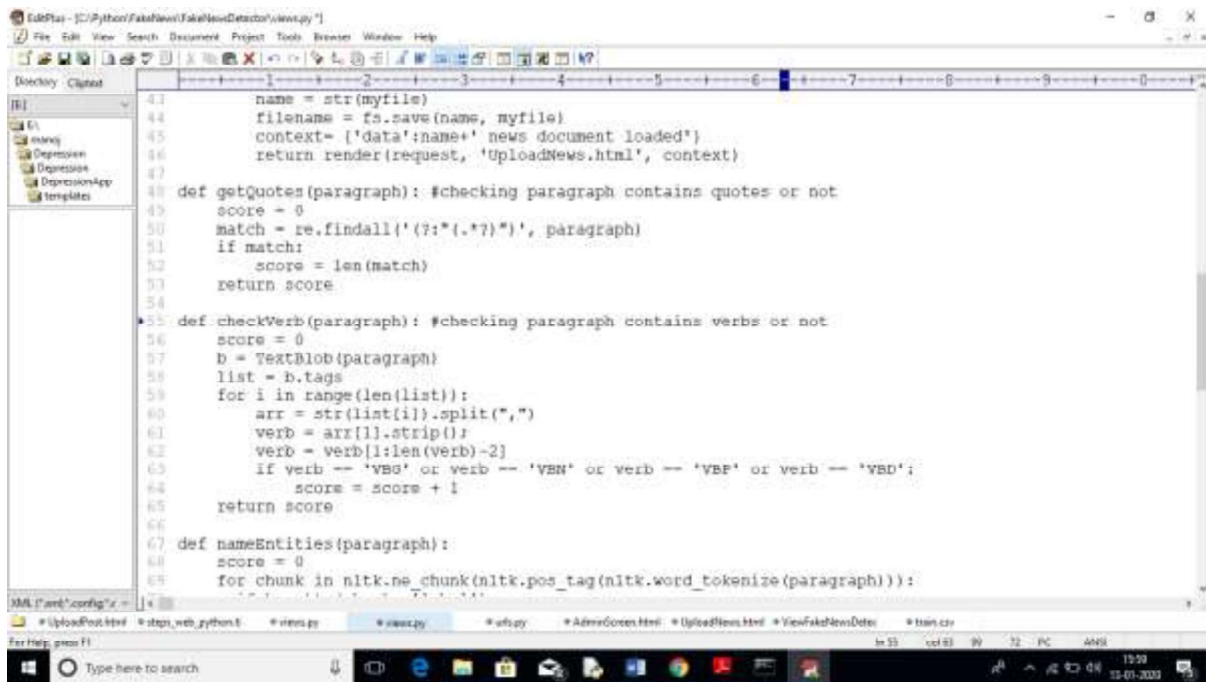
News Text	Detection Result	Fake Rank Score
Says the Anties Life political group supports third-trimester abortions on demand.	Fake News	0.8333333333333333
When did the decline of coal start? It started when natural gas took off that started to begin in (President George W.) Bush's administration.	Real News	2.142857142857143
"Hillary Clinton agrees with John McCain" by voting to give George Bush the benefit of the doubt on Iran.	Real News	3.076923076923077
Health care reform legislation is likely to mandate fine and change surgeries.	Fake News	0.7692307692307693
The economic turnaround started at the end of my term.	Real News	0.9090909090909091
The Chicago Bears have had more starting quarterbacks in the last 10 years than the total number of tenured (UW) faculty fired during the last two decades.	Real News	1.3333333333333333
Jim Dornan has not lived in the district he represents for years now.	Real News	2.142857142857143
I'm the only person on this stage who has worked actively just last year passing, along with Russ Feingold, some of the toughest ethics reform since Watergate."	Real News	1.5151515151515151
"However, it took \$49.5 million in Oregon Lottery funds for the Port of Newport to eventually land the new NOAA Marine Operations Center-Pacific."	Real News	2.142857142857143
Says GOP primary opponents Glenn Grothman and Joe Leibman cut a compromise vote that cost \$788 million in higher electricity costs.	Real News	2.1730120434782608
For the first time in history, the share of the national popular vote margin is smaller than the Latino vote margin."	Fake News	0.8
"Since 2000, nearly 12 million Americans have slipped out of the middle class and into poverty."	Real News	1.5
"When Mitt Romney was governor of Massachusetts, we didn't just slow the rate of growth of our government, we actually cut it."	Real News	2.2222222222222223
The economy bled \$24 billion due to the government shutdown.	Fake News	0.8333333333333333
Most of the (Affordable Care Act) has already in some sense been waived or otherwise suspended.	Real News	2.1052631578947367
"In this last election in November, ... 63 percent of the American people chose not to vote. ... So percent of young people, (and) 75 percent of low-income workers chose not to vote."	Real News	0.973609736097361

In above screen first column contains news text and second column is the result value as 'fake or real' and third column contains score. If score greater > 0.90 then I am considering news as REAL otherwise fake.

Some neighborhood schools are closing.	Real News	0.3333333333333333
He told gay organizers in Massachusetts he would be a stronger advocate for special rights than even Ted Kennedy.	Real News	1.5
"The years that I was speaker, the Florida House consistently offered leaner budgets than the governor offered."	Real News	2.380952380952381
"We are already almost halfway to our 2010 goal of creating 700,000 new jobs in seven years."	Real News	1.5
Says the U.S. Supreme Court found that Social Security is not guaranteed.	Real News	3.8461538461538463
Says Michael Bennett wants to close Guantanamo Bay prison and bring terrorists right here to Colorado.	Real News	2.0000000000000005
Oregonians have an amazing no-cost way to fight abortion with free political donations	Fake News	0.7692307692307693
The president said he's going to bring in 250,000 (Syrian and Iraqi) refugees into this country.	Real News	2.380952380952381
Research shows that a vast majority of arriving immigrants today come here because they believe that government is the source of prosperity, and that's what they support."	Real News	1.6329032258064315
Newt Gingrich's immigration plan offers a new doorway to amnesty.	Real News	1.8181818181818183
Mr. Caprio is a career politician who has never worked in the private sector.	Real News	0.0
"In Rhode Island, 9 percent of workers use the states temporary disability insurance program each year while in New Jersey, the rate is only 3 percent."	Real News	1.3903225806431613
"In just 17 years, spending for Social Security, federal health care and interest on the debt will exceed ALL tax revenue!"	Fake News	0.7692307692307693
President Obama took more money from Wall Street in the 2008 campaign than anybody ever had.	Real News	2.3529411764705883
Donald Trump has said nuclear proliferation is OK.	Real News	0.3333333333333333
Hillary Clinton has taken over \$800,000 from lobbyists."	Real News	2.5
Barack Obama has never even worked in business.	Real News	1.3333333333333333
Save the Arizona immigration law expressly bans racial profiling	Real News	1.0
Says Gov. Rick Perry has been begging for the federal government to send the Coast Guard to patrol two lakes on the U.S.-Mexico border.	Real News	1.0230769230769231
"On the VA: Over 300,000 veterans have died waiting for care."	Real News	2.0000000000000005

For all 150 news text articles we got result as fake or real.

See below screen shots of code calculating quotes, name entity and verbs from news paragraphs



```
41 name = str(myfile)
42 filename = fs.save(name, myfile)
43 context= ['data':name+' news document loaded']
44 return render(request, 'UploadNews.html', context)
45
46 def getQuotes(paragraph): #checking paragraph contains quotes or not
47 score = 0
48 match = re.findall('(?:\".*?\")', paragraph)
49 if match:
50 score = len(match)
51 return score
52
53 def checkVerb(paragraph): #checking paragraph contains verbs or not
54 score = 0
55 b = TextBlob(paragraph)
56 list = b.tags
57 for i in range(len(list)):
58 arr = str(list[i]).split(",")
59 verb = arr[1].strip()
60 verb = verb[1:len(verb)-2]
61 if verb == 'VBS' or verb == 'VBN' or verb == 'VBP' or verb == 'VBD':
62 score = score + 1
63 return score
64
65 def nameEntities(paragraph):
66 score = 0
67 for chunk in nltk.ne_chunk(nltk.pos_tag(nltk.word_tokenize(paragraph))):
```

4. CONCLUSION

This paper presented the results of a study that produced a limited fake news detection system. The work presented herein is novel in this topic domain in that it demonstrates the results of a full-spectrum research project that started with qualitative observations and resulted in a working quantitative model. The work presented in this paper is also promising, because it demonstrates a relatively effective level of machine learning classification for large fake news documents with only one extraction feature. Finally, additional research and work to identify and build additional fake news classification grammars is ongoing and should yield a more refined classification scheme for both fake news and direct quotes.

5. REFERENCES

- [1] H. Liu, T. Mei, J. Luo, H. Li, and S. Li, "Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing," in Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 9–18.
- [2] J. Li, X. Qian, Y. Y. Tang, L. Yang, and T. Mei, "Gps estimation for places of interest from social users' uploaded photos," IEEE Transactions on Multimedia, vol. 15, no. 8, pp. 2058–2071, 2013.
- [3] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model based collaborative filtering for personalized poi recommendation," IEEE Transactions on Multimedia, vol. 17, no. 6, pp. 907–918, 2015.
- [4] J. Sang, T. Mei, and C. Sun, J.T.and Xu, "Probabilistic sequential pois recommendation via check-in data," in Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, 2012.



- [5] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web*, vol. 5, no. 1, p. 5, 2011.
- [6] H. Gao, J. Tang, X. Hu, and H. Liu, "Content-aware point of interest recommendation on location-based social networks," in *Proceedings of 29th International Conference on AAAI*. AAAI, 2015.
- [7] Q. Yuan, G. Cong, and A. Sun, "Graph-based point-of-interest recommendation with geographical and temporal influences," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*. ACM, 2014, pp. 659–668.
- [8] H. Yin, C. Wang, N. Yu, and L. Zhang, "Trip mining and recommendation from geo-tagged photos," in *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2012, pp. 540–545.
- [9] Y. Gao, J. Tang, R. Hong, Q. Dai, T. Chua, and R. Jain, "W2go: a travel guidance system by automatic landmark ranking," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 123–132.
- [10] X. Qian, Y. Zhao, and J. Han, "Image location estimation by salient region matching," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4348–4358, 2015.
- [11] H. Kori, S. Hattori, T. Tezuka, and K. Tanaka, "Automatic generation of multimedia tour guide from local blogs," *Advances in Multimedia Modeling*, pp. 690–699, 2006.

A ROAD ACCIDENT PREDICTION MODEL USING DATA MINING TECHNIQUES

Sapparapu Naga Kishore (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr.I.R.Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Due to the exponentially increasing number of vehicles on the road, the number of accidents occurring on a daily basis is also increasing at an alarming rate. With the high number of traffic incidents and deaths these days, the ability to forecast the number of traffic accidents over a given time is important for the transportation department to make scientific decisions. In this scenario, it will be good to analyze the occurrence of accidents so that this can be further used to help us in coming up with techniques to reduce them. Even though uncertainty is a characteristic trait of majority of the accidents, over a period of time, there is a level of regularity that is perceived on observing the accidents occurring in a particular area. This regularity can be made use of in making well informed predictions on accident occurrences in an area and developing accident prediction models. In this paper, we have studied the inter relationships between road accidents, condition of a road and the role of environmental factors in the occurrence of an accident. We have made use of data mining techniques in developing an accident prediction model using Apriori algorithm and Support Vector Machines. Bangalore road accident datasets for the years 2014 to 2017 available in the internet have been made use for this study. The results from this study can be advantageously used by several stakeholders including and not limited to the government public work departments, contractors and other automobile industries in better designing roads and vehicles based on the estimates obtained.

1. INTRODUCTION

Intentionally deceptive content presented under the guise of legitimate journalism (or ‘fake news,’ as it is commonly known) is a worldwide information accuracy and integrity problem that affects opinion forming, decision making, and voting patterns. Most fake news is initially distributed over social media conduits like Facebook and Twitter and later finds its way onto

mainstream media platforms such as traditional television and radio news. The fake news stories that are initially seeded over social media platforms share key linguistic characteristics such as excessive use of unsubstantiated hyperbole and non-attributed quoted content. The results of a fake news identification study that documents the performance of a fake news classifier are presented and discussed in this paper.

2. EXISTING SYSTEM

Williams et al. [5] have found through their studies that the age and experience of a driver also play a major role in the occurrence of accidents. Suganya, E. and S. Vijayarani [6] in their paper have analysed the road accidents in India and compared the performance of different classification algorithms such as linear regression, logistic regression, decision tree, SVM, Naïve Bayes, KNN, Random Forest and gradient boosting algorithm using accuracy, error rate and execution time as a measure of performance. They have found the performance of KNN to be better than that of the others.

Sarkar et al. [7] have done a comparative study on the type of roads that are prominent in accidents. While exploring the other components associated with accidents, they have found that the occurrence of accidents in highways is more common than in a normal road similar to [4]. Stewart et al. [8] have utilized original data in building a neural network model to predict accidents. They found that this model was able to give quicker results than those being used in the models built on Indian roads.

Zheng et al. [9] have studied the range of injuries that come forth in a motor vehicle accident and have also analyzed the emotions of the drivers involved in the accidents that could have been a causal factor. Arun Prasath N and Muthusamy.

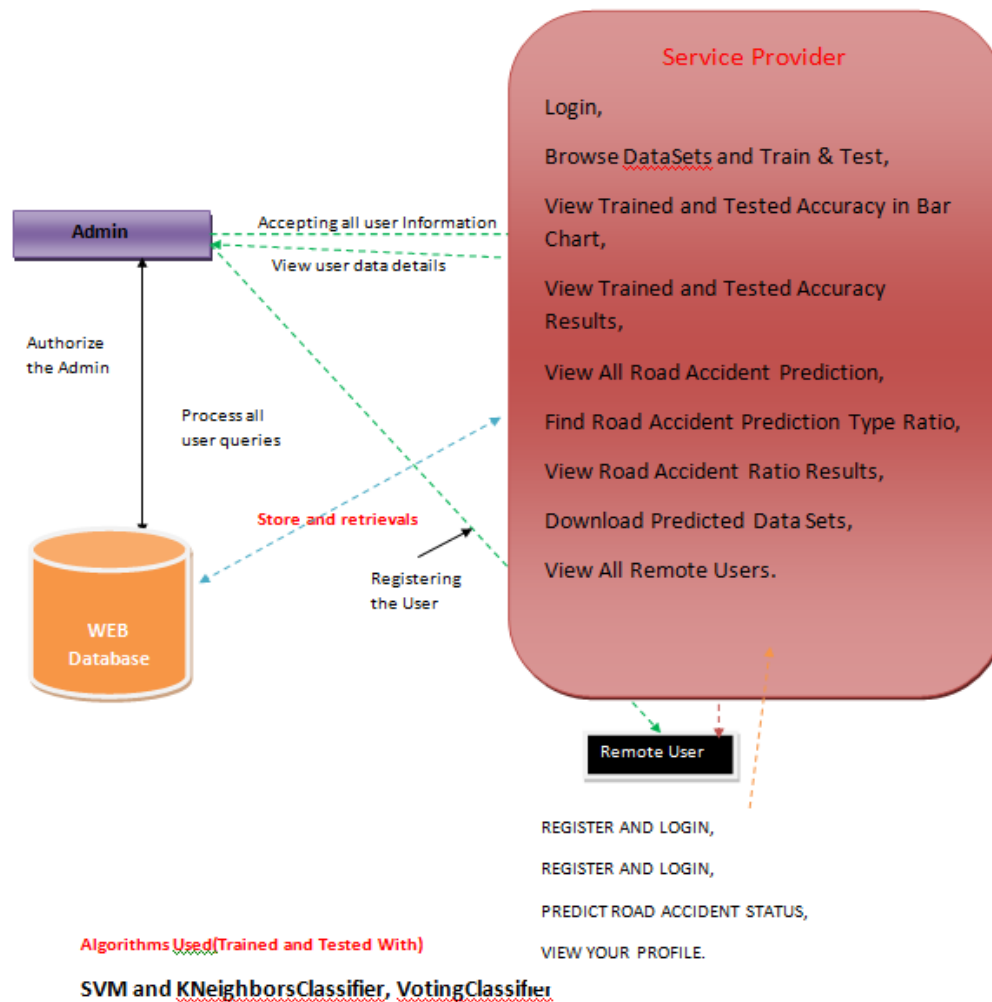
Punithavalli [10] have conducted an extensive survey on the different techniques used in road accident detection over the years, the approaches implemented in them and discusses their merits and de-merits.

George Yannis et al. [11], in their paper, have discussed about the current practices used in the development of accident prediction models on an international level. Detailed information on

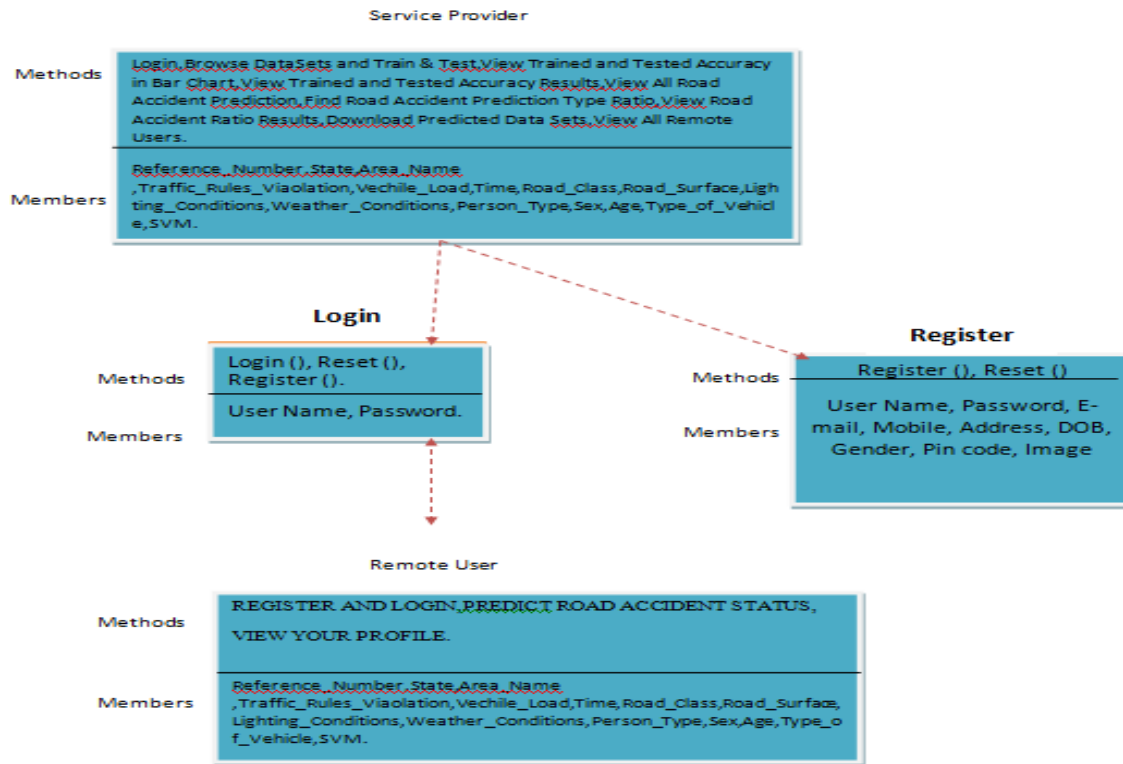
various models have been collected with the help of questionnaires and they have made use of this data to identify which could be the most useful model that can be applied for accident prediction.

Anand, J. V [12] has developed a method to determine the effect of different variables in the detection and prediction of atmospheric deterioration all over the world. Fuzzy C means clustering, R-studio, and the ARIMA frame work have been made use of in creating this method. A similar approach can also be tried in evaluating the impact of various factors on road accidents. Analyzing the original cause of accidents is important because this will tell us the impact factor and contribution of each attribute towards road accidents. Tiwari et al. [13] have made use of self-organizing maps, K-mode clustering techniques, Support Vector Machines, Naïve Bayes and Decision tree to classify the data from road accidents based on the type of road users.

Architecture Diagram



➤ **Class Diagram :**



3. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The

Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

4.3.3 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

4. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.

The application starts running when it is executed for the first time. The server has to be started and then the internet explorer is used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems

can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

5. CONCLUSION

An accident can change the lives of many people. It is up to each of us to bring down this increasing number. This can be made possible by adopting safe driving measures to an extent. Since all instances of accidents cannot be attributed to the same cause, proper precautionary measures will also need to be exercised by the road development authorities in designing the structure of roads as well as by the automobile industries in creating better fatality reducing vehicle models. One thing within our capability is to predict the possibility of an accident based on previous data and observations that can aid such authorities and industries. This project was successful in creating such an application that can help in efficient prediction of road accidents based on factors such as types of vehicles, age of the driver, age of the vehicle, weather condition and road structure, This model was implemented by making use of several data mining and machine learning algorithms applied over a dataset for Bangalore and has been successfully used to predict the risk probability of accidents over different areas with high accuracy.

The model can be further optimized in future to include several constraints that have been left out in the current study. These optimized models can be efficiently utilized by the government to reduce road accidents and to implement policies for road safety. Another scope of this work would be to develop a mobile app that will help the drivers in choosing a route for a ride. A call out to the driver through the maps service can also be implemented that would also announce the risk probability in a chosen route along with the directions. This can then be implemented by service provider companies such as Uber, Ola and so on in future. This will also be useful in having a better surveillance of accident prone areas and providing emergency services in the event of an accident. Better road safety instructions can also be installed along the highways taking into account the risks obtained from this model.

6. REFERENCES

- [1] <https://www.statista.com/topics/5982/road-accidents-in-india/>
- [2] Srivastava AN, Zane-Ulman B. (2005). Discovering recurring anomalies in text reports regarding complex space systems. In Aerospace Conference, IEEE. IEEE 3853-3862.

- [3] Ghazizadeh M, McDonald AD, Lee JD. (2014). Text mining to decipher free-response consumer complaints: Insights from the nhtsa vehicle owner's complaint database. *Human Factors* 56(6): 1189-1203. <http://dx.doi.org/10.1504/IJFCM.2017.089439>.
- [4] Chen ZY, Chen CC. (2015). Identifying the stances of topic persons using a model-based expectationmaximization method. *J. Inf. Sci. Eng* 31(2): 573-595. <http://dx.doi.org/10.1504/IJASM.2015.068609>
- [5] Williams T, Betak J, Findley B. (2016). Text mining analysis of railroad accident investigation reports. In 2016 Joint Rail Conference. American Society of Mechanical Engineers V001T06A009- V001T06A009. <http://dx.doi.org/10.14299/ijser.2013.01>.
- [6] Suganya, E. and S. Vijayarani. "Analysis of road accidents in India using data mining classification algorithms." 2017 International Conference on Inventive Computing and Informatics (ICICI) (2017): 1122-1126.
- [7] Sarkar S, Pateshwari V, Maiti J. (2017). Predictive model for incident occurrences in steel plant in India. In ICCCNT 2017, IEEE, pp. 1-5. <http://dx.doi.org/10.14299/ijser.2013.01>.
- [8] Stewart M, Liu W, Cardell-Oliver R, Griffin M. (2017). An interactive web-based toolset for knowledge discovery from short text log data. In International Conference on Advanced Data Mining and Applications. Springer, pp. 853-858. http://dx.doi.org/10.1007/978-3-319-69179-4_61.
- [9] Zheng CT, Liu C, Wong HS. (2018). Corpus based topic diffusion for short text clustering. *Neurocomputing* 275: 2444-2458. <http://dx.doi.org/10.1504/IJIT.2018.090859>.
- [10] ArunPrasath, N and Muthusamy Punithavalli. "A review on road accident detection using data mining techniques." *International Journal of Advanced Research in Computer Science* 9 (2018): 881-885.

META HEURISTIC OPTIMIZATION ALGORITHM BASED FEATURE SELECTION FOR CLINICAL BREAST CANCER DIAGNOSIS

Sheik Imran (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari
District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

Abstract :

The paper offers a crossbreed streamlining algorithm combining harmony search (HS) and simulated annealing (SA) known as harmony search and simulated annealing (HS-SA) for precise and accurate breast malignancy. Additionally, an improved wavelet-based contourlet transform (WBCT) system for feature extraction explores to get the highlights of the region of interest (ROI), permitting performance improvement over other standard methodologies. In the mined feature space, the projected HS-SA algorithm intends to diminish the feature dimensions and congregate at the unprecedented feature set. The SVM classifier backed with diverse kernel functions is used for classification, which is fed by the chosen features, and its exhibition contrasts with the conventional machine learning classification and optimization techniques. The actualized computer-aided diagnosis (CAD) learning mechanism is challenged by evaluating its findings. It examines two different breast mammographic datasets i) benchmark BCDR-F03 dataset and ii) local mammographic dataset. Trial reproductions, empirical outcomes, and measurable examinations likewise indicate that the proposed model is practical and advantageous for the arrangement of malignant breast growth. The findings show that the proposed CAD framework (HS-SA + kernel SVM) is better than different characterization accuracy procedures (with an accuracy of 99.89% for the local mammographic dataset and 99.76% for benchmark BCDR-F03 dataset, AUC of 99.41% for the local mammographic dataset and 99.21% for reference BCDR-F03 dataset), while keeping the feature space limited to just seven feature subsets and computational prerequisites as low as is prudent.

1.INTRODUCTION

Body cells change exponentially when they are influenced by malignant growth in breast cancer, inevitably turning into a protuberance or mass of tumor. Most of the diseases of breast cancer arise, in particular, in the segments of milk-creating organs, called lobules, and channels linked to lobules, and finally to the nipple. The breast includes greasy, connective, and lymphatic tissues.

Comparing to 22% of the new cases every year, breast disease is the most serious of all types of malignant growth, as it has occurred at a rapid pace. There are more than 200 kinds of cancer, but the pace and velocity by which breast cancer affects women worldwide is incomparable. Among US women in 2017, an expected 2,52,710 new instances of obtrusive breast malignancy, 63,410 new instances of breast carcinoma in situ, and 40,610 breast disease passing were witnessed as per the International Agency for Research on Cancer Disease (IARC) report. Breast disease represents 23% of the all-out malignant growth cases and 14% of the malignant growth demise in both developed and developing nations. The estimate is that more than 1.6 million new cases of breast cancer have occurred among women worldwide in 2010 alone. In 2011, almost 1.7 million people were targeted for breast disease. Measurements say that in the USA, 527 new instances of breast disease were analyzed every day, and 110 individuals died each day. In 2016–2017 insights, as per the American Cancer Society (ACS), many new instances of breast malignancy were expected in the US. It incorporates 18.3% of all malignant growth types in Egypt. In the event that a woman lives at the age of 85, there is one in eight possibilities (12%) for her to develop breast disease at some point in her life, because the risk of malignant breast growth increases significantly, paying little attention to her family ancestry as a woman ages. Indian females are most provoked by breast cancer by being the number one malignant growth among them, which compels the likewise inclination to its risks for a developing country. Factual research has shown a mortality of 12.7 per 100,000 women in parallel to an age adjusted rate as high as 25.8 per 100,000 women. In 2015–2016, women were influenced by malignant breast growths, and the passing rate was 50 per cent as communicated by the findings from the Indian Council of Medical Research (ICMR). Pakistan has to face an excess of 40,000 deaths annually because of it. After coronary disease and mishaps, malignant growth is the third cause

of death in Iran, accounting for 24.6 percent of all cancers. The average age of a woman who has breast cancer is 49.6 years.

Malignancy is the primary source of death around the world, representing 8.2 million in 2012. It is normal that the number of cases of annual disease will rise from 14 million in 2012 to 22 million in the next two decades. Malignant growth control and far-reaching avoidance plan are hence vital. Interim, more accentuation ought to be put on the early inference of maladies to add to the patient's lifetime. The timely seizure of breast cancer has significance for endurance, especially in low-wage countries where assets are extremely constrained, and the diagnosis of late-stage illness makes the burden more troublesome.

Mammography screening of asymptomatic women is a proven non-invasive technique for decreasing mortality from breast malignant by as much as 30%. Woman above 40, should undergo mammography screening once a year as recommended by the ACS. Quite often, mammogram discovers something that resembles malignancy, that results in false positive (FP). Also, radiologists may miss up to 30% of breast malignant growths contingent upon the thickness of the breast. In light of World health organization's (WHO) report, around 33% of disease is treated through early analysis unfurling the occurrence of breast cancer can be stopped provided early diagnosis becomes a reality. Improving malignancy anticipates a variety of techniques, that aims to enable the individuals to follow successful strategies for the prevention of disease. All these compelling factors have convinced a great deal of research over the last decades, focusing on the advancement of computational frameworks to assist the doctor in deciphering radiological pictures. These CAD frameworks have increased space in current medications, filled in as a data hotspot for authorities, and expanded the pace of correct discovery in the recognizable evidence of genuine infections, such as breast malignancy.

Notwithstanding, examinations found in writing utilize similar strategies and setups for both thick and non-thick masses, while those systems could be progressively fitting for a particular sort of thickness. Magnetic resource imaging (MRI), self and clinical breast checks, ultrasound, and mammography are some of the screening strategies utilized for malignant breast growth. The most accurate and straightforward system for distinguishing breast cancer is in favour of mammography. Film mammography is superseded by advanced digital mammography, where outstanding mechanized hardware is used to record patient breast images and for additional

treatments such as detection and classification. Microcalcifications and masses are the most well-known anomaly that stimulates breast malignant growth. The healthcare data analytics shifted the conventional healthcare services to recent evidence and cure based soon after the birth of machine learning and pattern recognition fields. The analysis of biomedical data (biomedical images like MRI, CT scan, PET, US, mammography, and biomedical signals like EEG, ECG, EMG) has led to the design of automated and smart CAD systems that help to diagnose early, accurate and precise diseases even before symptoms are visible externally. With the development of artificial intelligence techniques, the race of data-driven intelligent classification approaches have been applied for breast cancer diagnosis, such as Naive Bayesian, Neural Network, Support Vector Machine (SVM), ensemble methods, K-means, fuzzy and rough set techniques, PSO, semi-supervised techniques, deep learning, transfer learning Active learning or other hybrid algorithms.

2. LITERATURE SURVEY

In 2011, B. Zhang et al. proposed a random Subspace cascade with rejection options for the classification of microscopic biopsy images. A. E. Hassanien and T. Kim in 2012 familiarized an amalgam method that cartels the rewards of fuzzy sets, pulse coupled neural networks (PCNNs), and support vector machine, in conjunction with wavelet-based feature for best cancer classification. In 2013, V. Balanica et al. presented four new methods for mining the speculation feature of a perceived breast lesion on mammography. The authors in Ref. et al. displayed a fast-orthogonal search (FOS) that delivers competent iterative way to compute step by step regression, and can select features with a predictive value from a set of kinetic and texture candidate features computed from dynamic contrast-enhanced magnetic resonance images. L. Taifi et al. in 2014 presented a preprocessing method, grounded on homomorphic filtering and wavelet, for the removal of irregularities in mammographic images. W. Sun et al. planned a three-stage Semi-Supervised Learning (SSL) method for refining presentation of computerized breast cancer analysis with undiagnosed data. In 2015, N. P. Pérez et al. offered a new feature selection method (named uFilter) that advances the Mann-Whitney U test for tumbling dimensionality and ranking features in binary classification problems. In 2015, authors et al. proposed a novel local energy-based shape histogram (LESH) as the feature set for the appreciation of irregularities in mammograms. In 2016, H. Kong et al. proposed Jointly Sparse

Discriminant Analysis (JSDA) to sightsee the main factors in breast cancer to enlighten the accuracy in diagnosis and prediction. In 2017, W. Sun et al. developed a graph-based semi-supervised learning (SSL) scheme using a deep convolutional neural network (CNN) for breast cancer diagnosis. In 2018, S. Liu et al. utilized a Bayesian network (BN) modelling approach for breast cancer. In 2019, L. Tsochatzidis et al. investigated Deep convolutional neural networks (CNNs) in the context of computer-aided diagnosis (CADx) of breast cancer. Q. Xu et al. in 2019 developed a CAD based on a CNN network that aims to classify breast mass lesions in optical tomographic images. T. A. Shaikh et al. in 2020, proposed a LUPI-based CAD framework for breast cancer using privileged information that enhances the performance of a single-modal imaging-based CAD for breast cancer by relocating PI.

This research aims to diagnose breast cancer based on the characteristics of the extricated tumor. The classifiers are a vital ingredient in data mining methods whose quality of performance is fully reliant on the feature extraction and selection. The presentation of the estimator, either in terms of learning speed, generalization dimensions or straight forwardness, progressed the right choice of the features founded on classification and clustering methods, numerous approaches is applied for breast cancer diagnosis in the recent literature. However, the bulk rise in the amount of available data (both features and records) from recent years has constrained the old methodologies, and thus, the meta-heuristics approaches employ feature selection and dipping the number of features grew its usage in the data mining field. The current investigation condenses the prevailing feature space to diminish the computational cost for SVM training and preserve a parallel diagnosis accuracy. To excerpt beneficial information and diagnose the tumour, a hybrid of HS-SA search and support vector machine (K-SVM) algorithms is developed. The HS-SA algorithm is exploited to distinguish the unseen designs of the benign and malignant tumours distinctly and discover the best features for clinical breast cancer diagnosis. Seven subsets of abstract tumour features are mined from the original 21 features subsets for the training phase. Despite the fact that the K-SVM lessens dimensionality of input feature space, the high prediction accuracy is sustained. From the computation time perspective, the proposed strategy diminishes the training time altogether by diminishing the number of input features. To the best of our knowledge, the proposed amalgamation of HS-SA-SVM hybrid strategy with optimal feature subset is novel and treated as a new effort in the after-mentioned direction.

3. EXISTING SYSTEM

This section presents the detailed description of the feature selection problem with the mathematical model and the definitions, concepts and the classifications of metaheuristic algorithms.

A. FEATURE SELECTION

Feature selection deals with inappropriate, irrelevant, or unnecessary features. It is a process that extracts the best features from the datasets. Feature selection is one of the most critical and challenging problems in machine learning. The various applications of the feature selection problem can be demonstrated in different fields. There are some applications such as biomedical problems (to find the best gene from candidate gene); text mining (to find the best terms word or phrases); image analysis (to select the best visual contents pixels, colour) etc.

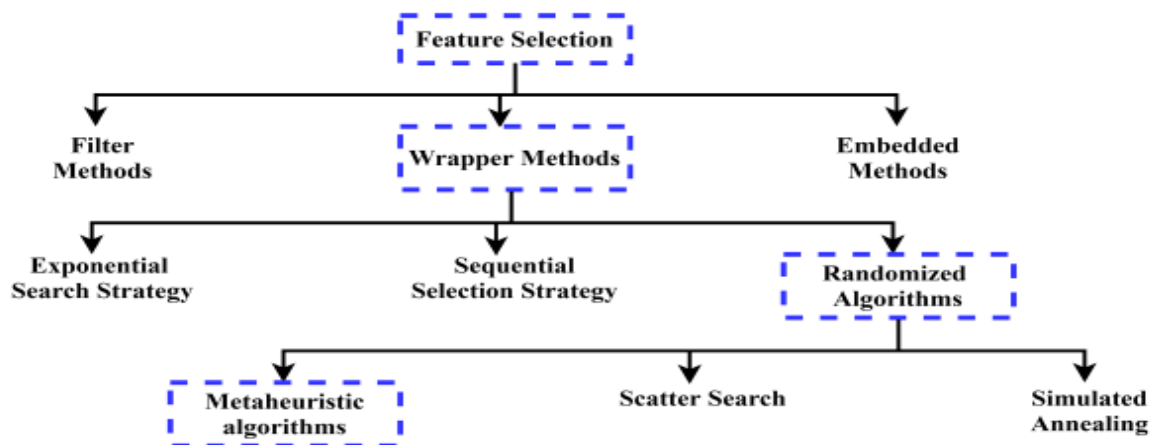
Mathematically, a feature selection problem can be formulated in the following way:

Assume a dataset S contains d number of features. Then the working mechanism of feature selection problem is to select relevant features among d features. Given dataset $S = \{f_1, f_2, f_3, \dots, f_d\}$ The objective is to select the best subsets of features from S . Extract Subset $D = \{f_1, f_2, f_3, \dots, f_n\}$ where, $n < d$ and $f_1, f_2, f_3, \dots, f_n$ represents the features/attributes of any dataset. Figure 2 depicts the working mechanism of the feature selection process. From the figure, it can be observed that there are five main components of the feature selection process, i.e. original dataset, selection of feature subset, evaluation of feature subset, selection criterion and validation.

Several feature selection methods are developed to obtain the best subset of features. Generally, the techniques are classified into three categories filter, wrapper and embedded methods. Filter methods are independent of learning or classification algorithm. It always focuses on the general characteristics of the data. Wrapper methods always include the classification algorithm and interact with the classifier. These are computationally expensive methods than the filter and also provide more accurate results as compared to filter methods. Embedded methods are a combination of filters and wrapper methods. In embedded methods, the feature selection is a part

of the training process and training process held with the classifier. Moreover, the embedded methods use learning algorithm in its process, they will be considered in wrapper approaches category.

Wrapper approaches present better results in comparison with filter methods, but they are slower than filters methods. Wrapper methods depend on the modelling algorithm in which every subset is generated and then evaluated. Subset generation in wrapper methods is based on the different search strategy. Jovic et al. differentiates search techniques into three categories; exponential, sequential and randomized selection strategy. In the exponential method, the number of evaluated features increases exponentially with the size of features. Although this method shows accurate results, it is not practically possible to apply because of the high computational cost. The examples for exponential search strategy are exhaustive search, branch and bound method . Sequential algorithms include or remove features sequentially. Once a feature is included or removed in the selected subset, it can not be further changed that leads to local optima. Some sequential algorithms are linear forward selection, floating forward or backward selection, best first etc. Randomized algorithms include randomness to explore the search space, which saves the algorithms from trapping into local optima. Randomized algorithms are commonly known as population-based approaches for example



simulated annealing, random generation, metaheuristic algorithms etc. . We do not present a detailed description of every method of the feature selection process. The detailed explanation of

each method can be found in . The flow chart of categorization of methods for solving feature selection is shown in Figure . In the figure, the dashed line box represents the methodology of this paper which describes how we reach to metaheuristic algorithms.

4. METAHEURISTIC ALGORITHMS

Metaheuristic algorithms are optimization methods that obtain the optimal (near-optimal) solution of optimization problems. These algorithms are derivative-free techniques and, have simplicity, flexibility and capability to avoid local optima . The behaviour of metaheuristic algorithms are stochastic; they start their optimization process by generating random solutions. It does not require to calculate the derivative of search space like in gradient search techniques. The metaheuristic algorithms are flexible and straightforward due to the simple concept and easy implementation. The algorithms can be modified easily according to the particular problem. The main property of metaheuristic algorithms is that they have a remarkable ability to prevent the algorithms from premature convergence. Due to the stochastic behaviour of algorithms, the techniques work as a black box and avoid local optima and explore the search space efficiently and effectively. The algorithms make a tradeoff between its two main essential aspects exploration and exploitation . In the exploration phase, the algorithms investigate the promising search space thoroughly, and exploitation comes for the local search of promising area(s) that are found in the exploration phase. They are successfully applied to various engineering and sciences problems, e.g. in electrical engineering (to find the optimal solution for power generation), industrial fields (scheduling jobs, transportation, vehicle routing problem, facility location problem), in civil engineering (to design the bridges, buildings), communication (radar design, networking), data mining (classification, prediction, clustering, system modelling) etc. Metaheuristic algorithms classify into the following two main categories;

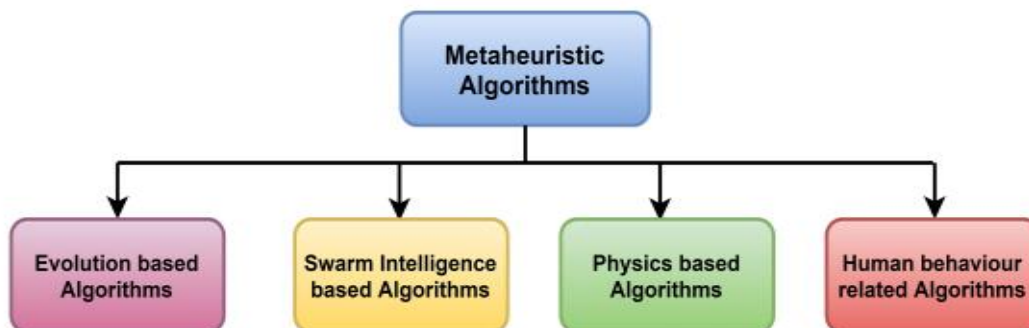
- (i) Single solution based metaheuristic algorithms: These techniques start their optimization process with one solution, and their solution is updated during the iterations. It may lead to trapping into local optima and also does not explore the search space thoroughly.
- (ii) Population (multiple) solution based metaheuristic algorithms: Initially, these algorithms generate a population of solutions and start their optimization process. The population of solutions update with the number of generations/iterations. The algorithms are beneficial for avoiding local optima as multiple solutions assist each other and have a

great exploration of search space. They also have the quality of jump towards the promising part of search space. Therefore, population-based algorithms use in solving most of the real-world problems

Researchers pay great attention to metaheuristic algorithms because of their characteristics. Several algorithms have been designed and solved different types of problems. Based on their behaviour, the metaheuristic algorithms can be divided into four categories; evolution-based, swarm intelligencebased, physics-based and human-related algorithms . The categorization of the algorithms is depicted in Figure

(1) Evolution based algorithms:

It is inspired from the natural evolution and start their process with randomly generated population of solutions. In these type of algorithms, the best solutions are put together to create new individuals. The new individuals are formed using mutation, crossover and select the best solution. The most popular algorithm in this category is Genetic algorithm (GA) that is based on Darwin evolution technique . There are other algorithms such as evolution strategy genetic programming , tabu search , differential evolution etc.



4.1 METAHEURISTIC ON FEATURE SELECTION :

It describes the metaheuristic algorithm, which has been used in solving the feature selection problem. Binary vectors representations are considered to obtain the relevant feature. In the

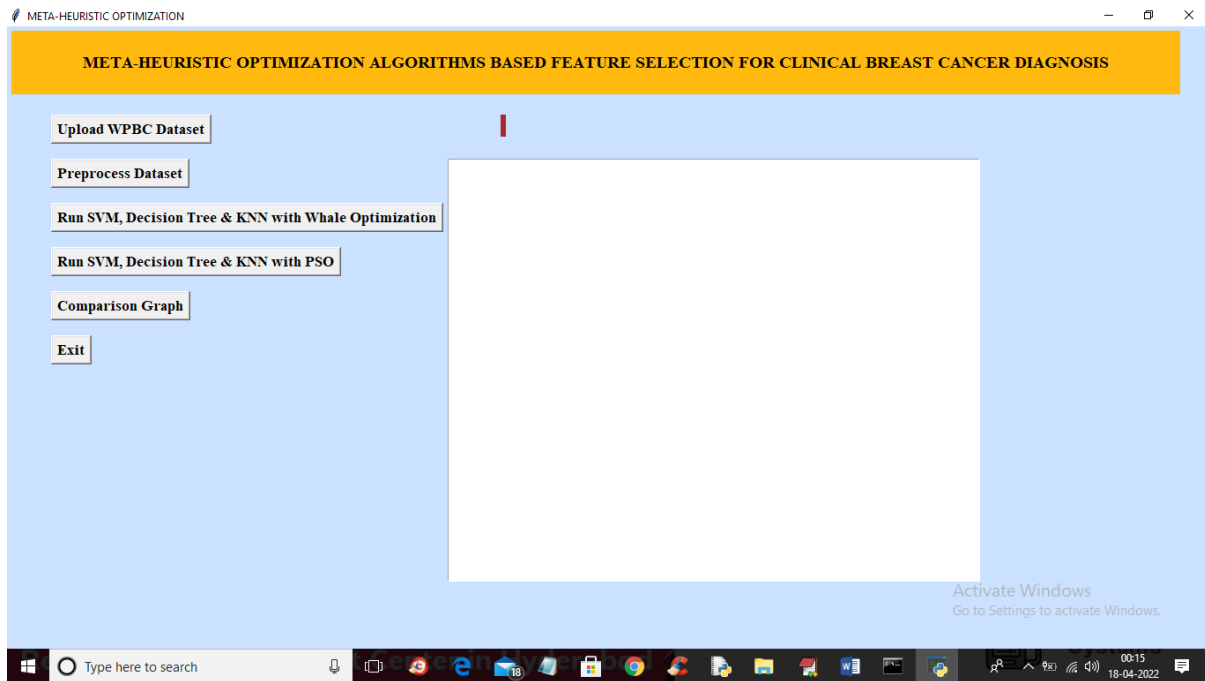
designed algorithm, a solution vector is represented by (10101100) this implies that 1 means that a particular feature is selected and 0 means that feature is not selected in the subset. Hence, this section investigates all binary variants of metaheuristic algorithms in detail. The first section describes the evolution-based algorithms; the second describes the swarm intelligence based algorithms, third demonstrates the physics-based algorithms, and the fourth one is for the human-related algorithm. And the last section is for the hybrid algorithms, which are a combination of two or more metaheuristic algorithms that have been used for classification problems.

A. EVOLUTION BASED ALGORITHMS

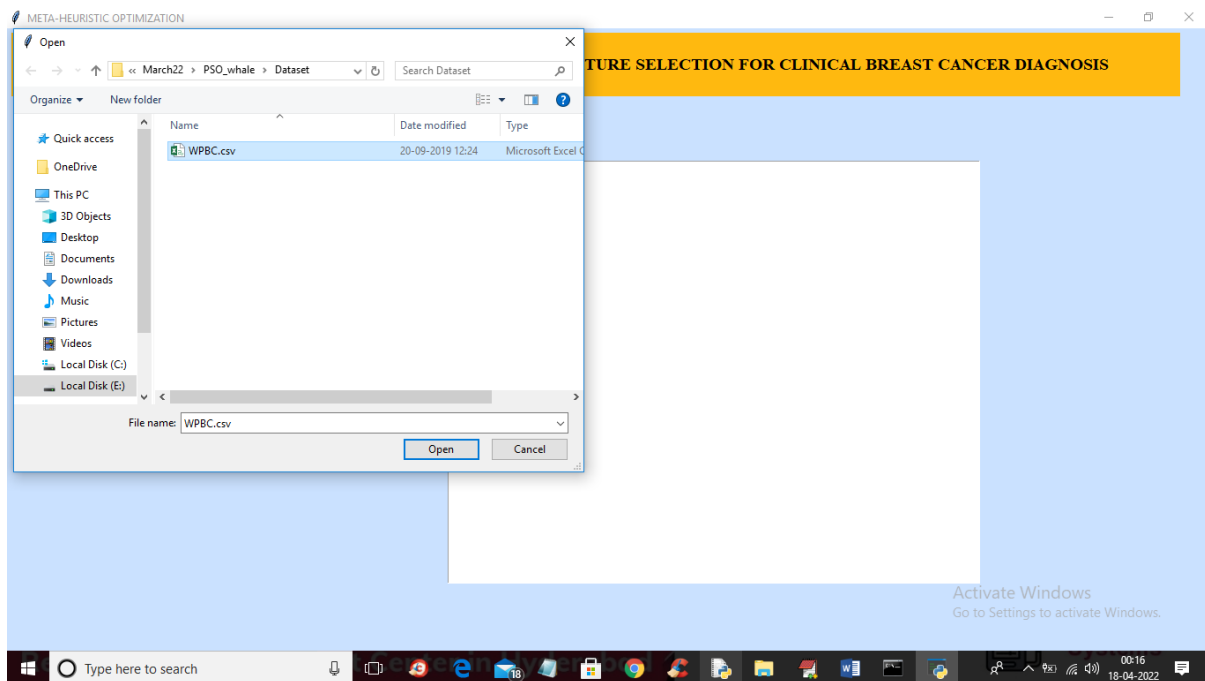
From Table 1, it can be seen that there are very few algorithms are developed in evolution based category from 2009- 2019. Gan and Duan proposed a chaotic differential search algorithm for image processing and it has been combined with lateral inhibition to edge extraction and image enhancement. Negahbani et al. used differential search algorithm for the diagnosis of coronary artery disease with fuzzy c-means that was used as a classifier. The performance of the proposed approach has been evaluated using accuracy, sensitivity and specificity measures. Zhang et al. proposed binary backtracking algorithm for wind speed forecasting in which extreme learning machine was employed for feature selection. Binary backtracking algorithm was developed using a sigmoidal function that transforms the continuous variables to binary variables. To identify the Leukemia cancer symptoms, Dhal et al. implemented the stochastic fractal search algorithm to provide optimal identification. The developed algorithm was compared with other classical methods and achieved high accuracy. Besides, a binary stochastic fractal search was developed to classify the galaxy colour images with extreme machine learning

5. SCREENSHOTS

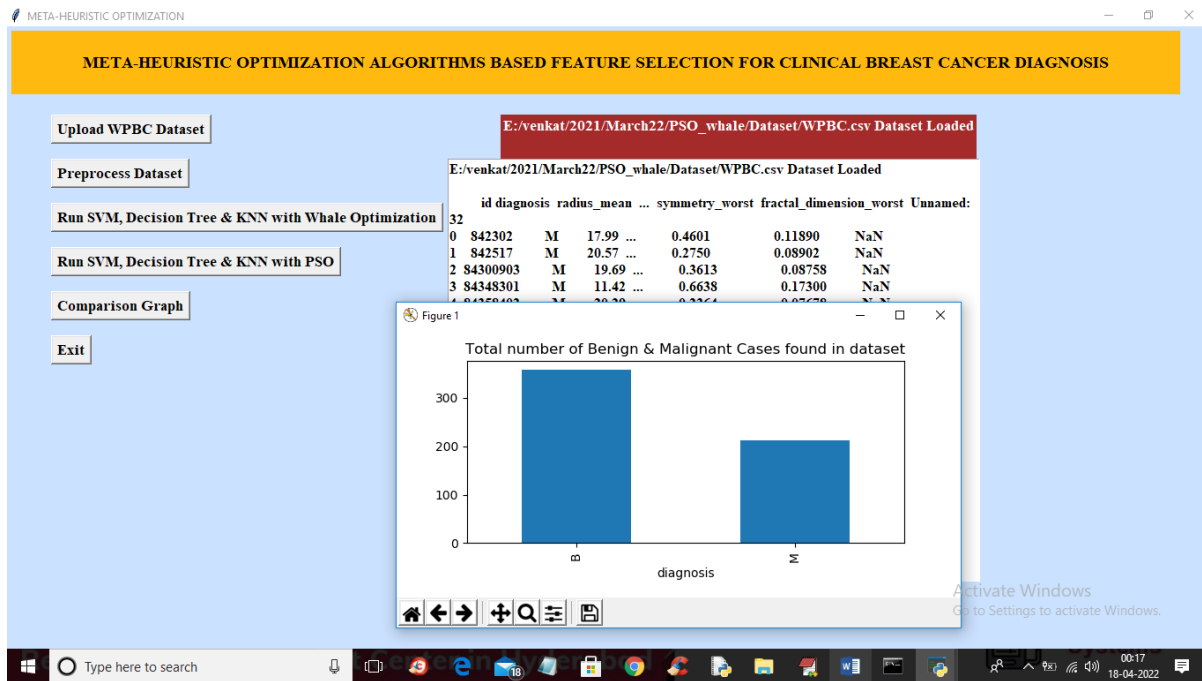
To run project double click on 'run.bat' file to get below screen



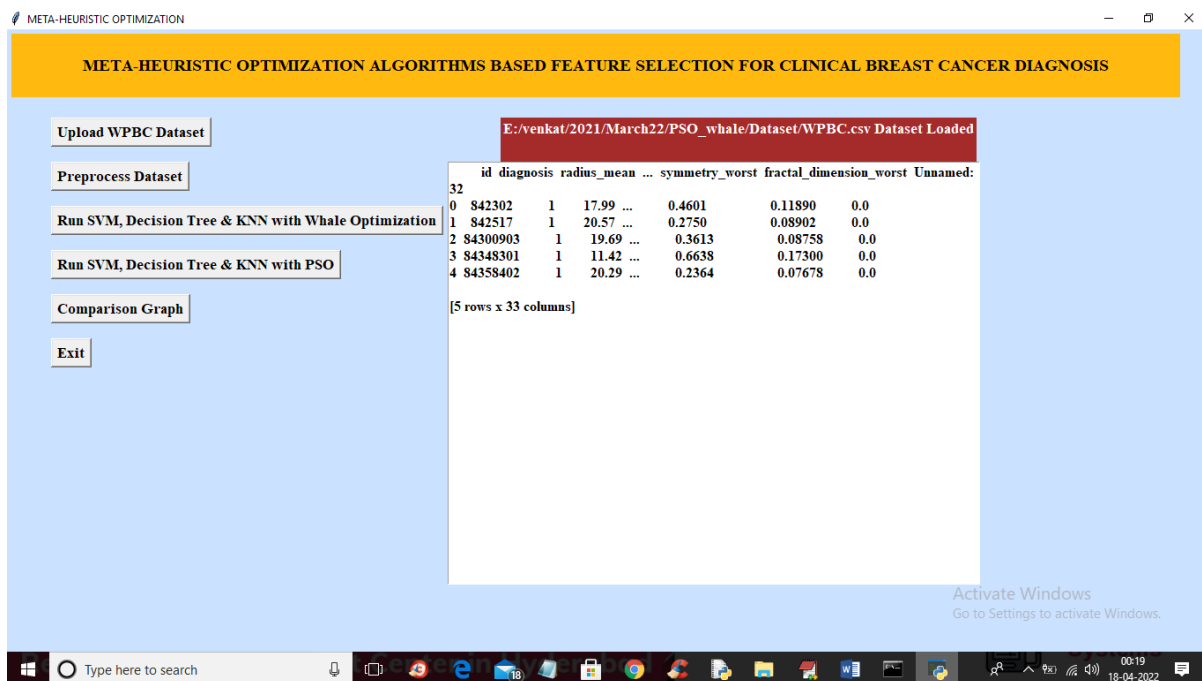
In above screen click on ‘Upload WPBC Dataset’ button to upload dataset and to get below screen



In above screen selecting and uploading dataset folder and then click on ‘Open’ button to load dataset and to get below screen

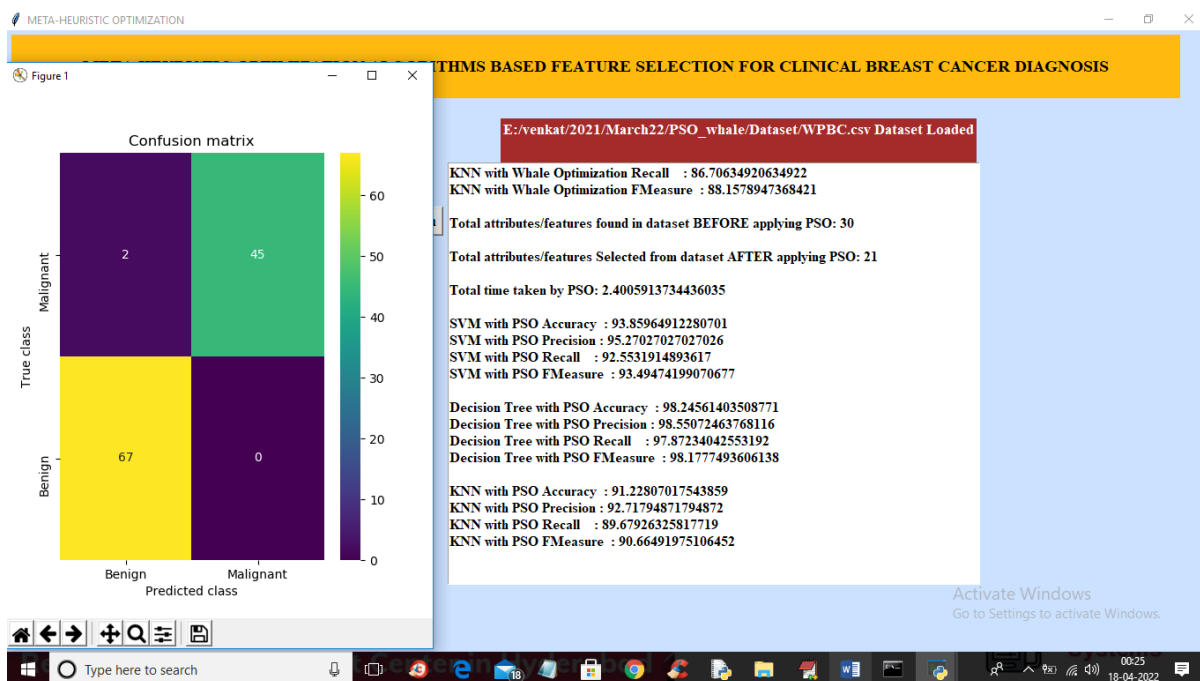


In above screen we can see dataset loaded and in dataset contains non-numeric values and missing NAN values and in above screen we can see graph showing number of benign and malignant cases found in dataset and now close above graph and then click on 'Preprocess Dataset' button to replace missing values and convert non-numeric data to numeric data and get below output



In above screen we can see all values are converted to numeric and now click on ‘Run SVM, Decision Tree & KNN with Whale Optimization’ button to apply whale optimization and train all ML algorithms to get below output

In above screen in first two lines we can see dataset contains 30 attributes and after applying whale we got 2 important attributes and then we can see accuracy of each algorithm on selected features and we can see execution time also and in above screen we can see prediction confusion matrix graph where application predict 67 records as benign correctly and only 6 records are incorrectly predicted and now close above graph and then click on ‘Run SVM, Decision Tree & KNN with PSO’ button to select features with PSO and train all algorithms to get below output



In above screen we can see PSO selected 21 attributes out of 30 and we can see accuracy of each algorithm on selected features and in confusion matrix we can see with PSO 67 records are correctly predicted as Benign and only 2 records are incorrectly predicted and now close above graph and then click on ‘Comparison Graph’ button to get below graph

6. CONCLUSION

This paper presents a comprehensive survey on metaheuristic algorithms that are developed from 2009 to the 2019 year and their binary variants, which have been applied to feature selection

problem. A detailed description and mathematical model of feature selection problem are given that could help researchers to understand the problem properly. Moreover, the techniques of solving feature selection problems are presented. Additionally, metaheuristic algorithms are considered in solving the feature selection problem. Therefore, basic definition, importance and the classification of metaheuristic algorithms are given. The evolution-based, swarm-based, physics-based category, human related algorithms have been developed and applied to feature selection problems.

However, metaheuristic algorithms have some following drawbacks:

- They suffer from slow convergence rate due to random generation movement.
- They explore the search space without knowing the search direction.
- They can trap into local optima, or they have some premature convergence.
- The values of the parameters used in the metaheuristic algorithms have to be adjusted, this may also lead to pre-mature convergence.

Besides, the limitation of the metaheuristic algorithms, the modified and enhanced version of the algorithms were developed which are successfully applied to the feature selection problems. Also, a categorization is presented based on the behaviour of algorithms; evolution-based, swarm-based, physics-related and human behaviour related algorithms. This paper benefits in such a way that a list of metaheuristic algorithms is presented based on their classification. It also benefits for the application point of view as it consists of a case study. The case study presents the eight benchmark datasets and the optimal feature subsets are found by implementing different metaheuristic algorithms.

evolution and human-related category, but there are several algorithms have been designed in the swarm and physics-related algorithms. It implies that there is a scope to develop or propose new metaheuristic algorithms in these categories. This paper mainly focuses on solving the feature selection problem using binary variants of metaheuristic algorithms. Hence, extensive literature is presented in every class of metaheuristic algorithms. All binary variants of all reviewed algorithms regarding feature selection problems are pointed. In swarm-based category, all binary variants of Cuckoo search, Bat algorithm, Firefly algorithm, flower

pollination algorithm, Krill herd algorithm, Grey wolf optimizer, Ant lion optimizer, Dragonfly algorithm, Whale optimization algorithm, Grasshopper optimization algorithm, Salp swarm algorithm are reviewed with the key factor of solving feature selection problem. Moreover, hybrid approaches are also reviewed in the process of solving the feature selection problem.

It can be concluded that there is some area(s) which are less explored, such as spam detection, theft detection and weather prediction. However, lots of research has been done on the well-known datasets of UCI repository and in medical diagnosis (cancer classification), intrusion detection systems, text classification, multimedia etc. Hence, researchers should pay great attention to explore this area with metaheuristic algorithms. Moreover, there are some algorithms in the literature for which binary variants are not developed yet such as PFA, CGS, TCO, ES, HSO, WSA, BMO, OptBees, TGSr, EVOA, VCS, EPC, GbSA, CSO, WEO, LCA, EMA, VPL. These algorithms benefit classification after developing their binary version. From the literature, it can be observed that the researcher has to face many challenges to obtain the best feature subset of the considered classification problem. A good choice of classifier has a significant impact of the quality of obtained solution such KNN classifier is the most used classifier in getting the best subset with well-known datasets of UCI repository. After that, SVM classifier used to classify in different applications such as medical diagnosis, pattern recognition, image analysis etc. There are some other classifiers which are less used in terms of classification. Hence, this another gap to use different classifiers in classification problem and compared with most used ones. Finally, researchers will get the benefit of this study as they could find all the key factors in solving the feature selection problem using metaheuristic algorithms under one roof.

7. REFERENCES

- [1] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.
- [2] P. Y. Lee, W. P. Loh, and J. F. Chin, "Feature selection in multimedia: The state-of-the-art review," *Image Vis. Comput.*, vol. 67, pp. 29–42, Nov. 2017.

- [3] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Comput. Biol. Med.*, vol. 112, Sep. 2019, Art. no. 103375.
- [4] M. Sharma and P. Kaur, “A comprehensive analysis of nature-inspired meta-heuristic techniques for feature selection problem,” *Arch. Comput. Methods Eng.*, pp. 1–25, Feb. 2020, doi: 10.1007/s11831-020-09412-6.
- [5] M. Z. Asghar, A. Khan, S. Ahmad, and F. M. Kundi, “A review of feature extraction in sentiment analysis,” *J. Basic Appl. Sci. Res.*, vol. 4, no. 3, pp. 181–186, 2014.
- [6] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.
- [7] V. Bolón-Canedo and A. Alonso-Betanzos, “Ensembles for feature selection: A review and future trends,” *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019.
- [8] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering,” *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.
- [9] S. Ahmed, M. Zhang, and L. Peng, “Enhanced feature selection for biomarker discovery in LC-MS data using GP,” in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2013, pp. 584–591.
- [10] M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, “Text feature selection using ant colony optimization,” *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6843–6853, Apr. 2009.
- [11] A. Ghosh, A. Datta, and S. Ghosh, “Self-adaptive differential evolution for feature selection in hyperspectral image data,” *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1969–1977, Apr. 2013.
- [12] M. Dash and H. Liu, “Feature selection for classification,” *Intell. Data Anal.*, vol. 1, nos. 1–4, pp. 131–156, 1997.
- [13] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.
- [14] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, “Feature selection: An ever evolving frontier in data mining,” in *Feature Selection in Data Mining*. Hyderabad, India, 2010, pp. 4–13.

PREDICTION OF AIR POLLUTION BY USING MACHINE LEARNING ALGORITHM

Siddireddy Sai Lakshmi Annapurna (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District,
Andhra Pradesh, India, 534202.

ABSTRACT

controlling and defensive the higher air greatness has gotten one in everything about first imperative occasions in different creating and metropolitan districts at the present. The greatness of air is adversely contacting collectible to the different styles of tainting influenced through the transportation, power, powers consumptions, and so forth. In our country population is a big problem as day by day population is increasing, so the rapid increasing in population and economic upswing is leading environment problems in city like air pollution, water pollution etc. In some of air pollution and air pollution is direct impact on human body. As we know that major pollutants are arising from Nitrogen Oxide, Carbon Monoxide & Particulate matter (PM), SO₂ etc. Carbon Monoxide is arising due to the deficient Oxidization of propellant like as petroleum, gas, etc. nitrogen oxide (NO) is arising due to the ignition of thermal fuel; Sulphur Dioxide(SO₂) is major spread in air, So₂ is a gas which is present more pollutants in air, it's affect more in human body. the predominance of air is overstated by multidimensional impacts containing spot, time and vague boundaries. The goal of this improvement is to take a gander at the AI basically based ways for air quality expectation. In this paper we will predict of air pollution by using machine learning algorithm.

1. INTRODUCTION

The Environment describe about the thing which is everything happening in encircles the Environment is polluted by human daily activities which include like air pollution, noise pollution. If humidity is increasing more than automatically environment is going more hotter. Major cause of increasing pollution is increasing day by day transport and industries there are 75 % NO or other gas like CO, SO₂ and other particle is exist in environment.. The expanding scene, vehicles and creations square measure harming all the air at a feared rate.

Therefore, we have taken some attributes data like vehicles no., Pollutants attributes for prediction of pollution in specific zone of Delhi

2. EXISTING SYSTEM

The Air Pollution Forecasting System: Air Quality Index (AQI) is a record that gives the public the degree of contamination related with its wellbeing impacts. The AQI centers around the different wellbeing impacts that individuals may encounter dependent fair and square and long stretches of introduction to the poison concentration. The AQI values are not quite the same as nation to nation dependent on the air quality norm of the country.

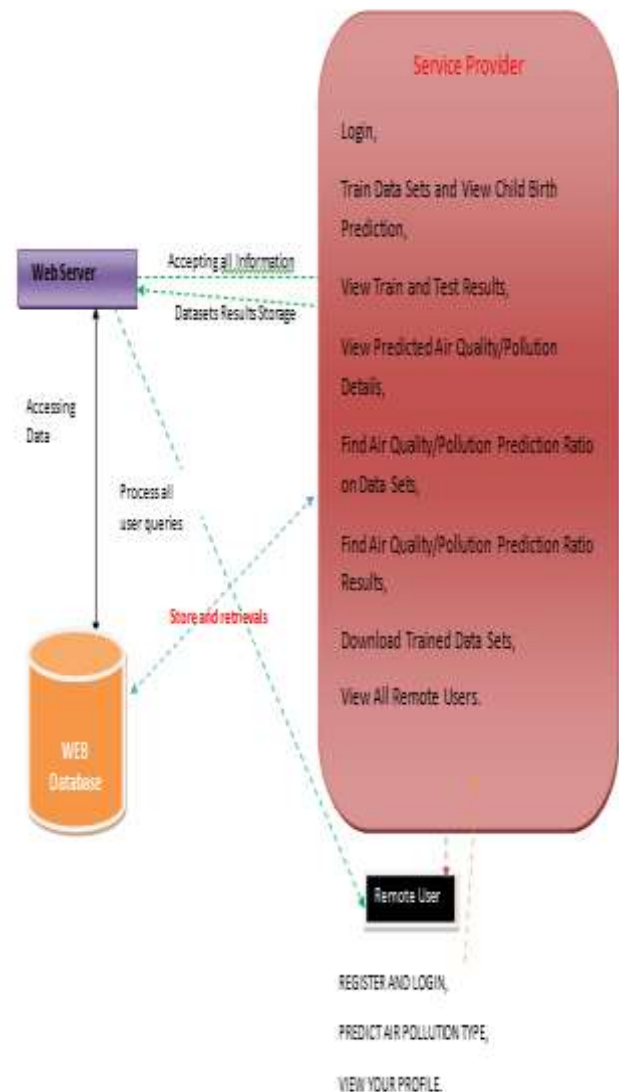
The higher the AQI level more noteworthy is the danger of wellbeing related problems. The by and large point of this venture is to make a student calculation that will have the option to foresee the hourly contamination focus. Additionally, an Android application will be built up that will provide the clients about the constant contamination convergence of PM2.5 alongside the hourly forecasted value of the toxin fixation from the student calculation. The Android application will also recommend data of the less dirtied[1].

3. PROPOSED SYSTEM

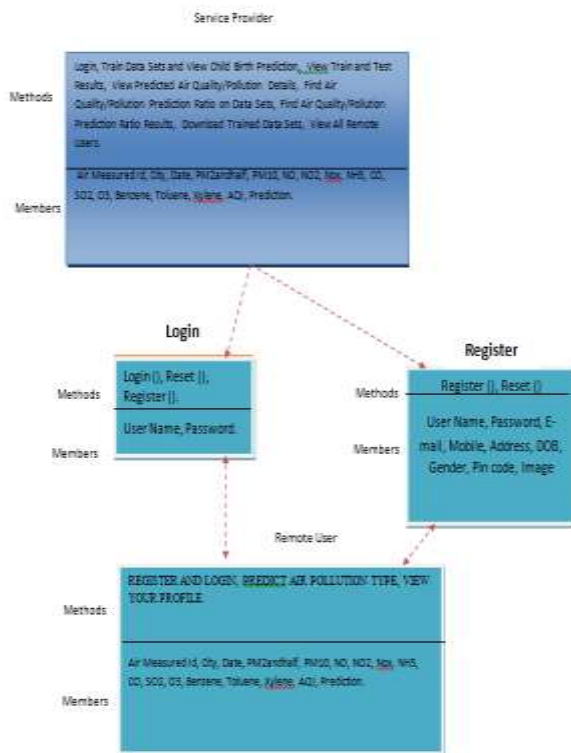
- 1) Data assortment: There is a different method from which we collected data from various dependable sources like Delhi Gov. site.
- 2) Exploratory examination: We research and explore examination with various parameter like ID of outliers, consistency check, missing qualities, and so on, it's totally occurred in this period of the venture.
- 3) Data Manipulation control: In period of data control stage the required missing data need to insert in utilizing the mean estimations of that characteristic of information. [2]
- 4) Prediction of boundaries utilizing by gauge model: For appropriate data indirect relapse we have to keep future qualities for different boundaries just
- 5) Implementation of straight relapse: Whenever all the boundaries become in active mode or they are accessible mode, the direct relapse calculation would be used in anticipate the air quality index (AQI).

- 6) Data accuracy investigation: We have to analyze that used model is being fit for overall data or not so we have to cross check root mean error, absolute percentage error then after we have to assume this factor is good for accuracy or not.

Architecture Diagram



> |Class Diagram :



CONCLUSION

Precision of our model is very acceptable. The anticipated AQI has a precision of 96%. Future upgrades incorporate expanding the extent of district and to incorporate whatever number locales as could be allowed as of now this venture targets foreseeing the AQI estimations of various areas of close by New Delhi. Further, by utilizing information of various urban areas the extent of this venture can be exhausted to anticipate AQI for different urban communities also.

REFERENCES

[1] Ni, X.Y.; Huang, H.; Du, W.P. "Relevance analysis and short-term prediction of PM 2.5 concentrations in

Beijing based on multi-source data." Atmos. Environ. 2017, 150, 146-161.

[2] G. Corani and M. Scanagatta, "Air pollution prediction via multi-label classification," Environ. Model. Softw., vol. 80, pp. 259-264,2016.

[3] Mrs. A. GnanaSoundariMtech, (Phd), Mrs. J. GnanaJeslin M.E, (Phd), Akshaya A.C. "Indian Air Quality Prediction And Analysis Using Machine Learning". International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11, 2019 (Special Issue).

[4] Suhasini V. Kottur , Dr. S. S. Mantha. "An Integrated Model Using Artificial Neural Network

[5] RuchiRaturi, Dr. J.R. Prasad . "Recognition Of Future Air Quality Index Using Artificial Neural Network".International Research Journal ofEngineering and Technology (IRJET) .e-ISSN: 2395-0056 p-ISSN: 2395-0072 Volume: 05 Issue: 03 Mar-2018

[6] Aditya C R, Chandana R Deshmukh, Nayana D K, Praveen Gandhi Vidyavastu ." Detection and Prediction of Air Pollution using Machine Learning Models". International Journal o f Engineering Trends and Technology (IJETT) - volume 59 Issue 4 - May 2018

[7] Gaganjot Kaur Kang, Jerry ZeyuGao, Sen Chiao, Shengqiang Lu, and Gang Xie." Air Quality Prediction: Big Data and Machine Learning Approaches". International Journal o f Environmental Science and Development, Vol. 9, No. 1, January 2018

[8] PING-WEI SOH, JIA-WEI CHANG, AND JEN-WEI HUANG," Adaptive Deep Learning-Based Air Quality Prediction



Model Using the Most Relevant Spatial-Temporal Relations,” IEEE ACCESS July 30, 2018. Digital Object Identifier 10.1109/ACCESS.2018.2849820.

[9] Gaganjot Kaur Kang, Jerry Zeyu Gao, Sen Chiao, Shengqiang Lu, and Gang Xie, “Air Quality Prediction: Big Data and Machine Learning Approaches,” International Journal of Environmental Science and Development, Vol. 9, No. 1, January 2018.

[10] Haripriya Ayyala somayajula, Edgar Gabriel, Peggy Lindner and Daniel Price, “Air Quality Simulations using Big Data Programming Models,” IEEE Second International Conference on Big Data Computing Service and Applications, 2016.

FINDING PSYCHOLOGICAL INSTABILITY USING MACHINE LEARNING

Snc Soundaryavalli (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr.I.R.Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

As we know that people around the globe work hard to keep up with this racing world. However, due to this each individual is dealing with different health issues, one of the most known issue is depression or stress which may eventually lead to death or other brutal activities. These abnormalities can be termed as the Bipolar disorder which can be treated by undergoing some treatment suggested by specialists. For this research, data has been collected from working people which comprises of all kinds of questions for despondent detection and the dataset has been run through some machine learning algorithms. Random Forest algorithm gives the highest accuracy as 87.02% compared to the other algorithms.

1. INTRODUCTION

Mental health can influence everyday living, relations, and physical health. In any case, this connection additionally works the other way. Factors in individuals' lives, relational associations, and physical variables would all be able to add to mental health disturbances. Caring for mental issues can improve a person's perspective over life in a positive way.

Doing this can help in achieving harmony in life. Conditions, for example, stress, despondency, and nervousness would all be able to influence mental health and disturb an individual's everyday practice. Despite the fact that the term mental health is in like manner use, numerous conditions that specialists perceive as mental issue have physical roots. Modifiable variables for mental health issue include: financial conditions, such whether work is accessible in the neighborhood occupation a person's level of social consideration education

living quality Non-modifiable variables include: □ gender □ age Mental disorders impact around 25 percent of elders; just about 6 percent are truly disabled and named having real mental sickness.

These disorders are habitually associated with endless physical infirmities, for instance, coronary disease and diabetes. They in like manner increase the peril of physical injury and going through disasters, severity, and suicides. Suicide alone was at risk for 35,345 deaths in the U.S in 2019 (the latest year for which last data are available), making it the tenth driving explanation behind death. Among adolescents and young adults, suicide is responsible for extra deaths than the blend of harmful development, heart ailment, innate irregularities, respiratory disorder, influenza, , iron deficiency, and kidney and liver disease. The treatment of mental affliction has been held somewhere around the inclination that disorders of feeling, thinking, and direct somehow need realness and rather reflect particular weakness or poor life choices. Most crisis offices are sick prepared to address the issues of patients amidst mental health emergencies. Most protection plans see mental ailment and dependence as special cases to standard thought, not part of it. Regardless of a general social move towards sympathy, our overall population in spite of everything will when all is said in done view the mentally wiped out and those with propensity as morally broken instead of as wiped out.

2. EXISTING SYSTEM

Sridharan et al. presented the detection diagnostics on online social media with the assistance of Convolution Neural Networks (CNN) where accentuation was to get information posted by different clients while also ensuring algorithm protects the security with the assistance of separating agents which deal with information.

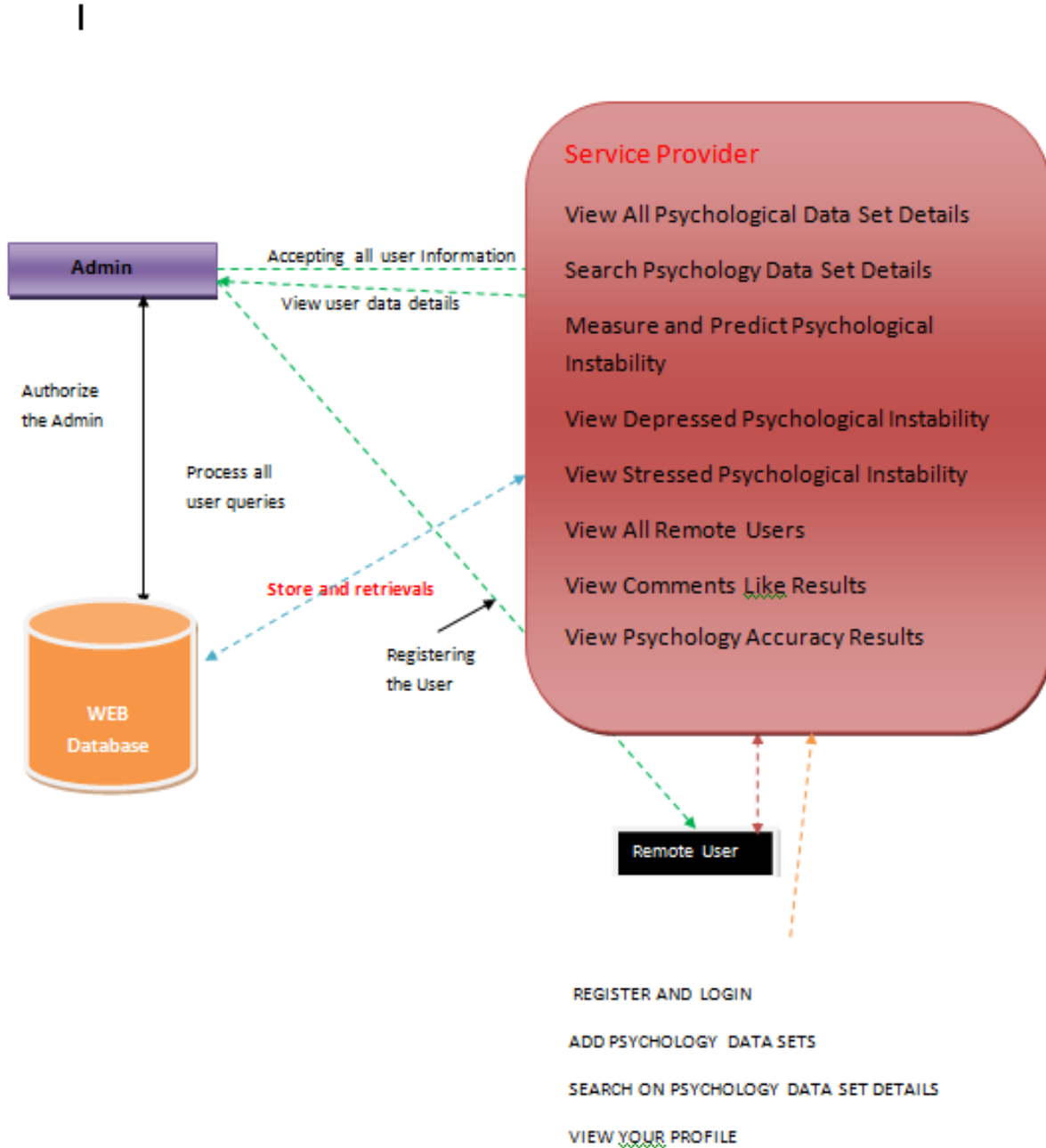
M N Stollar, M Lechh, S J Stollar, N B Allen approach utilizes an upgraded spectral move off parameters for detection of the depression side effects from discourse signals on the clinical dataset obtained. The classification of these highlights is done with the assistance of basic SVM classifier. In past investigation, gender dependence has improved depression classification either best for females, males and fluctuated amongst highlights. In this examination depression detection was more viable in males than females.

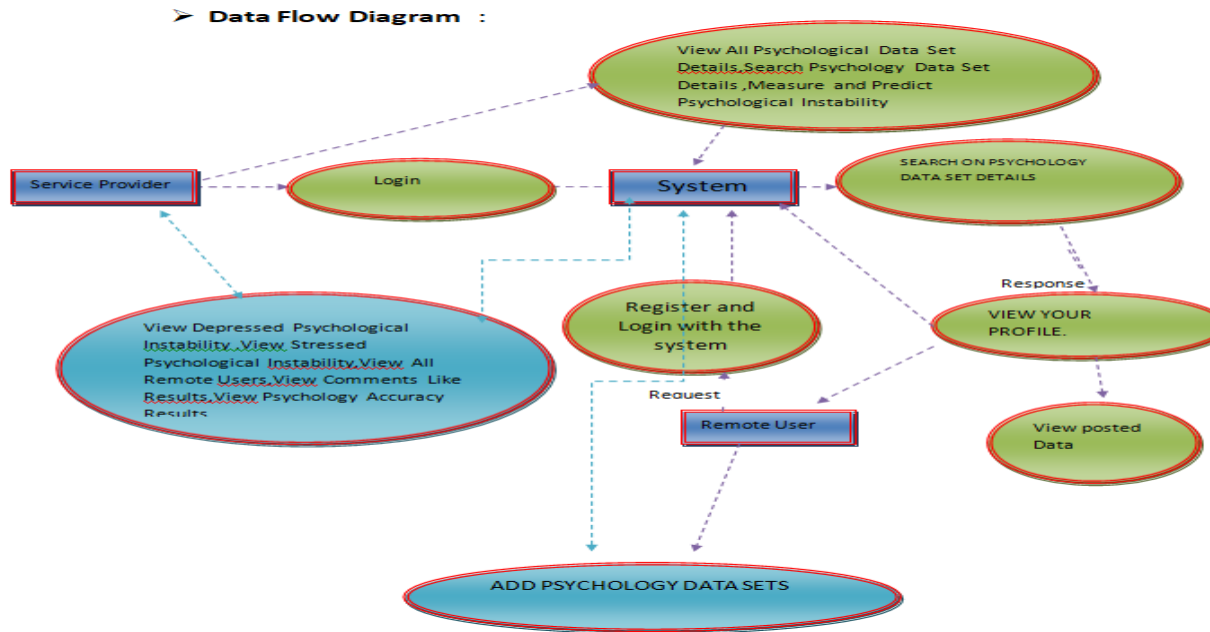
3. PROPOSED SYSTEM

The proposed system considers the stress detection among the tech people. The dataset considered is a survey among the working people, which considered all possible question for stress detection.

The designed approach utilizes the ML algorithm for stress identification; SVM, DT and Random forest are used on the dataset for learning and detection. The proposed approach finds the suitable algorithm for mental disorder prediction.

Architecture Diagram





PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly

use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform

Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

4. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.

The application starts running when it is executed for the first time. The server has to be started and then the internet explorer is used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

5. CONCLUSIONS

There are various methods which are utilized for detection of mental illness among individuals of various ages. The method utilized by these systems utilizes the method of detection via analyzing the mental issue detection through the set of questionnaires, in order to anticipate the downturn levels among various age groups. The machine learning algorithms are utilized for mental confusion detection. The dataset with 1200 samples are considered for study. We utilized SVM, Decision Tree and Random woodland for learning and detection. The experimental outcomes demonstrated that the Random Forest achieves the most elevated accuracy around 87%. In

future, we are intrigued to expand the work with some profound learning models, for example, Neural Networks or convolution neural networks.

6. REFERENCES

- [1] Mental Disorder Detection : Bipolar Disorder Scrutinization using Machine Learning, published in 2019.
- [2] Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm Azar, Ghassan & Gloster, Clay & El- Bathy, Naser & Yu, Su & Neela, Rajasree & Alothman, Israa. (2015). Intelligent data mining and machine learning for mental health diagnosis using genetic algorithm. 201-206. 10.1109/EIT.2015.7293425
- [3] A Framework for Classifying Online Mental Health-Related Communities With an Interest in Depression B. Saha, T. Nguyen, D. Phung and S. Venkatesh, "A Framework for Classifying Online Mental Health-Related Communities With an Interest in Depression," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no. 4, pp. 1008- 1015, July 2016.
- [4] Detecting Cognitive Distortions Through Machine Learning Text Analytics T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez and C. Giraud-Carrier, "Detecting Cognitive Distortions Through Machine Learning Text Analytics," 2017 IEEE International Conference on Healthcare Informatics (ICHI), Park City, UT, 2017, pp. 508-512.
- [5] Machine Learning Framework for the Detection of Mental Stress at Multiple Levels Subhani, Ahmad & Mumtaz, Wajid & MOHAMA SAAD, MOHAMAD NAUFAL & Kamel, Nidal & Malik, Aamir. (2017). Machine Learning Framework for the Detection of Mental Stress at Multiple Levels. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2723622.
- [6] Prediction of Mental Health Problems Among Children Using Machine Learning Techniques Sumathi, Ms & B., Dr. (2016). Prediction of Mental Health Problems Among Children Using Machine Learning Techniques. International Journal of Advanced Computer Science and Applications. 10.14569/IJACSA.2016.070176.

Implementation of Fruits Recognition Classifier using Convolutional Neural Network Algorithm for Observation of Accuracies for Various Hidden Layers

Talupuri Veera Venkata Siva Satya Prasad (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Dr. I. R. Krishnam Raju, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract—

Fruit recognition using Deep Convolutional Neural Network (CNN) is one of the most promising applications in computer vision. In recent times, deep learning based classifications are making it possible to recognize fruits from images. However, fruit recognition is still a problem for the stacked fruits on weighing scale because of the complexity and similarity. In this paper, a fruit recognition system using CNN is proposed. The proposed method uses deep learning techniques for the classification. We have used Fruits-360 dataset for the evaluation purpose. From the dataset, we have established a dataset which contains 17,823 images from 25 different categories. The images are divided into training and test dataset. Moreover, for the classification accuracies, we have used various combinations of hidden layer and epochs for different cases and made a comparison between them. The overall performance losses of the network for different cases also observed. Finally, we have achieved the best test accuracy of 100% and a training accuracy of 99.79%.

Keywords—Fruit Recognition, Convolutional Neural Network (CNN), Fruits-360 dataset, Adam optimizer, Cost function, Hidden layers and epochs

1. INTRODUCTION

With the lively improvement of our human society, additional attention has been paid to the superiority of our life, particularly the food we eat. Over the last few years, computer visions have been widely used in fruit recognition methods. In the field of image recognition and classification, Deep Neural Network (DNN) is used to identify fruits from images. DNN

performs better than other machine learning algorithms. Convolutional Neural Networks (CNNs) are classified as a deep learning algorithm. In deep learning, CNN [1, 2] are the most commonly used type of Artificial Neural Networks (ANNs). It is being used several visual recognition analyzing which includes video and image recognition [3], face recognition [4], handwritten digit recognition [5], and fruit recognition [6] etc. The accuracies in these fields including fruit recognition using CNN have reached human-level perfection. Mammalian visual systems' biological model is the one by which the architecture of the CNN is inspired. It was found by D. H. Hubel et al. in 1962 that the cells in the cat's visual cortex are refined to a minute area of the visual field which is recognized as the receptive field [7]. In 1980, the neocognitron [8] introduced by Fukushima was the pattern recognition model inspired by the work of D. H. Hubel et al. [9] was the first computer vision. However, CNN is categorized by a network architecture which consists of convolution and pooling layers to extract and combine high-level features from 2D input.

CNN has a very similar architecture as ANN. There are several neurons in each layer in ANN. Hence, the weighted sum of all the neurons of a layer becomes the input of a neuron of the next layer adding a biased value. In CNN the layer has three dimensions. Here all the neurons are not fully connected instead they are connected to the local receptive field. A cost function is generated in order to train the network. It compares the network's output with the desired output. Accurate and efficient fruit recognition is of great importance in the field of robotic harvesting and yield mapping. An ideal fruit recognition system is accurate that can be trained on an easily available dataset, shows real-time predictions and acclimates various types of fruits. Therefore, in our research, we implement a fruit recognition classifier using CNN. The input image is taken as 100×100 pixels of RGB image. For the networks best performance, we used various combinations of hidden layers for five cases and observe the accuracies. The final experiment result shows the much-improved fruit recognition rate. The mathematical model of the network is executed in python with tensor flow.

2. PREVIOUS WORK

A number of factors made fruit recognition a challenging task which includes fruits that arise in scenes of fluctuating brightness, obstructed by other objects, sharpen edge, texture, reflectance properties etc. Many kinds of research exist to help fruit recognition challenges. Fruit recognition can be considered as an image segmentation problem. Several works are available in the

literature addressing the problem of fruit recognition as an image segmentation problem. Wang et al. [10] established a system that detects apples based on their color. They surveyed the issue of apple detection for yield prediction. Hung et al. [11] proposed a five-class segmentation method for almond segmentation using conditional random fields. This method learned features using a Sparse Autoencoder (SAE). Later, these features were used within the CRF framework and shown remarkable segmentation performance. A novel approach for detecting fruit using the deep convolutional neural network is presented in paper [6]. In this paper, the author adapts a Faster Region-based CNN through transfer learning. They trained the model using RGB and NIR (Near-Infrared) images. The combination of RGB and NIR discovers the early and late fusion methods. Fruit recognition method based on deep fruit system [6] achieved a remarkable milestone in the development of deep learning approaches for fruit detection. In another research [12], where the neural networks trained by two backpropagation algorithms on images of Apple Gala variety trees in order to calculate the yield for the forthcoming seasons. Furthermore, detection in relation with the angle of the camera [13], Scale Invariant Feature Transform (SIFT), improved ChanVese level-set model [14] are also used in the literature for the fruit recognition.

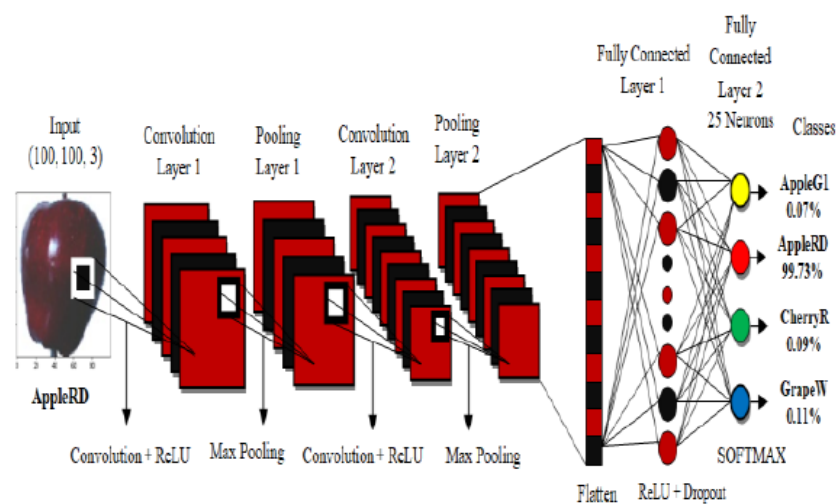


Fig. 1. Schematic of the architecture of a convolutional neural network

3. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is

not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

3.1.1 ECONOMICAL FEASIBILITY:

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

3.1.2 TECHNICAL FEASIBILITY:

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

3.1.3 SOCIAL FEASIBILITY:

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the

users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5. SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

6. CONCLUSION

This paper explores a fruits recognition classifier based on CNN algorithm. The accuracy and loss curves were generated by using various combinations of hidden layers for five cases using fruits-360 dataset. The recognition rate has greatly improved throughout the experiment. Among all the cases, the model achieved the best test accuracy of 100% in case 4 from 11 to 15 epochs and best training accuracy of 99.79% in case 1 at epoch 15. This type of higher accuracy will cooperate to stimulate the overall performance of the machine more adequately in fruits recognition. On the contrary, the highest and the lowest performance loss were found without and with the presence of dropout. The highest loss is approximately 0.5881 found in case 2 without dropout. Besides, the lowest performance loss is approximately 0.0032 in the presence of conv1, pool1, conv2, pool2 with dropout. This low loss will provide CNN better performance to attain better fruit recognition. In the future, our plan is to perform the segmentation process on the image before recognition and then applying it on CNN.

7. REFERENCES

- [1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, pp. 541-551, 1989.

- [2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.
- [3] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725-1732.
- [4] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988-1996.
- [5] F. Siddique, S. Sakib, and M. A. B. Siddique, "Handwritten Digit Recognition using Convolutional Neural Network in Python with Tensorflow and Observe the Variation of Accuracies for Various Hidden Layers," 2019.
- [6] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, p. 1222, 2016.
- [7] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, pp. 106-154, 1962.
- [8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, p. 436, 2015.
- [9] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, ed: Springer, 1982, pp. 267-285.
- [10] Q. Wang, S. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *Experimental robotics*, 2013, pp. 745-758.
- [11] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 5314-5320.

- [12] H. Cheng, L. Damerow, Y. Sun, and M. Blanke, "Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks," *Journal of Imaging*, vol. 3, p. 6, 2017.
- [13] J. Hemming, J. Ruizendaal, J. Hofstee, and E. van Henten, "Fruit detectability analysis for different camera positions in sweet-pepper," *Sensors*, vol. 14, pp. 6032-6044, 2014.
- [14] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, pp. 266-277, 2001.
- [15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [16] M. A. Nielsen, *Neural networks and deep learning* vol. 25: Determination press USA, 2015.
- [17] T. Schaul, S. Zhang, and Y. LeCun, "No more pesky learning rates," in *International Conference on Machine Learning*, 2013, pp. 343-351.

[View publication](#)

PHISHING WEB SITES FEATURES

CLASSIFICATION ON MACHINE LEARNING

Tangella Ravi Kiran (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT:

Phishing are one of the most common and most dangerous attacks among cybercrimes. The aim of these attacks is to steal the information used by individuals and organizations to conduct transactions. Phishing websites contain various hints among their contents and web browser-based information. The purpose of this study is to perform Extreme Learning Machine (ELM) based classification for 30 features including Phishing Websites Data in UC Irvine Machine Learning Repository database. For results assessment, ELM was compared with other machine learning methods such as Support Vector Machine (SVM), Naïve Bayes (NB) and detected to have the highest accuracy of 95.34%

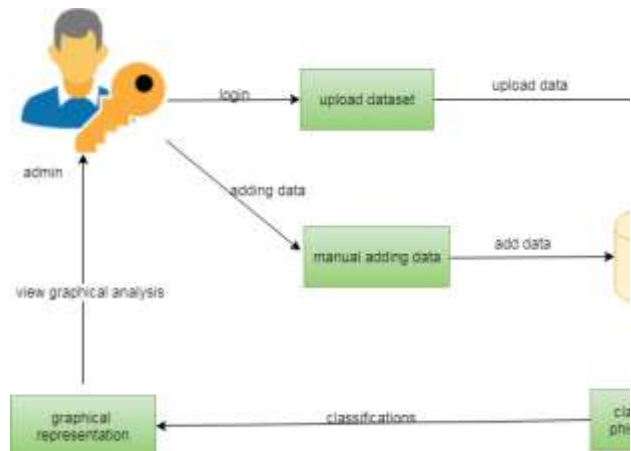
1. INTRODUCTION

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints among their contents and web browser-based information. Individual(s) committing the fraud sends the fake website or e-mail information to the target address as if it

comes from an organization, bank or any other reliable source that performs reliable transactions. Many articles have been published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below. In the first section we defined rules and we gave equations of web features. We need these equations in order to explain phishing attacks characterization. In this study, features in the database created for phishing websites are classified by determining the input and output parameters for the ELM classifier. Results obtained by ELM show that ELM has higher achievement compared to other classifier (SVM and NB) methods. This study is considered to be an applicable design in automated systems with high

performing classification against the phishing activity of websites.

2. ARCHITECTURE



3. EXISTING SYSTEM

Internet use has become an essential part of our daily activities as a result of rapidly growing technology. Due to this rapid growth of technology and intensive use of digital systems, data security of these systems has gained great importance. The primary objective of maintaining security in information technologies is to ensure that necessary precautions are taken against threats and dangers likely to be faced by users during the use of these technologies. Phishing is defined as imitating reliable websites in order to obtain the proprietary information entered into websites every day for various purposes, such as usernames, passwords and citizenship numbers. Phishing websites contain various hints among their contents and web browser-based information. Individual(s) committing the fraud sends the fake website or e-mail information to the target address as if it comes from an organization, bank or any other reliable source that performs reliable transactions. Many articles have been

published about how to predict the phishing websites by using artificial intelligence techniques. We examined phishing websites and extracted features of these web sites. Guidelines regarding the extracted features of this database are given below. In the first section we defined rules and we gave equations of web features. We need these equations in order to explain phishing attacks characterization.

4. PROPOSED SYSTEM

In this study, features in the database created for phishing websites are classified by determining the input and output parameters for the ELM classifier. Results obtained by ELM show that ELM has higher achievement compared to other classifier (SVM and NB) methods. This study is considered to be an applicable design in automated systems with high performing classification against the phishing activity of websites. Furthermore, in literature comparisons, this study is observed to be high-performing by having a high performance of 92.18% that is also the highest test performance in the publication. The topics addressed in this section are the two measures that affect the performance of the model and the algorithm used, the first one being the division of data set into training and test data set and the second one being the definition of expressions measuring the performance. In the first measure, the data set is divided into three parts as training, validation and test data by three-phase division in K-Fold method, and model selection and performance status are simultaneously performed. In the second measure, performance assessment of



classifier models generally uses a validation value. phishing attacks are vishing, smishing, search engine phishing, spear phishing, whaling attacks.

5. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are,

- ◆ **ECONOMICAL FEASIBILITY**
- ◆ **TECHNICAL FEASIBILITY**
- ◆ **SOCIAL FEASIBILITY**

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will

lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TEST

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TYPES OF TESTS

Unit testing



Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures : interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

White Box Testing

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at



least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

Black Box Testing

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format

- No duplicate entries should be allowed
- All links should take the user to the correct page.

Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g. components in a software system or – one step up – software applications at the company level – interact without error.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

6. CONCLUSION

In this paper, we defined features of phishing attack and we proposed a classification model in order to classification of the phishing attacks. This method consists of feature extraction from websites and classification section. In the feature extraction, we have clearly defined rules of phishing feature extraction and these rules have been used for obtaining features. In



order to classification of these feature, SVM, NB and ELM were used. In the ELM, 6 different activation functions were used and ELM achieved highest accuracy score.

7. REFERENCES

[1] AO Kaspersky lab. (2017). The Dangers of Phishing: Help employees avoid the lure of cybercrime. [Online] Available: <https://go.kaspersky.com/Dangers-Phishing-Landing-Page-Soc.html> [Oct 30, 2017].

[2] "Financial threats in 2016: Every Second Phishing Attack Aims to Steal Your Money" Internet:

<https://www.kaspersky.com/about/pressreleases/2017-financial-threats-in-2016>. Feb 22, 2017 [Oct 30, 2017].

[3] Y. Zhang, J. I. Hong, and L. F. Cranor, "Cantina: A Content-based Approach to Detecting Phishing Web Sites," New York, NY, USA, 2007, pp. 639-648.

[4] M. Blasi, "Techniques for detecting zero day phishing websites." M.A. thesis, Iowa State University, USA, 2009.

[5] R. S. Rao and S. T. Ali, "PhishShield: A Desktop Application to Detect Phishing Webpages through Heuristic Approach," *Procedia Computer Science*, vol. 54, no. Supplement C, pp. 147-156, 2015.

[6] E. Jakobsson, and E. Myers, *Phishing and Counter-Measures: Understanding the Increasing Problem of Electronic Identity Theft*. Wiley, 2006, pp.2-3.

[7] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, "Detecting phishing web sites: A heuristic URL-based approach," in 2013 International Conference on Advanced Technologies for Communications (ATC 2013), 2013, pp. 597-602.

[8] Z. Zhang, Q. He, and B. Wang, "A Novel Multi-Layer Heuristic Model for Anti-Phishing," New York, NY, USA, 2017, p. 21:1-21:6.

[9] N. Sanglerdsinlapachai and A. Rungsawang, "Web Phishing Detection Using Classifier Ensemble," New York, NY, USA, 2010, pp. 210-215.

[10] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, "CANTINA+: A Feature- Rich Machine Learning Framework for Detecting Phishing Web Sites," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 21:1-21:28, Sep. 2011.

[11] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Predicting phishing websites based on self-structuring neural network," *Neural Comput & Applic*, vol. 25, no. 2, pp. 443-458, Aug. 2014.

[12] Pradeepthi K V and Kannan A, "Performance study of classification techniques for phishing URL detection," in 2014 Sixth International Conference on Advanced Computing (ICoAC), 2014, pp. 135-139.

Prediction of Modernized Loan Approval System Based on Machine Learning Approach

Turu Jyothi (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S.K.Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Technology has boosted the existence of humankind the quality of life they live. Every day we are planning to create something new and different. We have a solution for every other problem we have machines to support our lives and make us somewhat complete in the banking sector candidate gets proofs/ backup before approval of the loan amount. The application approved or not approved depends upon the historical data of the candidate by the system. Every day lots of people applying for the loan in the banking sector but Bank would have limited funds. In this case, the right prediction would be very beneficial using some classes-function algorithm. An example the logistic regression, random forest classifier, support vector machine classifier, etc. A Bank's profit and loss depend on the amount of the loans that is whether the Client or customer is paying back the loan. Recovery of loans is the most important for the banking sector. The improvement process plays an important role in the banking sector. The historical data of candidates was used to build a machine learning model using different classification algorithms. The main objective of this paper is to predict whether a new applicant granted the loan or not using machine learning models trained on the historical data set.

1. INTRODUCTION

Prediction of modernized loan approval system based on machine learning approach is a loan approval system from where we can know whether the loan will pass or not. In this system, we take some data from the user like his monthly income, marriage status, loan amount, loan duration, etc. Then the bank will decide according to its parameters whether the client will get the loan or not.

So there is a classification system, in this system, a training set is employed to make the model and the classifier may classify the data items into their appropriate class. A test dataset is created that trains the data and gives the appropriate result that, is the client potential and can repay the loan.

Prediction of a modernized loan approval system is incredibly helpful for banks and also the clients. This system checks the candidate on his priority basis. Customer can submit his application directly to the bank so the bank will do the whole process, no third party or stockholder will interfere in it. And finally, the bank will decide that the candidate is deserving or not on its priority basis. The only object of this research paper is that the deserving candidate gets straight forward and quick results.

2. EXISTING SYSTEM

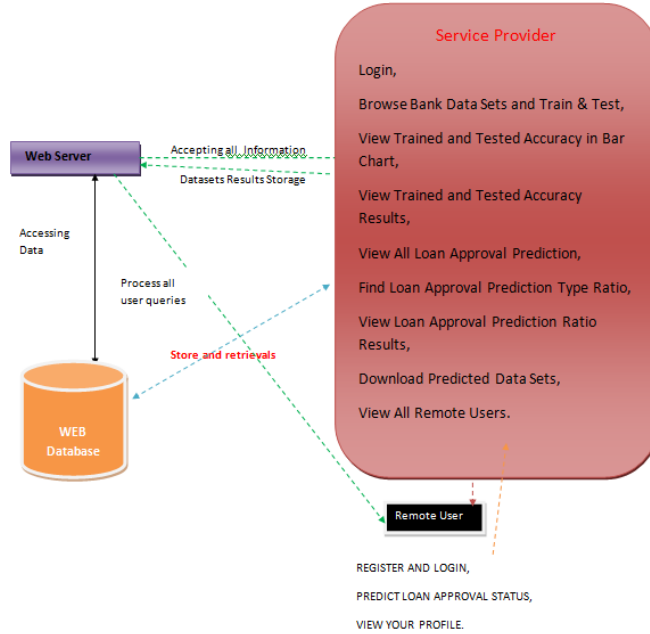
Loan approval is a very important process for banking organizations. Banking Industry always needs a more accurate predictive modeling system for many issues. Predicting credit defaulters is a difficult task for the banking industry. The system approved or rejects the loan applications. Recovery of loans is a major contributing parameter in the financial statements of a bank. It is very difficult to predict the possibility of payment of loan by the customer. Machine Learning (ML) techniques are very useful in predicting outcomes for large amount of data. In the proposed system, three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are applied to predict the loan approval of customers. The experimental results conclude that the accuracy of Decision Tree machine learning algorithm is better as compared to Logistic Regression and Random Forest machine learning approaches.

3. PROPOSED SYSTEM

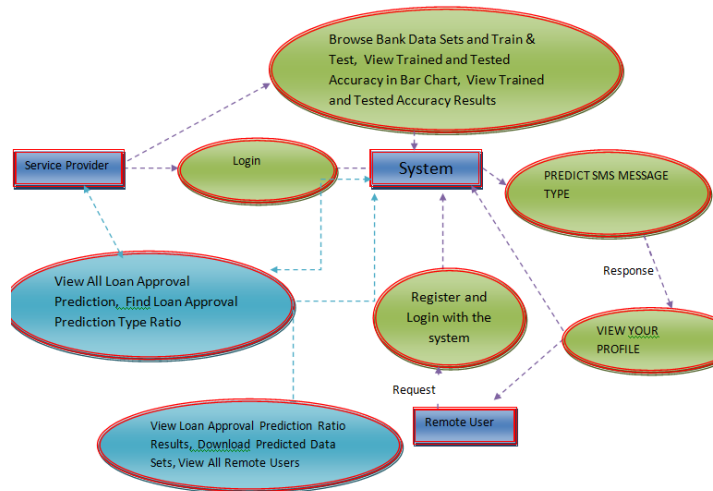
- ❖ This proposed model will characterize the behavior of customers on the Basis of their record. These records is taken from the customers, and create a data set. With the help of These data sets and training machine learning model, we predict that the customer's loan will passed or not.
- ❖ This Machine algorithms predict the possibility of a customer would be able to repay the loan or not and In this, we are going to discuss the advantage of loan prediction. In this system, we are going to predict that the person who is applying for a loan can repay or

not. If the client can repay then we predict that yes, eligible for a loan. And if the candidate fails then we predict that client is not eligible.

Architecture Diagram



Data Flow Diagram :



4. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also

including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

5. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

OUTPUT DESIGN

The Output from the computer is required to mainly create an efficient method of communication within the company primarily among the project leader and his team members, in other words, the administrator and the clients. The output of VPN is the system which allows the project leader to manage his clients in terms of creating new clients and assigning new projects to them, maintaining a record of the project validity and providing folder level access to each client on the user side depending on the projects allotted to him. After completion of a project, a new project may be assigned to the client. User authentication procedures are maintained at the initial stages itself. A new user may be created by the administrator himself or a user can himself register as a new user but the task of assigning projects and validating a new user rests with the administrator only.

The application starts running when it is executed for the first time. The server has to be started and then the internet explorer is used as the browser. The project will run on the local area network so the server machine will serve as the administrator while the other connected systems

can act as the clients. The developed system is highly user friendly and can be easily understood by anyone using it even for the first time.

6. CONCLUSIONS

According to this research paper prediction accuracy is sweet for both datasets. In some situations like client going through some disaster so here the algorithm cannot predict the appropriate result. This research paper can find out the client is potential and repay the loan and the accuracy is good. loan duration, loan amount, age, income are the most important factors for finding out there (whether the client would have been). 'zip code' and 'credit history' are the foremost important factors for predicting the category of the loan Applicant.

7. REFERENCES

- [1] Amruta S. Aphale and R. Prof. Dr. Sandeep. R Shinde, "Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval", International Journal of Engineering Trends and Applications (IJETA), vol. 9, issue 8, 2020)
- [2] Loan Prediction Using Ensemble Technique, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016
- [3] Exploratory data analysis https://en.wikipedia.org/wiki/Exploratory_data_analysis
- [4] Pandas Library <https://pandas.pydata.org/pandas-docs/stable/>
- [5] MeanDecreaseAccuracy <https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html>

NORMALIZATION OF DUPLICATE RECORDS FROM MULTIPLE SOURCES

VANAPALLI SANTHI

PG Scholar, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India.

S.K. ALISHA

Associate Professor, Dept of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India.

Abstract:

Data consolidation is a challenging issue in data integration. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The promise of Big Data hinges upon addressing several big data integration challenges, such as record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there is limited work on creating a uniform, standard record from a group of records corresponding to the same real-world entity. We refer to this task as record normalization. Such a record representation, coined normalized record, is important for both front-end and back-end applications. In this paper, we formalize the record normalization problem, present in-depth analysis of normalization granularity levels (e.g., record, field, and value-component) and of normalization forms (e.g., typical versus complete). We propose a comprehensive framework for computing the normalized record. The proposed framework includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record. We conducted extensive empirical studies with all the proposed methods. We indicate the weaknesses and strengths of each of them and recommend the ones to be used in practice.

Keywords : Data integration, Standards, Task analysis, Databases, Google, Data mining, Terminology

1.INTRODUCTION

1.1 Introduction:

The usefulness of Web data increases exponentially (e.g., building knowledge bases, Web-scale data analytics) when it is linked across numerous sources. Structured data on the Web resides in Web databases and Web tables. Web data integration is an important component of many applications collecting data from Web databases, such as Web data warehousing (e.g., Google and Bing Shopping; Google Scholar), data aggregation (e.g., product and service reviews), and met searching.

Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record matching problem and the truth discovery problem. The record matching problem is also referred to as duplicate record detection, record linkage, object identification, entity resolution, or de-duplication and the truth discovery problem is also called as truth

finding or fact finding - a key problem in data fusion.

This work assumes that the tasks of record matching and truth discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. It calls the generated record the normalized record. It call the problem of computing the normalized record for a group of matching records the record normalization problem (RNP), and it is the focus of this work.

RNP is another specific interesting problem in data fusion. Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or (ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user,

because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) it run the risk of presenting a record with missing or incorrect pieces of data. Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data.

This work aims to develop a framework for constructing normalized records systematically. This work includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record.

1.2 Purpose:

Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data. For example, Table 1

contains four records corresponding to the same entity (publication). They are extracted from different websites. Record Rnorm is constructed by hand for illustration purposes. One notices that the same publication has different representations in different websites.

1.3 Scope:

Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the record matching problem and the truth discovery problem. The record matching problem is also referred to as duplicate record detection, record linkage , object identification, entity resolution ,or deduplication and the truth discovery problem is also called as discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. We call the generated record the normalized record. We call the problem of computing the normalized record for a group of matching records the record normalization problem

(RNP), and it is the focus of this work. RNP is another specific interesting problem in data fusion.

1.4 Motivation:

Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or(ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) we run the risk of presenting a record with missing or incorrect pieces of data.

1.5 Overview:

We identify three levels of normalization granularity: record, field, and value-component. Record level assumes that the values of the fields within a record are governed by some hidden criterion and that together create a cohesive unit that is user-

friendly. As a consequence, this normalization favors building the normalized record from entire records among the set of matching records rather than piecing it together from field values of different records. Thus, any of the matching records (ideally, that has no missing values) can be the normalized record. Using our running example in Table 1, the record R_c is a possible choice for the normalized record with this level of normalization granularity. Field level assumes that record level is often inadequate in practice because records contain fields with incomplete values. Recall that these records are the products of automatic data extraction tools, which are not perfect and thus may produce errors [18]. This normalization level ignores the cohesion factor in the record normalization level and assumes that a user is better served when each field of the normalized record has as easy to understand a value as possible, selected from among the values in the set of matching records.

2. RELATED WORK

Sanghyeon Baeg [1] 2008, Power consumption is the most critical issue for low-power ternary content-addressable memory (TCAM) in match lines designs. In the proposed match-line architecture, the match line present in each TCAM word is

partitioned into four segments and is selectively pre-charged to reduce the match-line power consumption. The match lines which are partially charged are evaluated to determine the final comparison result by sharing the charges deposited in various parts of the partitioned segments.

B. Heller et al, [2] 2010, Built ElasticTree, which through data-center-wide traffic management and control, introduces energy proportionality in today's non-energy proportional networks. They will likely essentially decrease this quickly developing vitality cost. Compare multiple strategies for finding the minimum-power network [20]. The framework is vitality proficiency, best execution, and adaptation to non-critical failure. The system worked near its ability will build the possibility of dropped and postponed bundles.

A.R. Curtis et al, [3] 2011, DevoFlow proposition enables administrators to target just the streams that issue for their administration issue. DevoFlow handles most miniaturized scale streams in the information plane and consequently enables us to make the most out of switch resources. DevoFlow takes care of the issue by permitting a clonable trump card principle to choose a

yield port. Multipath steering to statically stack balance movement with no utilization of the control-plane. These procedures don't spare much vitality on elite systems.

P. Porraset al, [4] 2012, Incorporates several critical components that are necessary for enabling security applications in Open Flow networks including role-based authorization, rule reduction, conflict evaluation, and policy synchronization. FortNOX is a critical initial phase in enhancing the security of Open Flow systems. It shows the achievability and suitability of our nom de plume set guideline decrease approach [18]. It is unable to handle the dynamic matching process.

Zahid Ullah et al, [5] 2012, Hybrid partitioned static random is a memory architecture in which access memory-based ternary content addressable memory (HP SRAM-based TCAM), which involves TCAM functionality with conventional SRAM, where we are eliminating the inherited disadvantages of conventional TCAMs. HP SRAM-based TCAM is a technique in which they logically dissect conventional TCAM table in a hybrid way (column-wise and row-wise) into TCAM sub-tables, which are then processed to be

mapped to their corresponding SRAM memory units.

H. Kim and N. Feamster et al, [6] 2013, Designed and implemented Procera, an event-driven network control framework based on SDN. Additionally, utilize the OpenFlow convention to impart between the Procera controller and the hidden system switches. It gives better permeability and command over undertakings for performing system. This SDN can improve common network management tasks [19]. Procera experiences the characteristic deferral caused by the communication of the control plane and the information plane.

M. Yu, L. Jose et al, [7] 2013, OpenSketch empowers a straightforward and proficient approach to gather estimation information. It utilizes information plane estimation natives dependent on ware switches and an adaptable control plane so administrators can without much of a stretch execute variable estimation calculations. It has a simple, efficient way to control switches [16]. Sketches more flexible in supporting various measurement tasks. Delay of each measurement pipeline component is large.

Weirong Jiang et al, [8] 2013, Random access memory i.e. (RAM)-based

Ternary Content Addressable Memory i.e.(TCAM) architecture is design for efficient implementation on state-of-the-art FPGAs. We give a formal study on RAM-based TCAM to disclose the ideas and the algorithms behind it. To face the timing challenge, we propose a modular architecture consisting of arrays of small-size RAM-based TCAM units.

Jacobson et al, [9] 2014, Novel control plane architecture called OpenNF that addresses these challenges through careful API design. OpenNF enables applications to settle on reasonable decisions in meeting their destinations. NF software is always Up-to-Date. The system has High performance on network monitoring.

M. Moshref et al, [10] 2014, DREAM enables operators and cloud tenants to flexibly specify their measurement tasks in a network and dynamically allocates TCAM resources to these tasks based on the resource-accuracy. User-specified high level of accuracy. DREAM can support more concurrent tasks. DREAM needs to dismiss almost half of the assignments and drop about 10%.

N. Katta et al, [11] 2014, CacheFlow system is a system which “caches” the most popular rules in the small TCAM, in which

they are relying on software to handle the small amount of “cache miss” traffic. But, we cannot blindly apply existing cache-replacement algorithms, because of dependencies between rules with overlapping patterns.

Naga Katta et al, [12] 2014, Instead of creating long dependency chains to cache smaller groups of rules in which semantics of the network policy are preserved. There are mainly four types of criteria for it. Elasticity which combines the best of hardware and software switches. Transparency which faith-fully supporting native OpenFlow semantics, including traffic counters. Fine-grained rule caching which places popular rules in the TCAM, despite dependencies on less-popular rules. Adaptability which enables incremental changes to the rule caching as the policy changes.

3. EXISTING SYSTEM

In existing, TCAM Razor, DomainFlow and Palette algorithms are used. Which is both power hungry and highly limited in capacity . Most TCAM-capable commodity switches support only a few thousand wildcard entries. Although certain products recently reported an ability to support up to 125k wildcard entries, enlarging the memory with

enhanced control capability significantly increases the cost To improve scalability, two approaches have been taken: proactively allocating rules on multiple switches to load balance the memory consumption , and reactively caching rules on each switch individually.

3.1 Disadvantages: These existing works provides poor caching ratio and less hit ratio. These existing works provides poor caching ratio and less hit ratio

4. PROPOSED SYSTEM

To deal with existing disadvantages, this work proposed a novel wildcard-rule caching algorithm and a cache replacement algorithm to make use of TCAM space efficiently. TCAM can look up a packet’s header and compare the matching patterns of the packet to the match field of all rules in the flow table in parallel. Our wildcard-rule caching algorithm repeats caching a set of important rules into TCAM until there is no TCAM space. Our cache replacement algorithm takes temporal and spatial traffic localities into consideration, which could make hit ratio high.

4.1 Advantages

The proposed wildcard-rule caching algorithm could have better caching ability than the other existing algorithms.

Furthermore, the proposed cache replacement algorithm could have higher hit ratio than the other existing algorithms.

5. IMPLEMENTATION

5.1 Load conference name dataset:

This module load conference name dataset. This dataset contain rid, label and conference name. This dataset contains 3683 records.

5.2 Mining Abbreviation Definition pairs:

This module use a number of heuristics to determine whether given two value components s and t , s is an abbreviation of t . In this section, a value component is a word (or term). As we mentioned previously, in this module we consider only fields with the string data type. We define the neighboring context of a word w within the set of values of a field f_j as the set of pairs (left neighbor word, right neighbor word) with the property that the substring left neighbor word w right neighbor word is a substring of a value of f_j in some record in Re . If w is the beginning word of a field value, we use a special start-symbol “< s >” to mark left neighbor word. If it is the last word in the field value, we use the special end-symbol “</ s >” to mark right neighbor word. For example, the words “proceedings” and

“proc” occur many times in the field venue, and they share a good fraction of their neighboring contexts, such as (in, of), (< s >, of), (in, acm). “proc” is also the prefix of “proceedings”, so we become increasingly confident that “proc” is a possible abbreviation of “proceedings”.

5.3 Mining Template Collocation-SubCollocation Pairs (MTS):

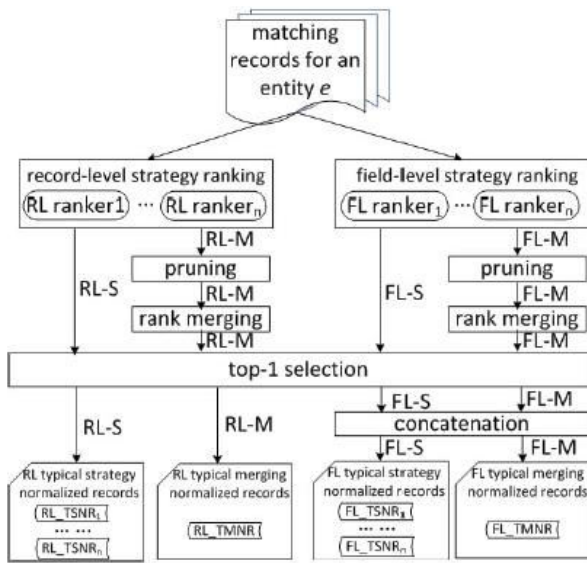
This module aim to find all template collocations and their subcollocations. The template collocations become the candidates with which it can expand (replace) the subcollocations. They will be used to generate the normalized component values for a field. Let an n -collocation tc be a template collocation and a k collocation kc be its subcollocation ($k < n$).

5.4 Mining Most Frequently Co-occurring Template Collocation:

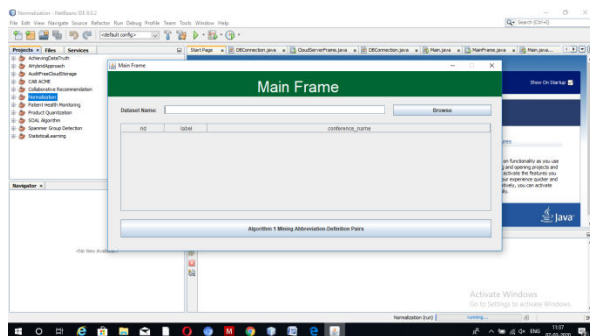
The above module, discussed how to obtain the template collocations and their corresponding subcollocations. We notice that some of the template collocations co-occur frequently. For example, among the values of the field venue, the template collocation “conference on” co-occurs most frequently with “in proceedings of the.” We also observe that template collocation co-

occurrence is not always bidirectional. For example, the template collocation “symposium on” co-occurs most often with “in proceedings of the”, but “in proceedings of the” co-occurs most frequently with “conference on.”

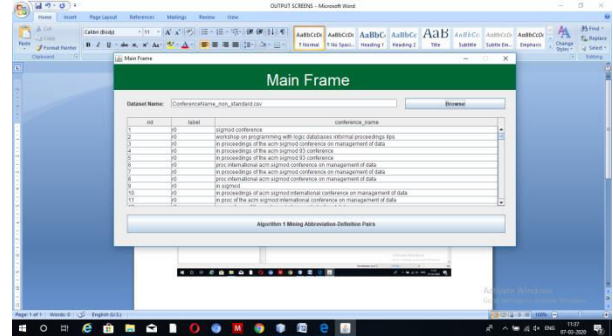
6. Architecture



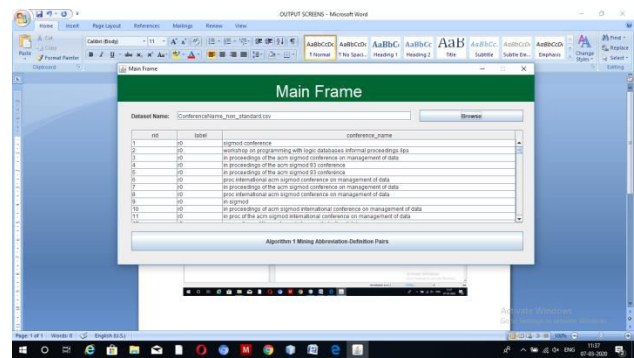
7. OUTPUT RESULTS



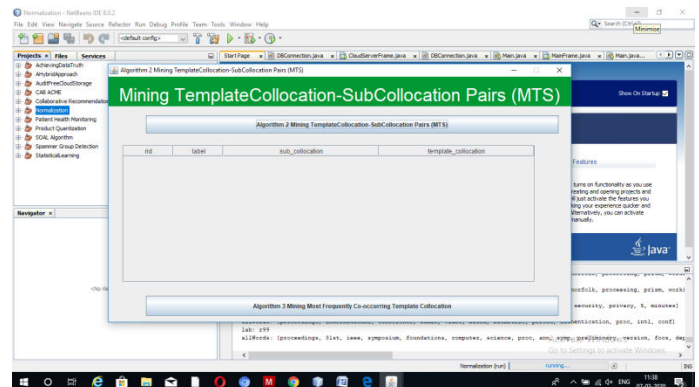
Load Dataset Screen



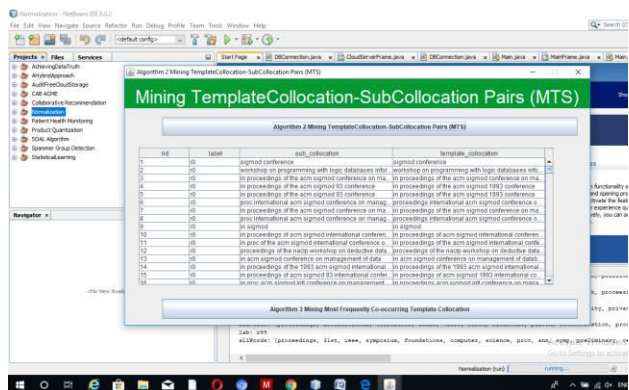
Show Dataset screen



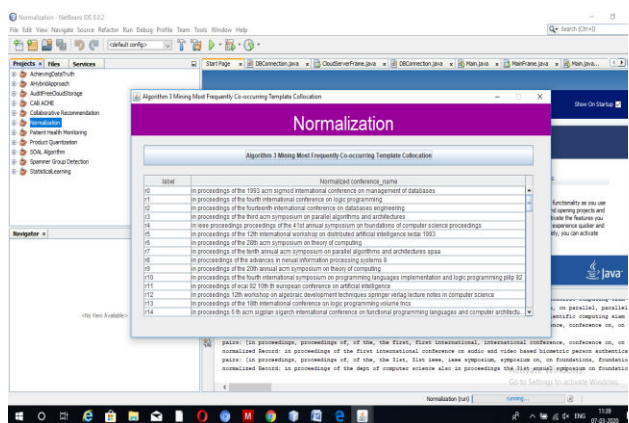
Implement Algorithm-1



Implement Algorithm-2



Minig Template collection-sub collection pairs screen



Implement Algorithm-3

8. CONCLUSION AND FUTURE ENHANCEMENT

This work studied the problem of record normalization over a set of matching records that refer to the same real-world entity. This work presented three levels of normalization granularities (record-level, field-level and value component level) and two forms of normalization (typical normalization and complete normalization). For each form of

normalization, this work proposed a computational framework that includes both single-strategy and multi-strategy approaches. This work proposed four single-strategy approaches: frequency, length, centroid, and feature-based to select the normalized record or the normalized field value. For multistrategy approach, this work used result merging models inspired from metasearching to combine the results from a number of single strategies. This work analyzed the record and field level normalization in the typical normalization. In the complete normalization, this work focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values. This work implemented a prototype and tested it on a real-world dataset. The experimental results demonstrate the feasibility and effectiveness of this approach. This method outperforms the state-of-the-art by a significant margin.

9. BIBLIOGRAPHY

[1] A. Culotta, M. Wick, R. Hall, M. Marzilli, and A. McCallum, "Canonicalization of database records using adaptive similarity measures," in SIGKDD, 2007, pp. 201–209.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: A generic approach to entity resolution," *VLDBJ*, vol. 18, no. 1, pp. 255–276, 2009.

[3] M. L. Wick, K. Rohanimanesh, K. Schultz, and A. McCallum, "A unified approach for schema matching, coreference and canonicalization," in *SIGKDD*, 2008, pp. 722–730.

[4] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," *Inf. Sys.*, vol. 26, no. 8, pp. 607–633, 2001.

[5] L. Wang, R. Zhang, C. Sha, X. He, and A. Zhou, "A hybrid framework for product normalization in online shopping," in *DASFAA*, vol. 7826, 2013, pp. 370–384.

[6] S. Chaturvedi and et al., "Automating pattern discovery for rule based data standardization systems," in *ICDE*, 2013, pp. 1231–1241.

[7] E. C. Dragut, C. Yu, and W. Meng, "Meaningful labeling of integrated query interfaces," in *VLDB*, 2006, pp. 679–690.

[8] S. Raunich and E. Rahm, "Atom: Automatic target-driven ontology merging," in *ICDE*, 2011, pp. 1276–1279.

QR CODE BASED SMART ATTENDANCE SYSTEM

Vardhanapu Subhakar Rao (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract - In this era of technology smartphones play a significant role in our day to day life. Nowadays smartphones can solve most of the problem very quickly and easily. It has made life of every person simple and easier with different social app, commercial app, problem solving app, app for education and marketing etc. Followed by the technology the paper purposed a system that will handle a problem for recording the attendance. The proposed system is a couple of two applications, one for generating the QR Code by entering the student details and second application for taking the attendance and generating the attendance in CSV or XLS format. The teacher will need to scan the QR code of the particular student in order to confirm their attendance. The paper discusses how the system verifies student identity to eliminate false registrations. The system deals with the management and evaluation of attendance of all students. The student QR code will be provided to professor for taking their attendance. The professor handling the subjects is responsible to mark the attendance for all students of the group or class. The attendance will be marked as 0 and 1, 0 for absent and 1 for present in the database of the particular student row in the table. The student attendance reports will be generated in CSV and XLS sheet for further use.

Keywords: QR, attendance, system, professor, student.

1. INTRODUCTION

Among the various types of attendance systems that have been developed, using punch cards, log books, fingerprint systems, barcodes, QR codes and also RFID still cause lots of problems such as providing incorrect information to the users. The purpose of the smartphone based attendance system is to computerize the traditional way of recording attendance and provide an easiest and smart way to track attendance in institutions nowadays, the most common device that have been come into account in marketing and business are smartphone devices. Moreover, it comprises lots of them running Android OS.

Main objective

“QR Code Based Attendance Management System” is a combination of two android applications developed for taking and storing the attendance of the students on the daily basis in the college. Here the professor, who is handling the subjects, will be responsible to mark the attendance of the students. Each staff will be given an android application that is used for taking attendance and generate the overall attendance status. An accurate report based on the student attendance is generated here. Report of the student’s attendance on weekly and monthly basis is generated as desired. The main objective of the automated attendance system is to computerize the traditional way of recording attendance and provide an efficient and automated method to track attendance in institutions. Advantages of QR Code Based Smart Attendance System:

- Provide better security.
- Maintenance of the system is easy and cost effective.
- Generate the result quickly.
- Provide accurate and efficient data.
- User friendly.

1.2. Problem statement

- Development of a SMART QR CODE BASED ATTENDANCE SYSTEM.
- Integrating Android device with QR code and SQLite to store attendance results.
- Analyzing the attendance on weekly and monthly basis.

1.3. Feasibility

- Economic feasibility:** The developed system is time effective because attendance is marked automatically. It is also cost effective because of no use of paperwork.
- Technical feasibility:** The system is economic and it does not use any other additional Hardware and software.
- Behavioral feasibility:** The system is user friendly.

1.4. Characteristic of proposed system

- User Friendly
- Reports are easily generated
- Very less paper work
- One spot solution for attendance calculation

2. METHODOLOGY

To achieve the above discussed objectives, a step-by-step methodology has been followed. The details of methodology are given below:

- Develop a QR code generator android app using the details of student such as roll number, student ID.
- Develop an Android app that take the attendance with respect to the specific subject and generate the student attendance sheet as per attendance details.

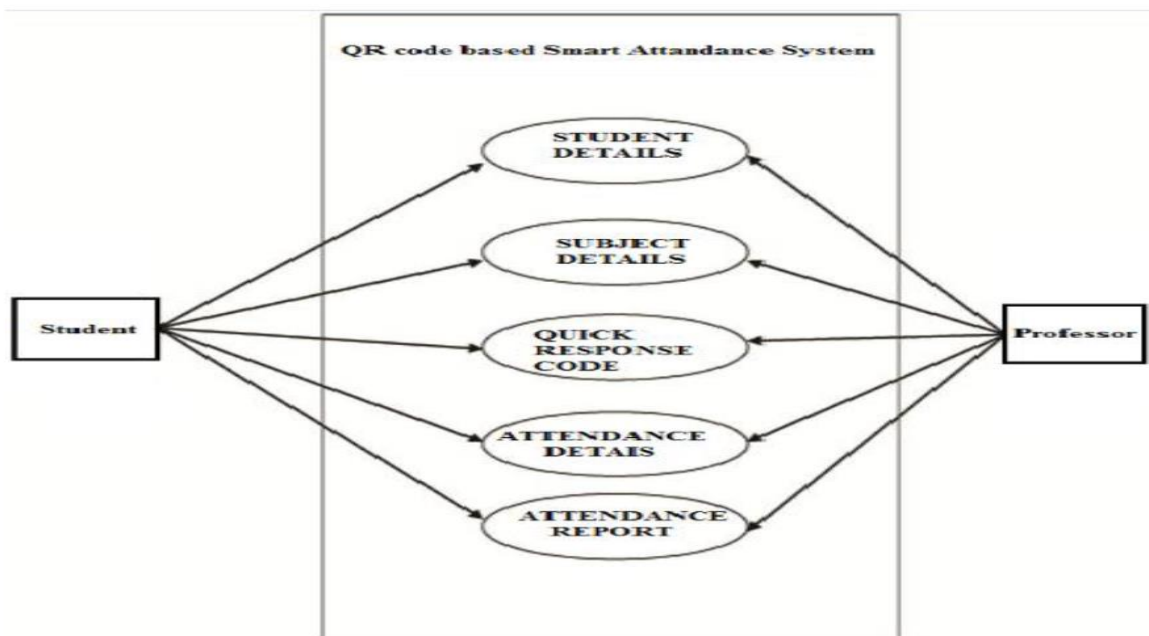


Figure 1. Use case diagram

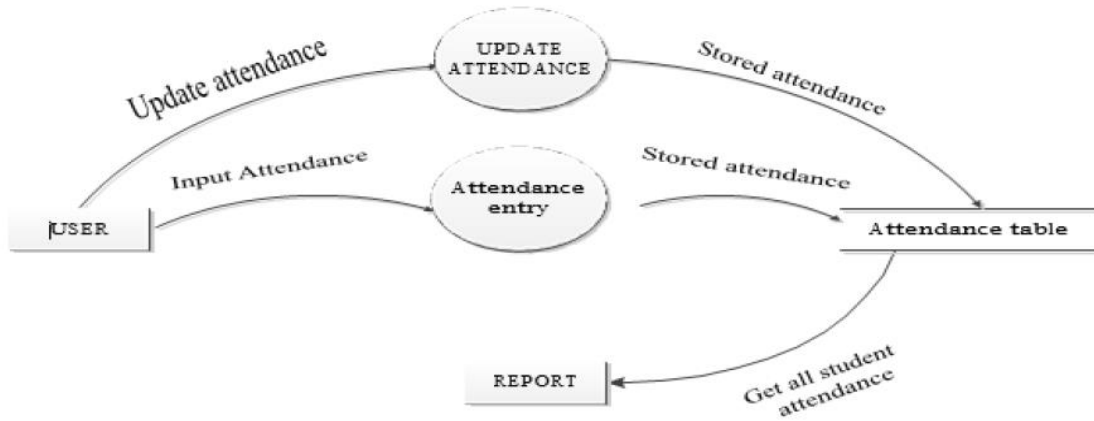


Figure 2. Data flow diagram

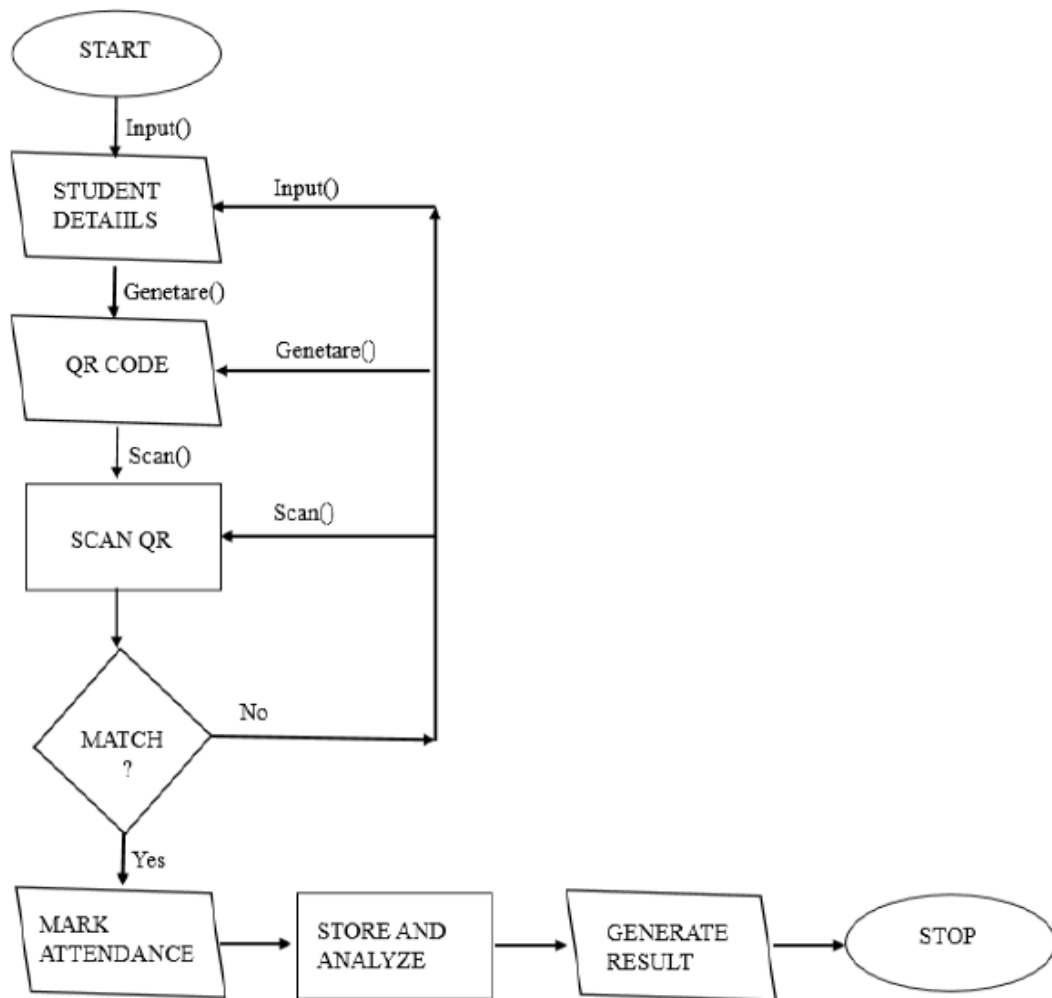
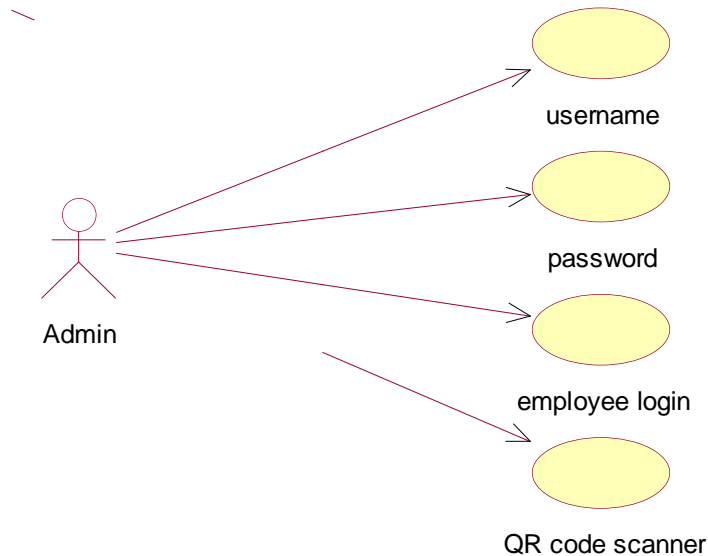


Figure 5. Flowchart of the application system

3. QR CODE BASED ATTENDANCE SYSTEM

1. REQUIRMENTS WORKFLOW

USECASE DIAGRAM :



USE CASE DESCRIPTION :

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

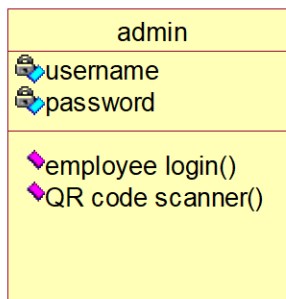
2.ANALYSIS WORKFLOW :

2.1 OBJECT ORIENT ANALYSIS :

Object-Oriented Analysis (OOA): Object-Oriented Analysis (OOA) is the first technical activity performed as part of object-oriented software engineering. OOA introduces new concepts to investigate a problem. It is based on a set of basic principles, which are as follows-

- 1.The information domain is modeled.
- 2.Behavior is represented.
- 3.The function is described.
4. Data, functional, and behavioral models are divided to uncover greater detail.
- 5.Early models represent the essence of the problem, while later ones provide implementation details.

CLASS DIAGRAM :



4. CONCLUSION

The developed system presented in this paper has been successfully designed and tested. The student's attendance status will be analysed and exported. Attendance monitoring system is very important in our daily life. It possesses a really great advantage, among the whole types of code scanning technology, QR Code Based Smart Attendance System is the most accurate. In this project report, we have given an introduction of Attendance monitoring system and its

advantage. It is an efficient method to store the attendance in the smart phone rather than wasting the paper.

5. FUTURE SCOPE

Our future work will focus on providing missed class topics and notes available to students. Full control to professor with more secured and enhanced options. Finally we conclude, if we integrates this attendance monitoring system with face identification tool then system will solve the real world attendance problem.

6. REFERENCES

- [1] "Android tutorials" [Online]. Available: <https://developer.android.com/training/index.html>
- [2] "Android tutorials" [Online]. Available: <https://www.tutorialspoint.com/android/>
- [3] "QR code integration with Android" [Online]. Available: <https://github.com/zxing/zxing>
- [4] "About Bar Code" [Online]. Available: http://files.microscan.com/whitepapers/barcode_basics.pdf
- [5] "ISS QR Code AIM Store: Historical Archive" [Online]. Available: Aimglobal.org
- [6] "Android Tutorial" [Online]. Available: <http://androidhive.com>



PREDICTION OF CARDIOVASCULAR DISEASE USING SUPERVISED LEARNING

Vasa Jothika Ganga (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT Heart Diseases have shown a tremendous hit in this modern age. As doctors deal with precious human life, it is very important for them to be right their results. Thus, an application was developed which can predict the vulnerability of heart disease, given basic symptoms like age, gender, pulse rate, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiographic results, exercise induced angina, ST depression ST segment the slope at peak exercise, number of major vessels colored by fluoroscopy and maximum heart rate achieved. This can be used by doctors to re heck and confirm on their patient"s condition. In the existing surveys they have considered only 10 features for prediction, but in this proposed research work 14 necessary features were taken into consideration. Also, this paper presents a comparative analysis of machine learning techniques like Random Forest (RF), Logistic Regression, Support Vector Machine (SVM), and Naïve Bayes in the classification of cardiovascular disease. By the comparative analysis, machine learning algorithm Random Forest has proven to be the most accurate and reliable algorithm and hence used in the proposed system. This system also provides the relation between diabetes and how much it influences heart disease

1. INTRODUCTION

Coronary illness has the biggest level of passing on the planet. In 2012, around 17.5 million individuals kicked the bucket from coronary illness, implying that it comprises of the 31% of every single worldwide passing. Besides, coronary illness loss of life rises each year. It is relied upon to develop more than 23.6 million by 2030. The exploration from the January 2017 demonstrated that the main source of death worldwide is cardiovascular infections. The cardiovascular malady is considered as a world's biggest killer and is currently taking the top position in the record of ten reasons for passing in the previous 15 years and in 2015 was numeration for fifteen million passing. Various human lives could be spared by diagnosing on schedule. Along these lines, diagnosing the

syndrome is significant and an exceptionally muddled undertaking. Mechanizing this procedure would conquer the issues with the diagnosis. The utilization of AI in ailment arrangement is normal and researchers are especially fascinated in the advancement of such frameworks for simpler following and analysis of cardiovascular diseases. Since MLpermits PC projects to ponder from information, building up a model to perceive ordinary examples and having the option to settle on choices dependent on assembled data, it doesn't have hitches with the deficiency of utilized medicinal database. The proposed model is to amass significant information relating all components identified with coronary illness and parameters impacting it, train the information according to the proposed calculation of AI andforesee how solid is



there a probability for a patient to get a coronary illness. The relationship with the diabetes related credits is considered to set up the impact

2. INPUT AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

What data should be given as input?

How the data should be arranged or coded?

The dialog to guide the operating personnel in providing input.

Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting

correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the

specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

Convey information about past activities, current status or projections of the

Future.

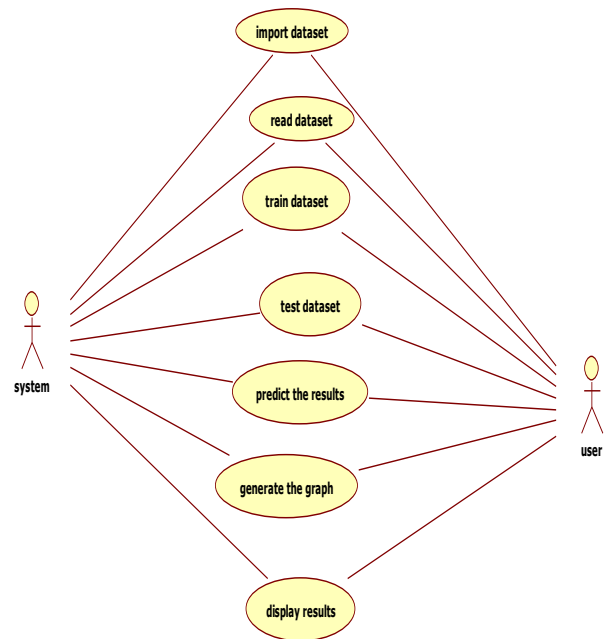
Signal important events, opportunities, problems, or warnings.

Trigger an action.

Confirm an action.

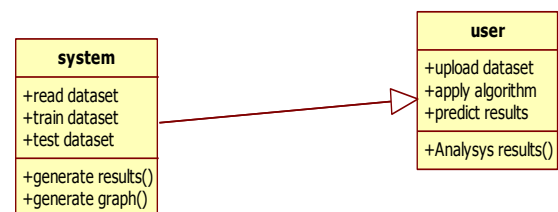
3. USE CASE DIAGRAM

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.



4. CLASS DIAGRAM

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.



5. CONCLUSION

Heart disease prediction which uses Machine learning algorithm provides users a prediction result if the user has heart disease. Recent advancements in technology made machine learning algorithms to evolve. In this proposed method Random Forest Algorithm was



used because of its efficiency and accuracy. This algorithm is also used to find the heart disease prediction percentage by knowing the correlation details between diabetes and heart diseases. The similar prediction systems can be built by calculating correlation between heart diseases and other diseases. Also new algorithms can be used to achieve increased accuracy. Better performance is obtained with more parameter used in these algorithms.

6. REFERENCES

- [1] Jaymin Patel, Prof. Tejal Upadhyay, Dr. Samir Patel "Heart disease prediction using Machine learning and Data Mining Technique" Volume 7. Number 1 Sept 2015-March 2016.
- [2] Thenmozhi.K and Deepika.P, Heart Disease Prediction using classification with different decision tree techniques. International Journal of Engineering Research & General Science, Vol 2(6), pp 6-11, Oct 2014.
- Igor Kononenko" Machine learning for medical diagnosis: history, state of art & perspective"
Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 904 - 912
Received 20 January 2021; Accepted 08 February 2021. 912 <http://annalsofrscb.ro>
- Elsevier -Artificial intelligence in Medicine, Volume 23, Aug 2001.
- [4] Gregory F. Cooper, Constantin F. Alfieris", Richard Ambrosino, John Aronisb, Bruce G. Buchanan, Richard Caruana', Michael J. Fine, Clark Glymour", Geoffrey Gordon", Barbara H. Hanusad, Janine E. Janoskyf, Christopher Meek", Tom Mitchell", Thomas Richardson", Peter Spirtes" An evaluation of machine-learning methods for predicting of pneumonia mortality"- Elsevier Feb 1997
- [5] Sana Bharti, Shailendra Narayan Singh" Analytical study of heart disease comparing with different algorithms": Computing, Communication & Automation (ICCCA), 2015 International Conference.
- [6] B. Dhomse Kanchan, M. Mahale Kishore "Study of Machine learning algorithms for special disease predictions using the principal of component analysis" Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC), 2016.
- [7] Matjaz Kuka, Igor Kononenko, Cyril Groselj, Katrina Kalif, Jure Fettich" Analysing and improving the diagnosis of ischaemic heart disease with machine learning" Elsevier -Artificial intelligence in Medicine, Volume 23, May 1999.
- [8] Geert Meyfroidt, Fabian Guiza, Jan Ramon, Maurice Brynooghe" Machine learning techniques to examine large patient databases"- Best practice & Research Clinical Anaesthesiology, Elsevier Volume 23 (1) Mar 1, 2009.
- [9] Gregory F. Cooper, Constantin F. Alfieris, Richard Ambrosino" An evaluation of Machine learning methods for predicting pneumonia mortality"- Elsevier, 1997.
- [10] Sanjay Kumar Sen" Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms", International Journal of Engineering And Computer Science ISSN:2319-7242 Volume 6 Issue 6 June 2017.
- [11] Abhishek Taneja" Heart Disease Prediction System Using Data Mining Techniques"- Vol. 6, No(4) December 2013

DETECTION OF FAKE AND CLONE ACCOUNTS IN TWITTER USING CLASSIFICATION AND DISTANCE MEASURE ALGORITHMS

Vedurupaka Hema Pavan Kumar (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

S.K.Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

Online Social Network (OSN) is a network hub where people with similar interests or real world relationships interact. As the popularity of OSN is increasing, the security and privacy issues related to it are also rising. Fake and Clone profiles are creating dangerous security problems to social network users. Cloning of user profiles is one serious threat, where already existing user's details are stolen to create duplicate profiles and then it is misused for damaging the identity of original profile owner. They can even launch threats like phishing, stalking, spamming etc. Fake profile is the creation of profile in the name of a person or a company which does not really exist in social media, to carry out malicious activities. In this paper, a detection method has been proposed which can detect Fake and Clone profiles in Twitter. Fake profiles are detected based on set of rules that can effectively classify fake and genuine profiles. For Profile Cloning detection two methods are used. One using Similarity Measures and the other using C4.5 decision tree algorithm. In Similarity Measures, two types of similarities are considered – Similarity of Attributes and Similarity of Network relationships. C4.5 detects clones by building decision tree by taking information gain into consideration. A comparison is made to check how well these two methods help in detecting clone profiles.

1. INTRODUCTION

ONLINE Social Networks (OSN) like Face book, Twitter, LinkedIn, Instagram etc are used by billions of users all around the world to build network connections. The ease and accessibility of social networks have created a new era of networking. OSN users share a lot of information in the network like photos, videos, school name, college name, phone numbers, email address,

home address, family relations, bank details, career details etc. This information if put into hands of attackers, the after effects are very severe.

Most of the OSN users are unaware of the security threats that exist in the social networks and easily fall prey to these attacks. The risks are more dangerous if the victims are children. In Profile Cloning attack, the profile information of existing users are stolen to create duplicate profiles and these profiles are misused for spoiling the identity of original profile owners[1- 6]. There are two types of Profile Cloning namely - Same Site and Cross Site Profile Cloning[1,7-9]. If user credentials are taken from one Network to create a clone profile in same Network then it is called Same Site profile cloning[1,10-12].

In Cross Site profile cloning, attacker takes the user information from one Network to create a duplicate profile in other Network in which the user is not having any account[1,13-15]. As the registration process in social networks have become very simple in order to attract more and more users, the creation of fake profiles are also increasing in an alarming rate. An attacker creates a fake profile in order to connect to a victim to cause malicious activities. And also to spread fake news and spam messages. The paper organized as below. Section II describes the literature survey. Section III explains the proposed methodology. Section IV discusses the results. At last, Section V concludes the paper with the conclusion.

2. EXISTING SYSTEM

- ❖ Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P Markatos [2] have proposed a prototype to check whether the users have become victim to cloning attack or not. Information is extracted from user profile and a search is made in OSN to find profiles which match to that of user profile and a similarity score is calculated based on commonality of attribute values. If the similarity score is above the threshold value then the particular profile is termed as clone.

- ❖ Brodka, Mateusz Sobas and Henric Johnson in their paper [3] have proposed two novel methods for detecting cloned profiles. The first method is based on the similarity of

attribute values from original and cloned profiles and the second method is based on the network relationships. A person who doubts that his profile has been cloned will be chosen as a victim. Then treating name as primary key, a search is made for profiles with the same name as that of victim, using query search. Potential clone (Pc) and the Victim profile (Pv) are compared and similarity S is calculated. If $S(Pc, Pv) > \text{Threshold}$, then profile is suspected to be a clone. In the verification step, the user does it manually as he knows which is his original profile and which one is a duplicate. Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M [4], in their paper have reviewed some of the most relevant existing features and rules (proposed by Academia and Media) for fake Twitter accounts detection. They have used these rules and features to train a set of machine learning classifiers. Then they have come up with Class A classifier which can effectively classify original and fake accounts.

- ❖ Ahmed El Azab, Amira M Idrees, Mahmoud A Mahmoud, Hesham Hefny [5], have proposed a classification method for detecting fake accounts on Twitter. They have collected some effective features for the detection process from different research and have filtered and weighted them in first stage. Various experiments are conducted to get minimum set of attributes which gives accurate results. From 22 attributes, only seven attributes were selected which can effectively detect fake accounts and have applied these factors on classification techniques. A comparison of the classification techniques based on results are made and the one which provides most accurate result is selected.

3. PROPOSED SYSTEM

Fake and clone profiles have become a very serious social threat. As information like phone number, email id, school or college name, company name, location etc are readily exposed in social networks, hackers can easily hack this information to create fake or clone profiles. They then try to cause various attacks like phishing, spamming, cyberbullying etc. They even try to defame the legitimate owner or the organisation. So, a detection method has been proposed which can detect both fake and clone profiles in order to

make the social life of the users more secure. The architecture of proposed system is as shown in the proposed system.

The proposed architecture consists of modules for Fake Profile detection and Clone Profile detection.

A. Fake Profile Detection

This module is used to detect fake Twitter profiles. Here fake profiles are detected based on rules that effectively distinguish fake profiles from genuine ones. Some of the rules that are used to detect fake profiles are - usually fake profiles do not have profile name or image. They do not include any description about the account. The geo-enabled field will be false as they do not want to expose their location in tweets.

They usually make large number of tweets or sometimes the profiles would not have made any tweets etc. The rules are applied on the profile, for each matching rule, a counter is incremented, if the counter value is greater than pre-defined threshold, then the profile is termed as fake.

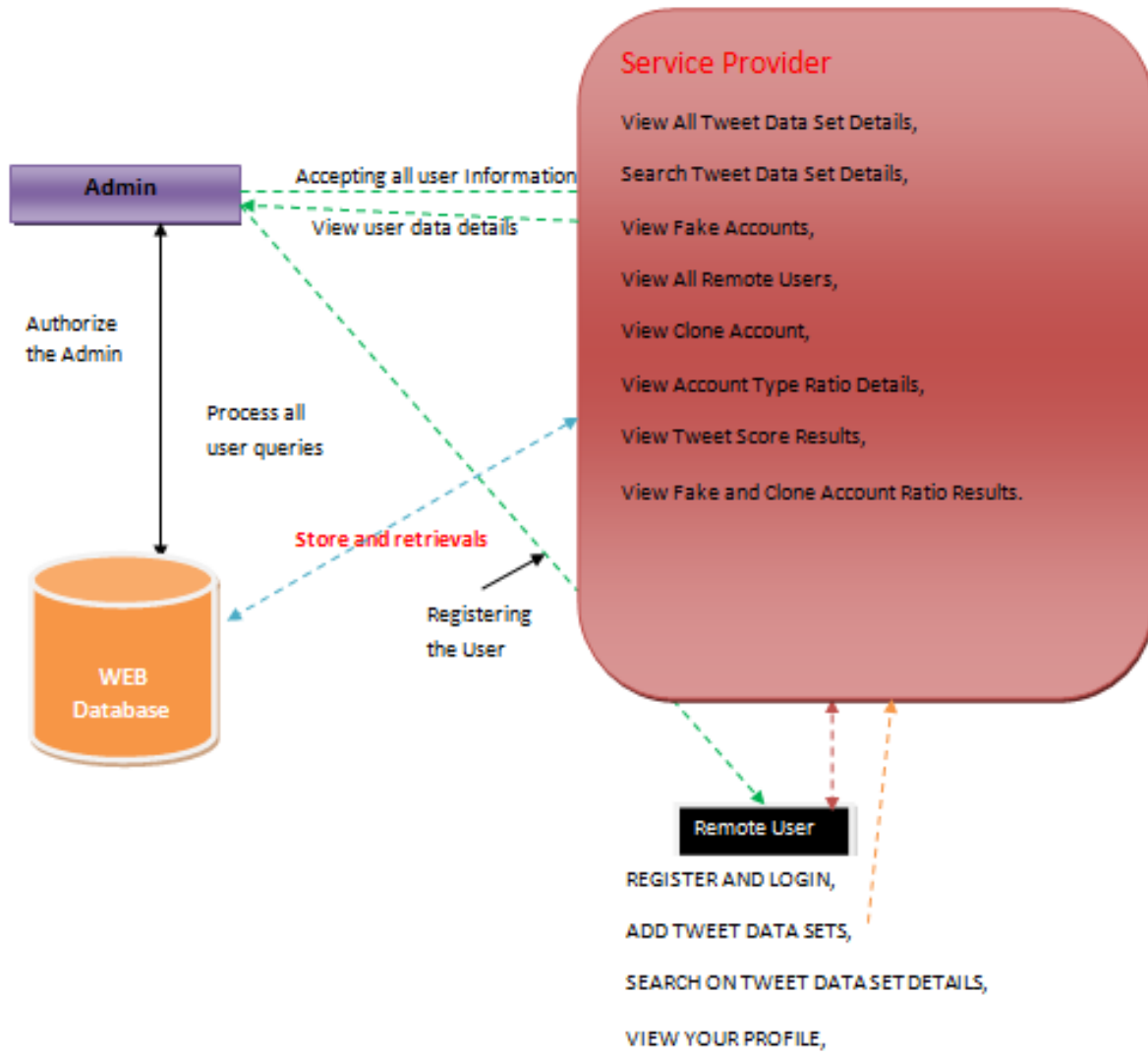
B. Clone Profile Detection using Similarity Measures

This module detects clones based on Attribute and Network similarity. User profile is taken as input. User identifying information are extracted from the profile. Profiles which are having attributes matching to that of user's profile are searched. Similarity index is calculated and if the similarity index is greater than the threshold, then the profile is termed as clone, else normal[1].

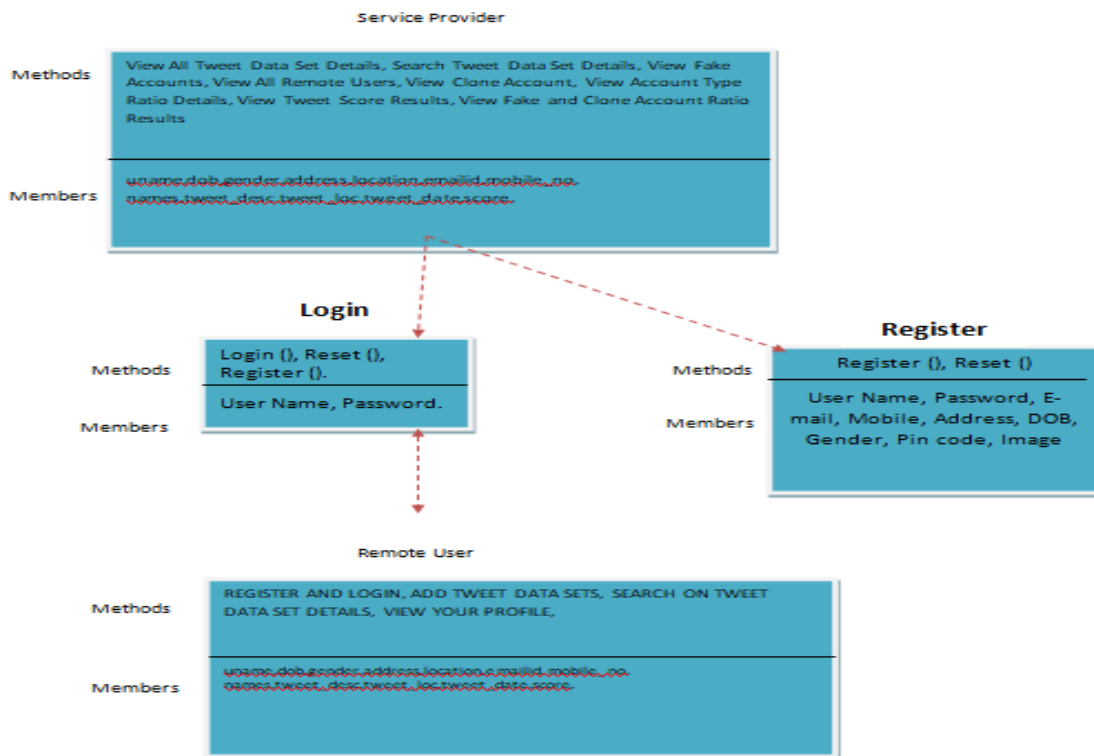
i) Attribute Similarity

Attribute similarity is calculated based on the similarity of attribute values between the profiles. The attributes that are considered for similarity measurement are Name, ScreenName, Language, Location and Time_zone. Two similarity measures are used to measure the similarity between the attributes – Cosine similarity and Levenshtein distance. Cosine similarity is used to find similarity between words and Levenshtein distance is used to find similarity between two sequences.

Architecture Diagram



➤ **Class Diagram :**



4. SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY

◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5. CONCLUSIONS

Fake and clone profiles have become a very serious problem in online social networks. We hear some or the other threats caused by these profiles in everyday life. So a detection method has

been proposed which can find both fake and clone Twitter profiles. For fake detection, a set of rules were used which when applied can classify fake and genuine profiles.

6. REFERENCES

- [1] Sowmya P and Madhumita Chatterjee ,” Detection of Fake and Cloned Profiles in Online Social Networks”, Proceedings 2019: Conference on Technologies for Future Cities (CTFC)
- [2] Georgios Kontaxis, Iasonas Polakis, Sotiris Ioannidis and Evangelos P. Markatos, “Detecting Social Network Profile Cloning”, 2013
- [3] Piotr Brodka, Mateusz Sobas and Henric Johnson, “Profile Cloning Detection in Social Networks”, 2014 European Network Intelligence Conference
- [4] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angello Spognardi, Maurizio Tesconi, “Fame for sale: Efficient detection of fake Twitter followers”, 2015 Elsevier’s journal Decision Support Systems, Volume 80
- [5] Ahmed El Azab, Amira M Idrees, Mahmoud A Mahmoud, Hesham Hefny, “Fake Account Detection in Twitter Based on Minimum Weighted Feature set”, World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering Vol:10, 2016
- [6] M.A.Devmane and N.K.Rana, “Detection and Prevention of Profile Cloning in Online Social Networks”, 2014 IEEE International Conference on Recent Advances and Innovations in Engineering
- [7] Kiruthiga. S, Kola Sujatha. P and Kannan. A, “Detecting Cloning Attack in Social Networks Using Classification and Clustering Techniques” 2014 International Conference on Recent Trends in Information Technology
- [8] Buket Erşahin, Ozlem Aktaş, Deniz Kilinc, Ceyhun Akyol, “Twitter fake account detection”, 2017 International Conference on Computer Science and Engineering (UBMK)
- [9] Arpitha D, Shrilakshmi Prasad, Prakruthi S, Raghuram A.S, “Python based Machine Learning for Profile Matching”, International Research Journal of Engineering and Technology (IRJET), 2018
- [10] Olga Peled, Michael Fire, Lior Rokach, Yuval Elovici, “Entity Matching in Online Social Networks”, 2013 International Conference on Social Computing

SOCIAL MEDIA AND MISLEADING INFORMATION IN A DEMOCRACY A MECHANISM DESIGN APPROACH

Vegesna Sai Prasanna Swarupa (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

ABSTRACT

In this paper, we present a resource allocation mechanism for the problem of incentivizing filtering among a finite number of strategic social media platforms. We consider the presence of a strategic government and private knowledge of how misinformation affects the users of the social media platforms. Our proposed mechanism incentivizes social media platforms to filter misleading information efficiently, and thus indirectly prevents the spread of fake news. In particular, we design an economically inspired mechanism that strongly implements all generalized Nash equilibria for efficient filtering of misleading information in the induced game. We show that our mechanism is individually rational, budget balanced, while it has at least one equilibrium. Finally, we show that for quasi-concave utilities and constraints, our mechanism admits a generalized Nash equilibrium and implements a Pareto efficient solution.

1. INTRODUCTION

For the last few years, political commentators have been indicating that we live in a *post-truth* era [1], wherein the deluge of information available on the internet has made it extremely difficult to identify facts. As a result, individuals have developed a tendency to form their opinions based on the *believability* of presented information rather than its truthfulness [2]. This phenomenon is exacerbated by the business practices of social media platforms, which often seek to maximize the *engagement* of their users at all costs. In fact, the algorithms developed by platforms for this purpose often promote conspiracy theories among their users [3].

The sensitivity of users of social media platforms to conspiratorial ideas makes them an ideal terrain to conduct political misinformation campaigns [4], [5]. Such campaigns are especially effective tools to disrupt democratic institutions, because the functioning of stable democracies relies on *com- mon knowledge* about the political actors and the processes they can

use to gain public support [6]. The trust held by the citizens of a democracy on common knowledge includes: (i) trust that all political actors act in good faith when contesting for power, (ii) trust that elections lead to a free and fair transfer of power between the political actors, and (iii) trust that democratic institutions ensure that elected officials wield their power in the best interest of the citizens. In contrast, citizens of democracies often have a *contested knowledge* regarding who should hold power and how they should use it [6]. The introduction of *alternative facts* can reduce the trust on common knowledge about democracy, especially if they become accepted beliefs among the citizens. Such disruptions on the trust on common knowledge can be found in the 2016 U.S. elections [7] and Brexit Campaign in 2016 [8], where the spread of misinformation through social media platforms resulted in a large number of citizens mistrusting the results of voting. To tackle this growing phenomenon of misinformation, in this paper, we consider a finite group of social media platforms, whose users represent the citizens in a democracy, and a democratic government. Every post in the platforms is associated with a parameter that captures its informativeness, which can take values between two extremes: (i) completely factual and (ii) complete misinformation. In our framework, posts that exhibit misinformation can lead to a decrease in trust on common knowledge among the users [9]–[12]. In addition, social media platforms are considered to have the technologies to *filter*, or label, posts that intend to sacrifice trust on common knowledge. Thus, the government seeks to incentivize the social media platforms to use these technologies and filter any misinformation included in the posts.

Motivated by capitalistic values, we induce a *misinformation filtering game* to describe the interactions between the social media platforms and the government. In this game, each platform acts as strategic player seeking to maximize their advertisement revenue from the engagement of their users [7], [13]. User engagement is a metric that can be used to quantify the interaction of users with a platform, and subsequently, how much time they spend on the platform. Recent efforts reported in the literature on misinformation in social media platforms have indicated that increasing filtering of misinformation leads to decreasing of user engagement [14]. There are many possible reasons for this phenomenon. First, filtering reduces the total number of posts propagating across the social network. Second, the users whose opinions are filtered may perceive this action as dictatorial censorship [15], and as a result, they may choose to express their opinions in other platforms. Finally, misinformation tends to elicit stronger

reactions, e.g., surprise, joy, sadness, as compared to factual posts [16], which may increase user engagement. Thus, each platform is reluctant to filter misinformation.

In our framework, we consider that the government is also a strategic player, whose utility increases as the trust of the users of social media platforms on common knowledge increases. Consequently, increasing filtering of misinformation by the social media platforms increases the utility of the government. Thus the government is willing to make an investment to incentivize the social media platforms to filter misinformation. In our approach, we use mechanism design to distribute this investment among the platforms optimally, and in return, implement an optimal level of filtering.

Mechanism design was developed for the implementation of system-wide optimal solutions to problems involving multiple rational players with conflicting interests, each with private information about preferences [17]. Note that this approach is different from traditional approaches to decentralized control with private information [18]–[21] because the players are not a part of the same time, but in fact, have private and competitive utilities. The fact that Mechanism design optimizes the behaviour of competing players has led to broad applications spanning different fields including economics, politics, wireless networks, social networks, internet advertising, spectrum and bandwidth trading, logistics, supply chain, management, grid computing, and resource allocation problems in decentralized systems [22]–[28].

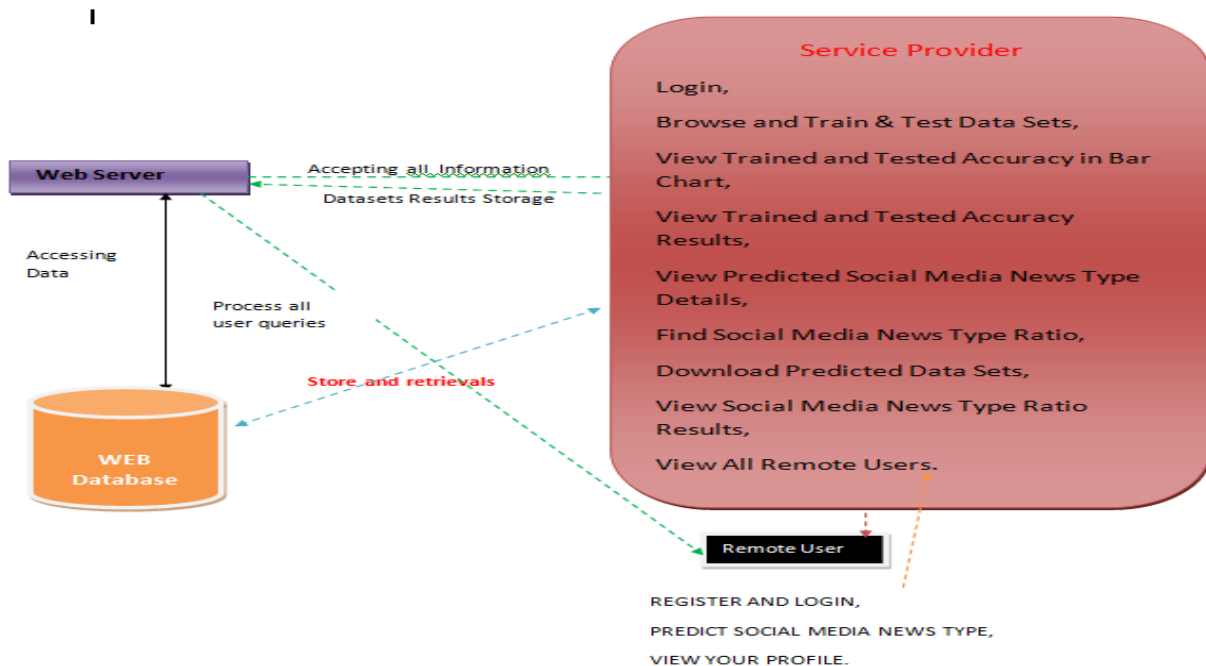
The contribution of this paper is as follows. We present an indirect mechanism to incentivize social media platforms to filter misleading information. We show that our proposed mechanism is (i) feasible, (ii) budget balanced, (iii) individual rational, and (iv) strongly implementable at the equilibria of the induced game. We prove the existence of at least one generalized Nash equilibrium and show that our mechanism induces a Pareto efficient equilibrium.

The rest of the paper is organized as follows. In Section II, we provide the modeling framework and problem formulation. In Section III, we present our mechanism, and in Section IV, we prove the associated properties of the mechanism. In Section V, we interpret the mechanism and present a descriptive example. Finally, in Section VI we conclude and present some directions for future research.

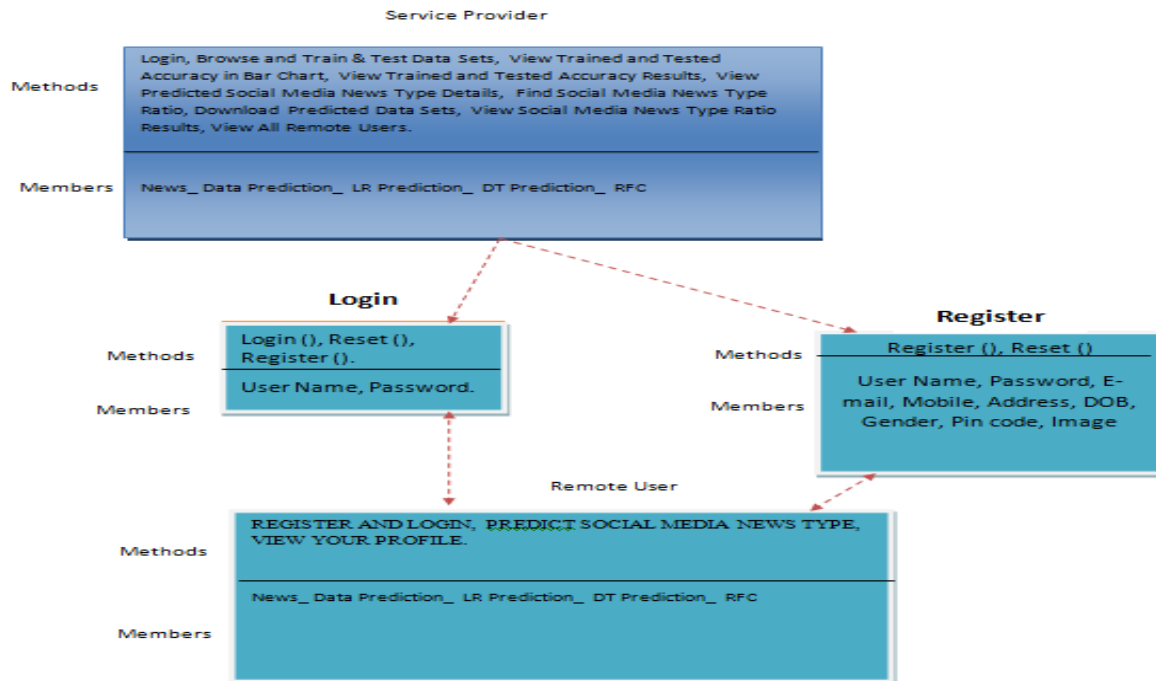
1.1 EXISTING SYSTEM

social media in particular, has generated extraordinary concern, in large part because of its potential effects on public opinion, political polarization, and ultimately democratic decision making. Recently, however, a handful of papers have argued that both the prevalence and consumption of “fake news” per se is extremely low compared with other types of news and news-relevant content. Although neither prevalence nor consumption is a direct measure of influence, this work suggests that proper understanding of misinformation and its effects requires a much broader view of the problem, encompassing biased and misleading—but not necessarily factually incorrect—information that is routinely produced or amplified by mainstream news organizations. In this paper, we propose an ambitious collective research agenda to measure the origins, nature, and prevalence of misinformation, broadly construed, as well as its impact on democracy. We also sketch out some illustrative examples of completed, ongoing, or planned research projects that contribute to this agenda.

Architecture Diagram



➤ **Class Diagram :**



2. PRELIMINARY INVESTIGATION

The first and foremost strategy for development of a project starts from the thought of designing a mail enabled platform for a small firm in which it is easy and convenient of sending and receiving messages, there is a search engine ,address book and also including some entertaining games. When it is approved by the organization and our project guide the first activity, ie. preliminary investigation begins. The activity has three parts:

- **Request Clarification**
- **Feasibility Study**
- **Request Approval**

REQUEST CLARIFICATION

After the approval of the request to the organization and project guide, with an investigation being considered, the project request must be examined to determine precisely what the system requires.

Here our project is basically meant for users within the company whose systems can be interconnected by the Local Area Network(LAN). In today's busy schedule man need everything should be provided in a readymade manner. So taking into consideration of the vastly use of the net in day to day life, the corresponding development of the portal came into existence.

FEASIBILITY ANALYSIS

An important outcome of preliminary investigation is the determination that the system request is feasible. This is possible only if it is feasible within limited resource and time. The different feasibilities that have to be analyzed are

- **Operational Feasibility**
- **Economic Feasibility**
- **Technical Feasibility**

Operational Feasibility

Operational Feasibility deals with the study of prospects of the system to be developed. This system operationally eliminates all the tensions of the Admin and helps him in effectively tracking the project progress. This kind of automation will surely reduce the time and energy, which previously consumed in manual work. Based on the study, the system is proved to be operationally feasible.

Economic Feasibility

Economic Feasibility or Cost-benefit is an assessment of the economic justification for a computer based project. As hardware was installed from the beginning & for lots of purposes thus the cost on project of hardware is low. Since the system is a network based, any number of employees connected to the LAN within that organization can use this tool from at anytime. The Virtual Private Network is to be developed using the existing resources of the organization. So the project is economically feasible.

Technical Feasibility

According to Roger S. Pressman, Technical Feasibility is the assessment of the technical resources of the organization. The organization needs IBM compatible machines with a graphical web browser connected to the Internet and Intranet. The system is developed for platform

Independent environment. Java Server Pages, JavaScript, HTML, SQL server and WebLogic Server are used to develop the system. The technical feasibility has been carried out. The system is technically feasible for development and can be developed with the existing facility.

2.2 REQUEST APPROVAL

Not all request projects are desirable or feasible. Some organization receives so many project requests from client users that only few of them are pursued. However, those projects that are both feasible and desirable should be put into schedule. After a project request is approved, its cost, priority, completion time and personnel requirement is estimated and used to determine where to add it to any project list. Truly speaking, the approval of those above factors, development works can be launched.

3. SYSTEM DESIGN AND DEVELOPMENT

INPUT DESIGN

Input Design plays a vital role in the life cycle of software development, it requires very careful attention of developers. The input design is to feed data to the application as accurate as possible. So inputs are supposed to be designed effectively so that the errors occurring while feeding are minimized. According to Software Engineering Concepts, the input forms or screens are designed to provide to have a validation control over the input limit, range and other related validations.

This system has input screens in almost all the modules. Error messages are developed to alert the user whenever he commits some mistakes and guides him in the right way so that invalid entries are not made. Let us see deeply about this under module design.

Input design is the process of converting the user created input into a computer-based format. The goal of the input design is to make the data entry logical and free from errors. The error in the input are controlled by the input design. The application has been developed in user-friendly manner. The forms have been designed in such a way during the processing the cursor is placed in the position where must be entered. The user is also provided with in an option to select an appropriate input from various alternatives related to the field in certain cases.

Validations are required for each data entered. Whenever a user enters an erroneous data, error message is displayed and the user can move on to the subsequent pages after completing all the entries in the current page.

4. CONCLUSIONS

Our primary goal in this paper was to design a mechanism to induce a GNE solution in the misinformation filtering game, where (i) each platform agrees to participate voluntarily, and (ii) the collective utility of the government and the platforms is maximized. We designed a mechanism and proved that it satisfies these properties along with budget balance. We also presented an extension of the mechanism with weaker technical assumptions.

Ongoing work focuses on improving the valuation and average trust functions of the social media platforms based on data. We also consider incorporating uncertainty in a platform's estimates of the impact of their filter. These refinements of the modeling framework will allow us to make our mechanism more practical for use in the real world.

Future research should include extending the results of this paper to a dynamic setting in which the social media platforms react in real-time to the proposed taxes/subsidies. In particular, someone could develop an algorithm that the players can use to iteratively arrive at the Nash equilibrium. In such an algorithm, the social planner can receive additional information from the players while they iteratively learn the GNE. Then, she can use this information to change her allocations dynamically, allowing us to relax either Assumption 5 on monitoring of average trust, or Assumption 6 on the excludability of the platforms.

REFERENCES

- [1] W. Davies, "The age of post-truth politics," *The New York Times*, vol. 24, p. 2016, 2016.
- [2] J. Cone, K. Flaharty, and M. J. Ferguson, "Believability of evidence matters for correcting social impressions," *Proceedings of the National Academy of Sciences*, vol. 116, no. 20, pp. 9802–9807, 2019.
- [3] Z. Tufekci, "Youtube, the great radicalizer," *The New York Times*, vol. 10, 2018.
- [4] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

- [5] J. Weedon, W. Nuland, and A. Stamos, "Information operations and facebook," *Retrieved from: <https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf>*, 2017.
- [6] H. Farrell and B. Schneier, "Common-knowledge attacks on democracy," *Berkman Klein Center Research Publication*, no. 2018-7, 2018.
- [7] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.
- [8] O. Analytica, "Russia will deny cyberattacks despite more us evidence," *Emerald Expert Briefings*, no. oxan-db, 2018.
- [9] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi, "Science vs conspiracy: Collective narratives in the age of misinformation," *PloS one*, vol. 10, no. 2, p. e0118093, 2015.
- [10] E. Brown, "Propaganda, misinformation, and the epistemic value of democracy," *Critical Review*, vol. 30(3–4), pp. 194–218, 2018.

DISEASE PREDICTION USING MACHINE LEARNING

Velamala Naga Durga Gayathri (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

Abstract

In this digital world, most of the people are prone to diseases, due to lack of healthy food, proper sleep and daily exercise. It is very crucial to know if we are suffering from a disease, at an early stage rather than discovering it at a later stage. Hence disease prediction system plays an important role as it predicts the diseases based on symptoms. This disease prediction system uses Machine Learning algorithm named Random Forest. This system also suggests drugs that are most commonly used to cure the disease.

Indexed Terms- Diseases, Drugs, Machine Learning, Prediction, And Random Forest.

1. INTRODUCTION

With the advancement in technology, Machine Learning is becoming more popular and commonly used technology by industry experts for solving problems faced in real life. Machine Learning is the scientific study of algorithms and statistical models that computer use to perform a specific task without using explicit instructions, relying on patterns and inference instead. Machine Learning is also used by the healthcare industry to bring advancement in their techniques so that they can provide better services to their patients. The disease prediction system predicts diseases based on patient's symptoms and also some commonly prescribed medicines for a particular disease.

1.1. EXISTING SYSTEM

There are many prevalent systems used for disease prediction. The existing systems

only predict the diseases. The various approaches used for predicting diseases is by using Machine Algorithms such as Naïve Bayes, Decision Tree, Random Forest, k-mean algorithm. Also, one of the approaches to build a disease prediction system is by using Big Data. Prediction using traditional disease risk model usually involves Machine Learning and supervised learning

algorithm which uses training data with the labels for the training of the models.

1.2. PROPOSED SYSTEM

In the proposed system, a disease prediction model is built using a Machine Learning algorithm that is Random Forest Algorithm. Based on the symptoms that are input by the user, the disease is predicted and the drug that is most commonly prescribed by the doctor is suggested.



2. MEDICAL DATA ANALYSIS USING ML

Since the arrival of advanced computing, the doctors' still requires the technology in various possible ways like surgical representation process and x-ray photography, but the technology perceptually stayed behind. The method still requires the doctor's information and experience due to alternative factors starting from medical records to weather conditions, atmosphere, blood pressure and numerous alternative factors. The huge numbers of variables are consider as entire variables that are required to understand the complete working process itself, however no model has analyzed successfully. To tackle this drawback, Medical decision support systems must be used. This system is able to assist the doctors to make the correct decision. Medical decision support system refers to both the process of attempting to determine or identify possible diseases or disorder and the opinion reached by this process. Thediagnostic opinion in the sense, it indicates either degree of abnormality on a continuum or a kind of abnormality in a classification. It's influenced by non medical factors such as power ethics and financial incentives for patient or doctor. It can be a brief summation or an extensive formulation, even taking the form of story or metaphor. It might be a means of communication such as computer code through which it triggers payment, prescription, notification, information or advice. Indication of medical diagnostic includes knowledge of what is normal and

measuring of patient's current condition. Automated decision support systems are rule based systems that are automatically providing solutions to repetitive management problems.

Medical decision could be extremely specialized and difficult job due to alternative factors or incase of rare diseases. The alternative factors include stress; tired misdiagnosis might vary from ignorance of doctors and incomplete information. Standard algorithm may go through the entire variables like prevailing conditions history of medical records, history of family records and various factors relating to the patient records, sheer magnitude of obtainable hidden factors. Differential diagnosis methods can be used to identify the presence of an entity where multiple alternatives are possible and also refers to include the candidate alternatives. This method is needs a process of elimination or obtaining information that shrinks the probability of candidate conditions to negligible levels. It contains four steps: 1) The doctor gather all information about the patients and create a symptoms list.2) The doctor should make a list of all possible causes of symptoms.3) The doctor should prioritize the list by which is the most dangerous possible cause of symptoms put in the top of the list.4) The doctor should rule out or treat the possible causes beginning with the most urgently dangerous conditions."Rule Out" in the sense to usethe test method or other scientific method. If there will be no such diagnosis means removing the diagnosis from the list and using tests that should have distinct results,



depends on which diagnosis is correct. This can be done based on the doctor's knowledge and experience. This method is very easy to implement.

To reduce the large number of variables and find the most probable diseases by using the K-Means algorithm. This algorithm is more suitable to cluster the more number of diseases. K-Means is one of the unsupervised learning algorithms which are used to solve the clustering problem. The main idea is to determine the k centroids, one for each cluster. Different tests performed on the patients will serve as attributes for clustering. By using this algorithm it reduces the number of iterations, boundaries of clusters are well defined without overlapping, to produce the accurate result for each and every diagnosis. This system uses Service oriented architecture (SOA), anyone can access with internet connections and LAMSTAR Network can be used to calculate the weight, to increase the accuracy of algorithm, overall speed test and produce the better result.

3. CONCLUSION

In this paper, algorithm used to predict the disease based on symptoms is discussed. Various symptoms are provided in the dropdown menu, out of which user selects any five of them and using algorithm the disease is predicted. The drugs that are commonly prescribed for a particular disease can also be suggested in this system.

The main aim is to predict the disease at the early stage and lead to early diagnosis. This system can also be used by doctors to avoid

confusion while predicting the disease. This system can provide assistance to doctors.

4. REFERENCES

- [1] Arthur Samuel, Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. Artificial Intelligence in Design '96. Springer, Dordrecht.
- [2] A. C. Jamgade, Prof. S. D. Zade (May 2019). Disease prediction using Machine Learning. International Research Journal of Engineering and Technology (IRJET). [Online]. Available: <https://www.irjet.net/archives/V6/i5/IRJET-V6I5977.pdf>
- [3] A. Borad (October 2018). How to develop Machine Learning applications for Business. eInfochips- An Arrow Company [Online]. Available: <https://www.einfochips.com/blog/how-to-develop-machine-learning-applications-for-business/>
- [4] Classification Algorithms – Random Forest. Tutorialspoint [Online]. Available: https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm

AN EXPERIMENTAL STUDY FOR SOFTWARE QUALITY PREDICTION USING MACHINE LEARNING METHODS

Vennavalli Charishma (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

Software quality estimation is an activity needed at various stages of software development. It may be used for planning the project's quality assurance practices and for benchmarking. In earlier previous studies, two methods (Multiple Criteria Linear Programming and Multiple Criteria Quadratic Programming) for estimating the quality of software had been used. Also, C5.0, SVM and Neural network were experimented with for quality estimation. These studies have relatively low accuracies. In this study, we aimed to improve estimation accuracy by using relevant features of a large dataset. We used a feature selection method and correlation matrix for reaching higher accuracies. In addition, we have experimented with recent methods shown to be successful for other prediction tasks. Machine learning algorithms such as Xgboost, Random Forest and Decision Tree are applied to the data to predict the software quality and reveal the relation between the quality and development attributes. The experimental results show that the quality level of software can be well estimated by machine learning algorithms.

Keywords: — Estimation, Machine Learning, Software Quality.

1. INTRODUCTION

Software applications may contain defects, originating from requirements analysis, specification and other activities conducted in the software development. Therefore, software quality estimation is an activity needed at various stages [1]. It may be used for planning the project based quality assurance practices and for benchmarking. In addition, the number of defects per unit is considered one of the most important factors that indicate the quality of the software [2].

There are two directly comparable studies on software quality prediction using defect quantities in ISBGS dataset. In the first study, the two methods (MCLP and MCQP) were experimented with the dataset and the results were compared [3]. The quality level was classified according to: number of minor defect + 2*number of major defect + 4*number of extreme defect. The quality of level was to be either high or low. They used k-fold cross-validation technique to measure MCLP and MCQP's performance on the ISBSG database. Release 10 Dataset (released in January 2007) which contained 4,017 records and 106 attributes was used. After preprocessing, 374 records and 11 attributes remained in the dataset.

In another study, the same data set was used again [4]. The software belonged to high quality class if it fulfills the following requirements: the extreme defects exist or the number of major defects is more than 1 or the number of minor defects is more than 10. The rest are assumed to belong to low quality class. After preprocessing, 746 projects and 53 attributes remained in the dataset. They used C5.0, SVM and Neural network for classification.

As an example to a more application oriented study Rashid et al. [5] used case based reasoning (CBR) for software quality estimation. CBR is a machine learning model which performs the learning process using the results of the previous experiments. Line of code, number of function, difficulty level, and development type and programmers experience are entered and these attributes are used for estimation. The deviation is calculated by using Euclidian distance (ED) or The Manhattan distance (MD). If the error in estimation is less than 10% then the record is saved to the database. Number of inputs that can be obtained from the user is limited. Also, it is necessary to have close values in the database in order to estimating precise values.

In these studies, quality estimation was done by binary classification. We tried to improve these prediction models, taking into account the size in terms of function points and using 4-level classification. We have experimented with recent classification methods shown to be successful for other prediction tasks.

2. EXISTING SYSTEM

Software quality estimation is an activity needed at various stages of software development. It may be used for planning the project's quality assurance practices and for benchmarking. In earlier previous studies, two methods (Multiple Criteria Linear Programming and Multiple

Criteria Quadratic Programming) for estimating the quality of software had been used. Also, C5.0, SVM and Neural network were experimented with for quality estimation. These studies have relatively low accuracies.

Demerits of Existing System

To improve estimation accuracy by using relevant features of a large dataset.

3. PROPOSED SYSTEM

In this paper We used a feature selection method and correlation matrix for reaching higher accuracies. In addition, we have experimented with recent methods shown to be successful for other prediction tasks. Machine learning algorithms such as Xgboost, Random Forest and Decision Tree are applied to the data to predict the software quality and reveal the relation between the quality and development attributes. The experimental results show that the quality level of software can be well estimated by machine learning algorithms.

Feasibility of operation

Proposed projects are beneficial only if they can be turned out into information system. That will meet the organization's operating requirements. Operational feasibility aspects of the project are to be taken as an important part of the project implementation. Some of the important issues raised are to test the operational feasibility of a project includes the following: -

- Is there sufficient support for the management from the users?
- Will the system be used and work properly if it is being developed and implemented?
- Will there be any resistance from the user that will undermine the possible application benefits?

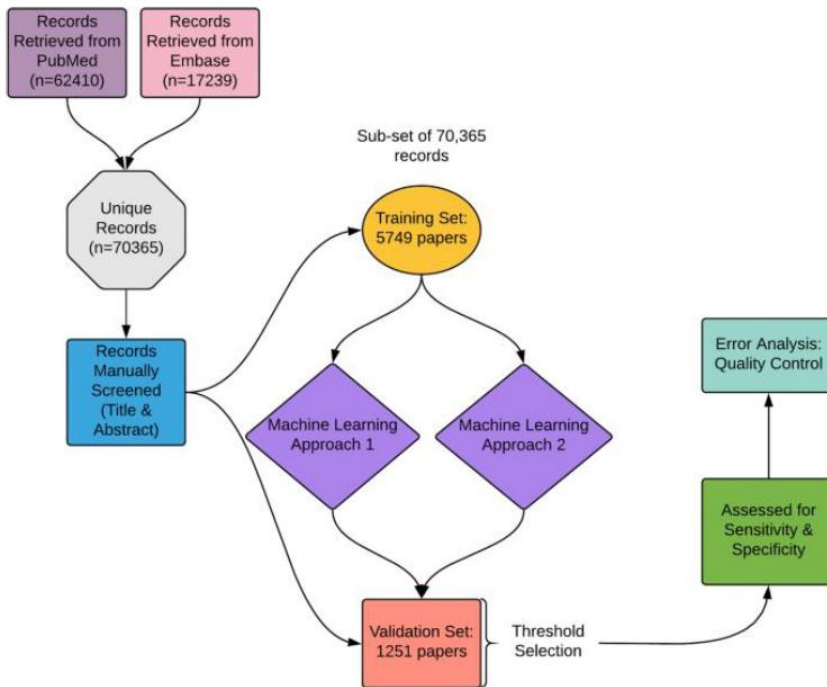
This system is targeted to be in accordance with the above-mentioned issues. Beforehand, the management issues and user requirements have been taken into consideration. So there is no question of resistance from the users that can undermine the possible application benefits. The well-planned design would ensure the optimal utilization of the computer resources and would help in the improvement of performance status.

4. SYSTEM DESIGN

Data flow diagrams delineate the way the data is stored within a system as well as input and return resources are concerned. It is easy to use information source diagrams to consider giving some market function. The strategy begins with an introductory image of the company and progresses by disintegrating each of the beneficial unmistakably outposts.

System Architecture

Software design sits at the technical kernel of the software engineering process and is applied regardless of the development paradigm and area of application. Design is the first step in the development phase for any engineered product or system. The designer’s goal is to produce a model or representation of an entity that will later be built. Beginning, once system requirement have been specified and analyzed, system design is the first of the three technical activities - design, code and test that is required to build and verify software.



Class diagram:

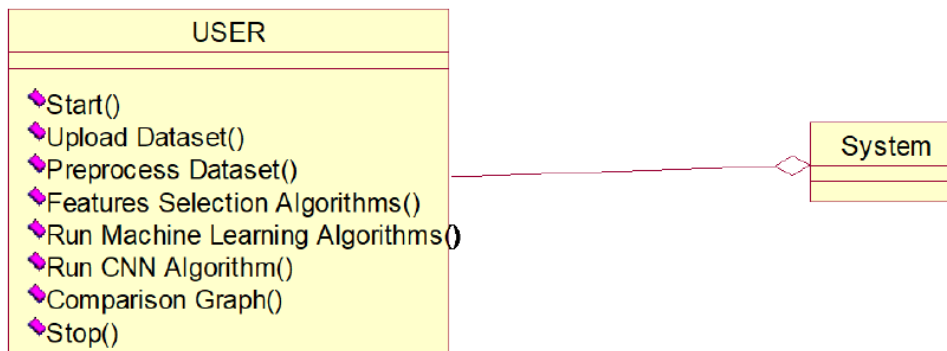
The class outline is the essential structure square of occasion driven information investigation. It is utilized both for by and large sensible confirmation of the application's exactness and for

detailed show making a model translation into software code. Equally, class graphs can be used to demonstrate information.

Every class in the class chart resembles both to the fundamental items, to the application participations and to the classes to be altered. A class has three locales; classes with boxes which contain three sections are addressed in the framework:

- The name is given by the first / top component
- The pivot section includes the class features
- the techniques or actions which the class may take or throw back is shown at the base

The underneath Fig 3.3 shows the class diagram of the project



1 FEASIBILITY STUDY:

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY

- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

5. CONCLUSION AND FUTURE ENHANCEMENT

In this paper we have experimented classification algorithms using Scikit-learn library on two dataset. We have experimented with recent algorithms that support multi-class classification. The accuracies achieved by using these algorithms are 92.28% on EBSPM Dataset and 92.22% on ISBSG Dataset. In comparison to previous directly comparable studies, acceptable level multiclass quality prediction could be achieved.

6. REFERENCES

- [1] Vijay, T. John, D. M. G. Chand, and D. H. Done. "Software quality metrics in quality assurance to study the impact of external factors related to time." International Journal of Advanced Research in Computer Science and Software Engineering, 2017.
- [2] D. Bowes, T. Hall, and J. Petrić, "Software defect prediction: do different classifiers find the same defects?." Software Quality Journal, 26(2), 2018, pp. 525-552.
- [3] X. Wang, Y. Zhang, L. Zhang and Y. Shi, "A Knowledge Discovery Case Study of Software Quality Prediction: ISBSG Database," 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, ON, 2010, pp. 219-222.
- [4] X. Wang, Y. Zhang, L. Zhang and Y. Shi, "A Knowledge Discovery Case Study of Software Quality Prediction Based on Classification Models: ISBSG Database," The 11th International Symposium on Knowledge Systems Sciences (KSS 2010), 2010
- [5] E. Rashid, S. Patnaik, and V. Bhattacharjee, "Software quality estimation using machine learning: Case-Based reasoning technique, " International Journal of Computer Applications, 2012
- [6] www.isbsg.org
- [7] <https://goverdson.nl/>

[8] H. Huijgens, "Evidence-based software portfolio management: a tool description and evaluation", 20th International Conference on Evaluation and Assessment in Software Engineering (EASE '16), 201

DEEP TEXTURE FEATURES FOR ROBUST FACE SPOOFING DETECTION

Viswanadham Durga Prabhu Prasad (MCA Scholar), B V Raju College, Vishnupur,
Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

Abstract:

Biometric systems are quite common in our everyday life. Despite the higher difficulty to circumvent them, nowadays criminals are developing techniques to accurately simulate physical, physiological, and behavioral traits of valid users, process known as spoofing attack. In this context, robust countermeasure methods must be developed and integrated with the traditional biometric applications in order to prevent such frauds. Despite face being a promising trait due to its convenience and acceptability, face recognition systems can be easily fooled with common printed photographs. Most of state-of-the-art antispoofing techniques for face recognition applications extract handcrafted texture features from images, mainly based on the efficient local binary patterns (LBP) descriptor, to characterize them. However, recent results indicate that high-level (deep) features are more robust for such complex tasks. In this brief, a novel approach for face spoofing detection that extracts deep texture features from images by integrating the LBP descriptor to a modified convolutional neural network is proposed. Experiments on the NUAA spoofing database indicate that such deep neural network (called LBPnet) and an extended version of it (n-LBPnet) outperform other state-of-the-art techniques, presenting great results in terms of attack detection.

1. INTRODUCTION

Biometrics, i.e., people automatic recognition based on their physical, physiological or behavioral traits, presents many advantages over the traditional knowledge and possess-based identification systems [1], [2]. However, nowadays criminals are already developing sophisticated techniques to accurately simulate traits of valid users, process known as spoofing attack. In this sense, countermeasures, i.e., robust spoofing detection techniques, must be developed and integrated with biometric systems in order to prevent such frauds [3]. There are

many points of attack in security systems that can be exploited by criminals. In the case of biometric systems, the great majority of attacks occur by fooling the capture sensor with synthetic traits since no knowledge regarding the inner working of the application is needed [4]. Among the main biometric traits, face is a promising one especially due to its convenience, low cost of acquisition and acceptability by users, being very suitable to a wide variety of environments, including mobile ones. However, despite all these advantages, face recognition systems are the ones that most suffer with spoofing attacks since they can be easily fooled even with common printed photographs obtained in the worldwide network. In this brief a novel approach for face spoofing detection that works with high-level (deep) texture features instead of handcrafted ones is proposed based on a modified Convolutional Neural Network (CNN) [5] by incorporating the LBP (Local Binary Patterns) [6] texture descriptor in its first layer. Besides the good results of LBP itself, CNNs have been increasingly used in such difficult tasks since they can extract and work accurately with high-level features, learned from the own set of training data, being more robust and suitable for activities such as attack detection, in which patterns are complex and can not be easily detected. Experiments show that the proposed deep neural network, called LBPnet, and its extended version, n-LBPnet (normalized LBPNet), outperform the state-of-the-art techniques based on handcrafted texture information, presenting great results in terms of attack detection.

2. TEXTURE-BASED FACE SPOOFING DETECTION

Likewise as in face recognition, texture plays an important role in face spoofing detection [7]. In general, the state-of-the-art antispoofing techniques extract handcrafted texture features from images in order to detect fake faces [7]–[10]. Among the main texture descriptors, the original LBP (Local Binary Patterns) [6] and its variations are the most commonly used due the good results they allow to reach and the efficient algorithm of LBP. Briefly explaining, given a neighborhood system $\{P, R\}$, where P corresponds to the number of neighbors to be considered and R the radius of the neighborhood, the LBP descriptor works by comparing the grayscale value of each pixel p of the image with the intensity of its neighbors and associating a new integer (grayscale) value to p based on such comparisons.

3. CONVOLUTIONAL NEURAL NETWORKS (CNN)

Convolutional Neural Networks (CNN) [5] are deep learning architectures constituted of layers in which different kind of filters (convolution and sampling) are applied to the input data,

initially two-dimensional images. The result of a given layer serves as input to the above one until the top of the network is reached. Differently from the fully connected networks, CNNs present a simplified topology and neurons of same layers use to share parameters, enabling an efficient learning. Besides convolutional and sampling operations, layers with neurons completely connected can be included at the top of the network for classification [12]–[14]. In practice, given a two-dimensional image, in each network layer a set of convolutional filters (kernels) are applied, obtaining different channels of the original input. Pooling, i.e., sampling operations are also performed in order to obtain certain kind of translational and scale invariance and reduce the amount of data being considered. At the top of the network it is obtained a high-level representation of the original image, which is more robust than the raw pixels information for many applications [3].

4. LBP-BASED CONVOLUTIONAL NEURAL NETWORK

Based on the well-referenced Lenet-5 [5] network model, in this brief a novel CNN architecture, called LBPnet, is proposed by integrating the LBP (Local Binary Patterns) [6] descriptor in its first layer in order to extract deep texture features, instead of handcrafted histograms, from images for a more robust face spoofing detection. The first layer of LBPnet incorporates LBP information as follows: the convolution operation actuates not only convolving the values of the kernels (weights of connections between neurons learned in training) with the image grayscale values, but also finding the LBP values of the image pixels before performing the convolution, i.e., the convolution is performed on the transformed LBP values of the pixels and not on their original grayscale values. This improves a lot the results of the proposed deep neural network since the method inherits the power of enhancing face spoofing cues from the LBP descriptor in a deeper and more robust architecture, working with high-level texture features learned from the training data. Fig. 1 shows the convolution of the first layer of LBPnet. The LBPnet presents the following configuration, from bottom to top, mainly inherited from Lenet-5: (i) Two layers with a convolution followed by a pooling operation - the first layer is modified, as said, by incorporating the LBP descriptor in the convolution step; (ii) a Rectified Linear Unit (ReLU) layer, that performs an inner product followed by a rectification (elimination of negative values) on the originated signals; and (iii) a Fully Connected (FC) layer, with two nodes, which also performs an inner product and classification (attack or not attack attempt) of the input image using the softmax function. A scheme of the architecture of LBPnet is shown in Fig. 2. Given a

detected and normalized grayscale facial image (in this brief resized to 66×66), the convolution operation in the first layer, CONV1, finds the pixels LBP-based values and produces 20 outputs with size 60×60 by convolving such values with 20 different kernels with size of 5×5 - each kernel generates an output and is applied with stride of 1 to the image.

5. RESULTS

No-a-days biometric systems are useful in recognizing person's identity but criminals change their appearance in behaviour and psychological to deceive recognition system. To overcome from this problem this paper introduce new technique called Deep Texture Features extraction from images and then building train model using CNN (Convolution Neural Networks) algorithm. This technique refer as LBPNet or NLBPNet as this technique heavily dependent on features extraction using LBP (Local Binary Pattern) algorithm.

In this project we are designing LBP Based Convolution Neural Network called LBPNET to detect face spoofing. Here first we will extract LBP from images and then train LBP descriptor images with Convolution Neural Network to generate training model. Whenever we upload new test image then that test image will be applied on training model to detect whether test image contains spoof image or non-spoof image. Below we can see some details on LBP.

Local binary patterns (LBP) is a type of visual descriptor used for classification in computer vision and is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications. It can be seen as a unifying approach to the traditionally divergent statistical and structural models of texture analysis. Perhaps the most important property of the LBP operator in real-world applications is its robustness to monotonic gray-scale changes caused, for example, by illumination variations. Another important property is its computational simplicity, which makes it possible to analyze images in challenging real-time settings.

The LBP feature vector, in its simplest form, is created in the following manner:

Divide the examined window into cells (e.g. 16x16 pixels for each cell).

For each pixel in a cell, compare the pixel to each of its 8 neighbors (on its left-top, left-middle, left-bottom, right-top, etc.). Follow the pixels along a circle, i.e. clockwise or counter-clockwise.

Where the center pixel's value is greater than the neighbor's value, write "0". Otherwise, write "1". This gives an 8-digit binary number (which is usually converted to decimal for convenience).

Compute the histogram, over the cell, of the frequency of each "number" occurring (i.e., each combination of which pixels are smaller and which are greater than the center). This histogram can be seen as a 256-dimensional feature vector.

Optionally normalize the histogram.

Concatenate (normalized) histograms of all cells. This gives a feature vector for the entire window.

The feature vector can now be processed using the Support vector machine, extreme learning machines, or some other machine learning algorithm to classify images. Such classifiers can be used for face recognition or texture analysis.

A useful extension to the original operator is the so-called uniform pattern,[8] which can be used to reduce the length of the feature vector and implement a simple rotation invariant descriptor. This idea is motivated by the fact that some binary patterns occur more commonly in texture images than others. A local binary pattern is called uniform if the binary pattern contains at most two 0-1 or 1-0 transitions. For example, 00010000 (2 transitions) is a uniform pattern, but 01010100 (6 transitions) is not. In the computation of the LBP histogram, the histogram has a separate bin for every uniform pattern, and all non-uniform patterns are assigned to a single bin. Using uniform patterns, the length of the feature vector for a single cell reduces from 256 to 59. The 58 uniform binary patterns correspond to the integers 0, 1, 2, 3, 4, 6, 7, 8, 12, 14, 15, 16, 24, 28, 30, 31, 32, 48, 56, 60, 62, 63, 64, 96, 112, 120, 124, 126, 127, 128, 129, 131, 135, 143, 159,

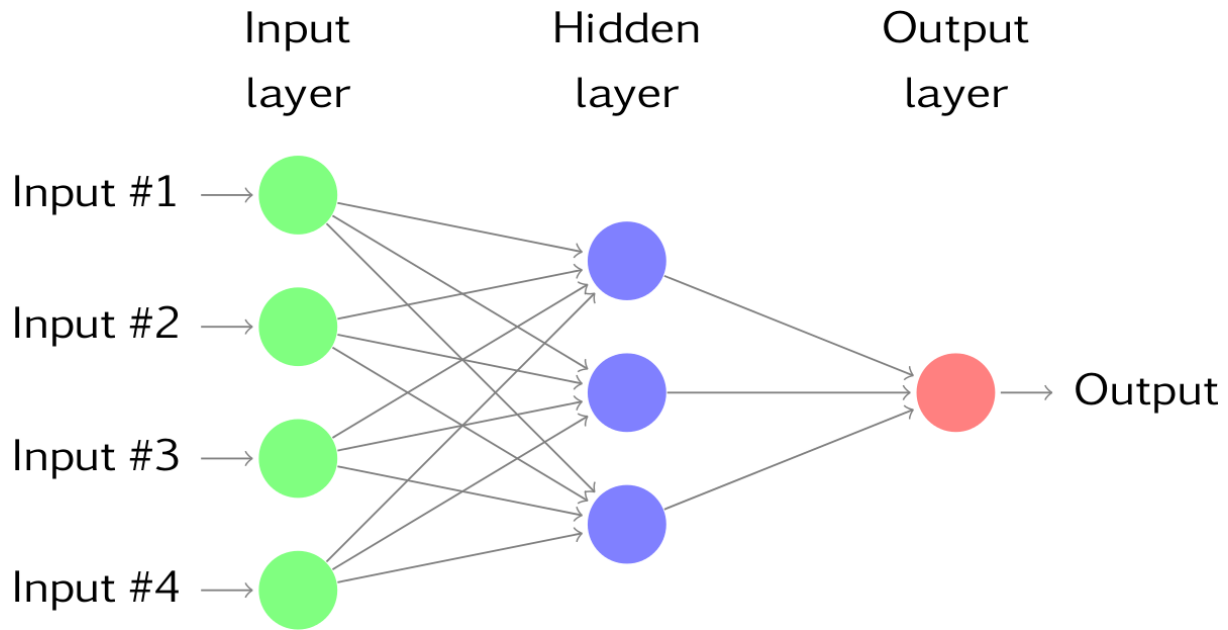
191, 192, 193, 195, 199, 207, 223, 224, 225, 227, 231, 239, 240, 241, 243, 247, 248, 249, 251, 252, 253, 254 and 255.

CNN working procedure

To demonstrate how to build a convolutional neural network based image classifier, we shall build a 6 layer neural network that will identify and separate one image from other. This network that we shall build is a very small network that we can run on a CPU as well. Traditional neural networks that are very good at doing image classification have many more parameters and take a lot of time if trained on normal CPU. However, our objective is to show how to build a real-world convolutional neural network using TENSORFLOW.

Neural Networks are essentially mathematical models to solve an optimization problem. They are made of neurons, the basic computation unit of neural networks. A neuron takes an input (say x), do some computation on it (say: multiply it with a variable w and adds another variable b) to produce a value (say; $z = wx + b$). This value is passed to a non-linear function called activation function (f) to produce the final output (activation) of a neuron. There are many kinds of activation functions. One of the popular activation function is Sigmoid. The neuron which uses sigmoid function as an activation function will be called sigmoid neuron. Depending on the activation functions, neurons are named and there are many kinds of them like RELU, TanH.

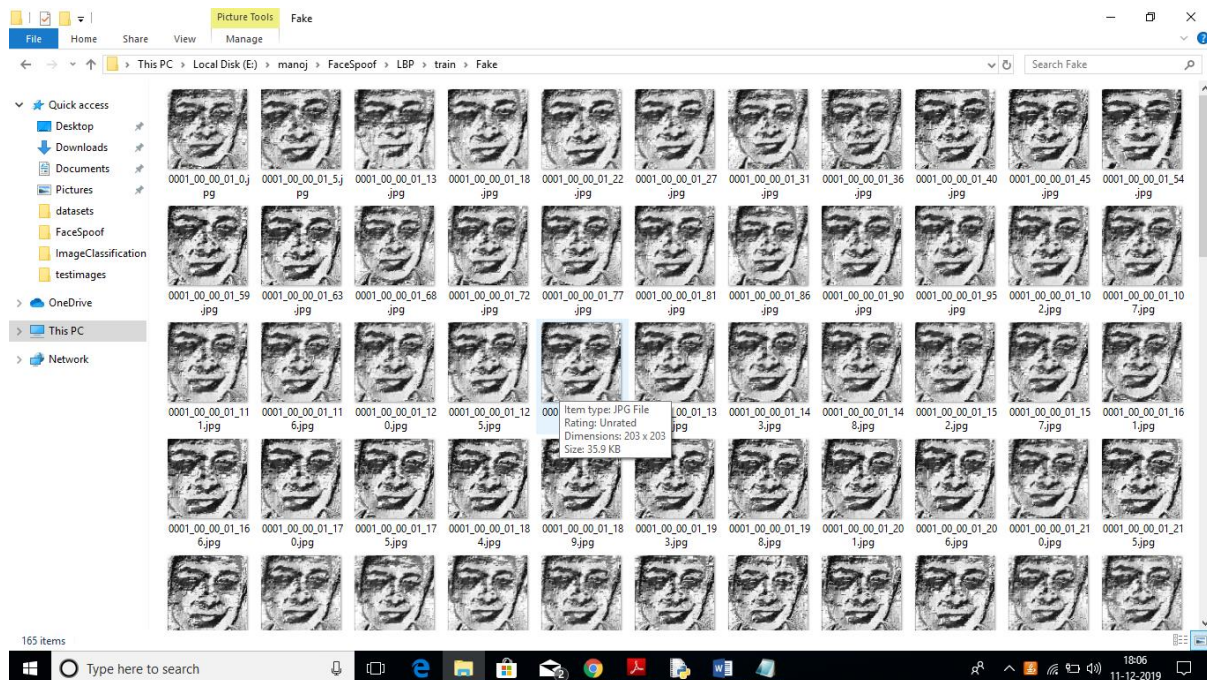
If you stack neurons in a single line, it's called a layer; which is the next building block of neural networks. See below image with layers



To predict image class multiple layers operate on each other to get best match layer and this process continues till no more improvement left.

Dataset Details:

In this paper author has used NUAA Photograph Imposter (fake) Database with images obtained from real and fake faces. We also used images and convert that image into LBP format. Below are some images from LBP folder



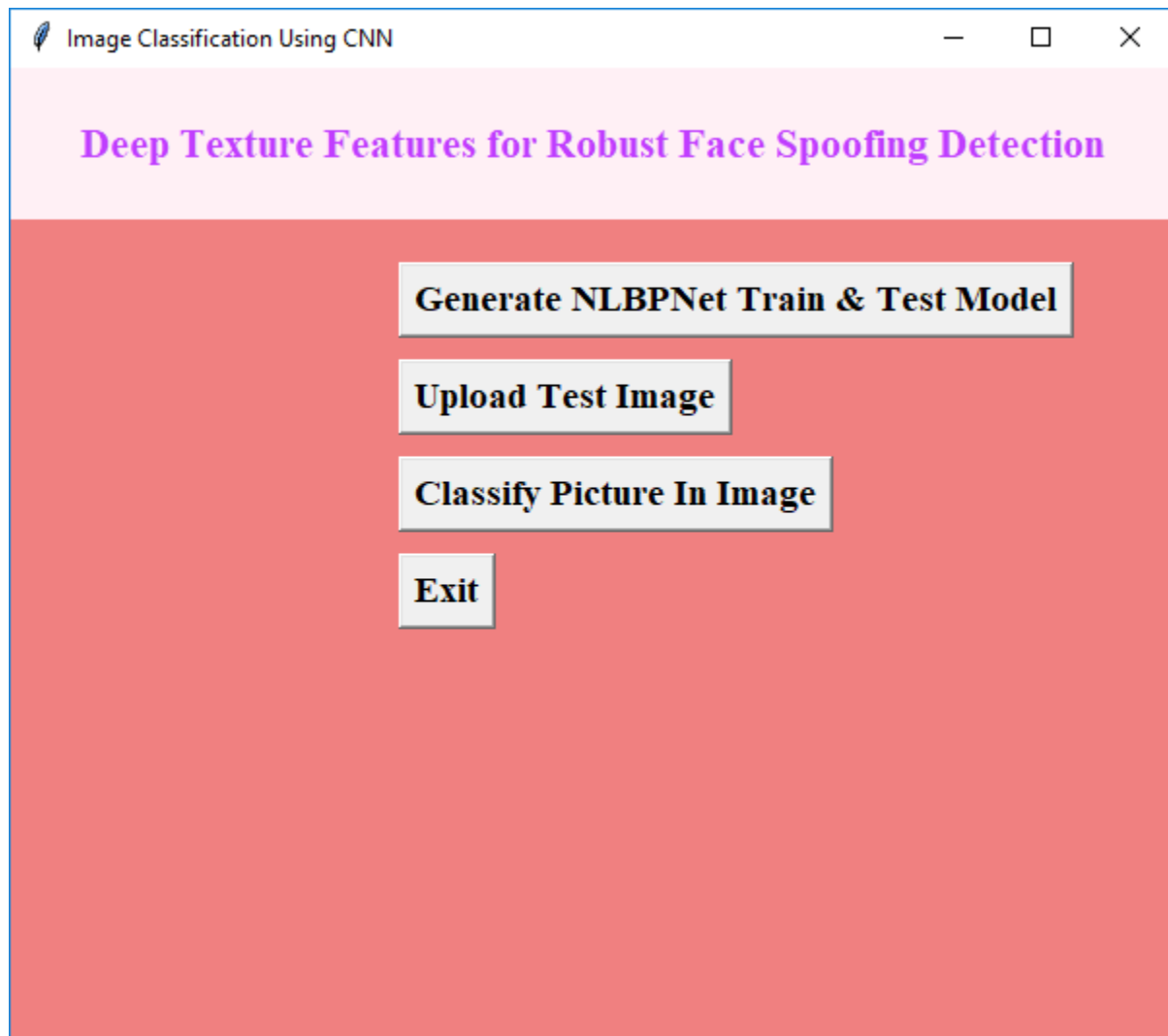
All this fake and real images you can see inside 'LBP/train' folder.

This project consists of following modules:

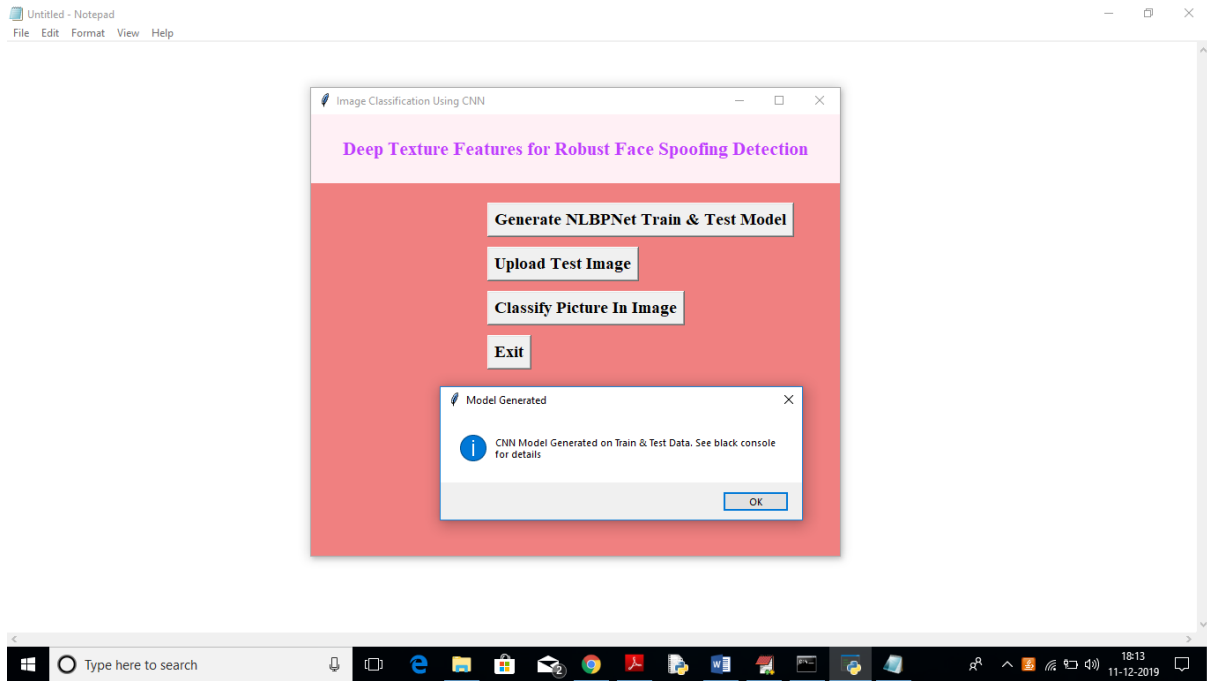
- 1) Generate NLBPNet Train & Test Model: in this module we will read all LBP images from LBP folder and then train CNN model with all those images.
- 2) Upload Test Image: In this module we will upload test image from 'testimages' folder. Application will read this image and then extract Deep Textures Features from this image using LBP algorithm.
- 3) Classify Picture In Image: This module apply test image on CNN train model to predict whether test image contains spooF or non-spooF face.

Screen shots

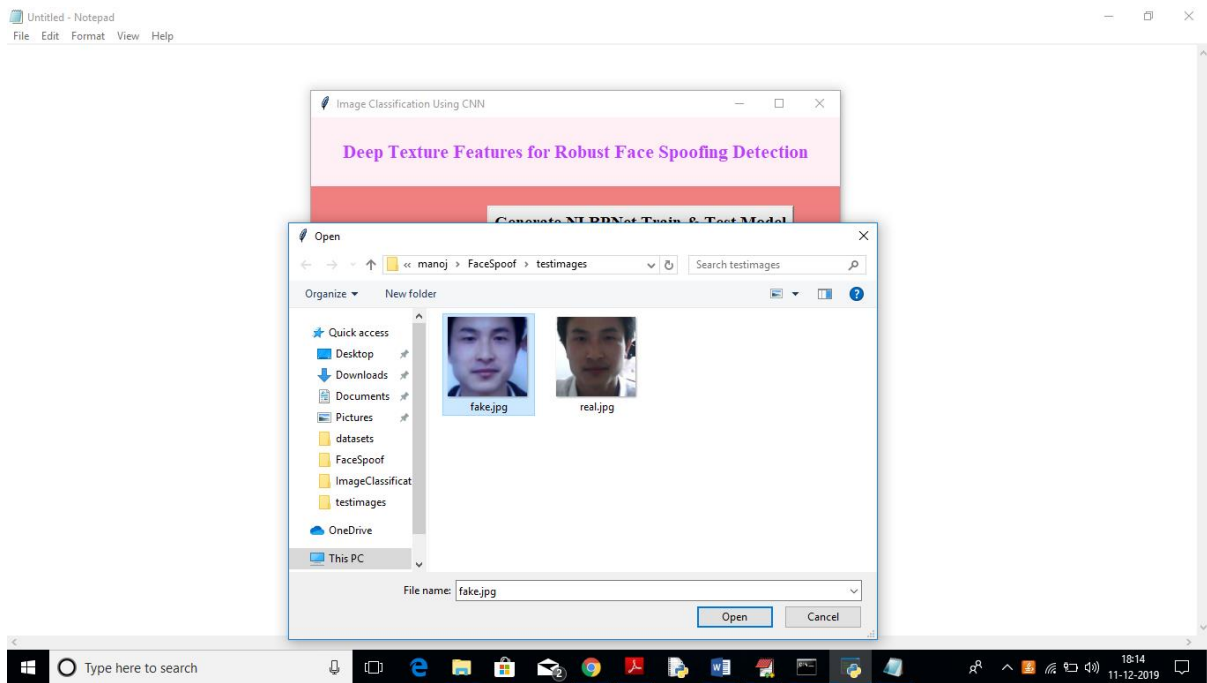
To run this project double click on 'run.bat' file to get below screen



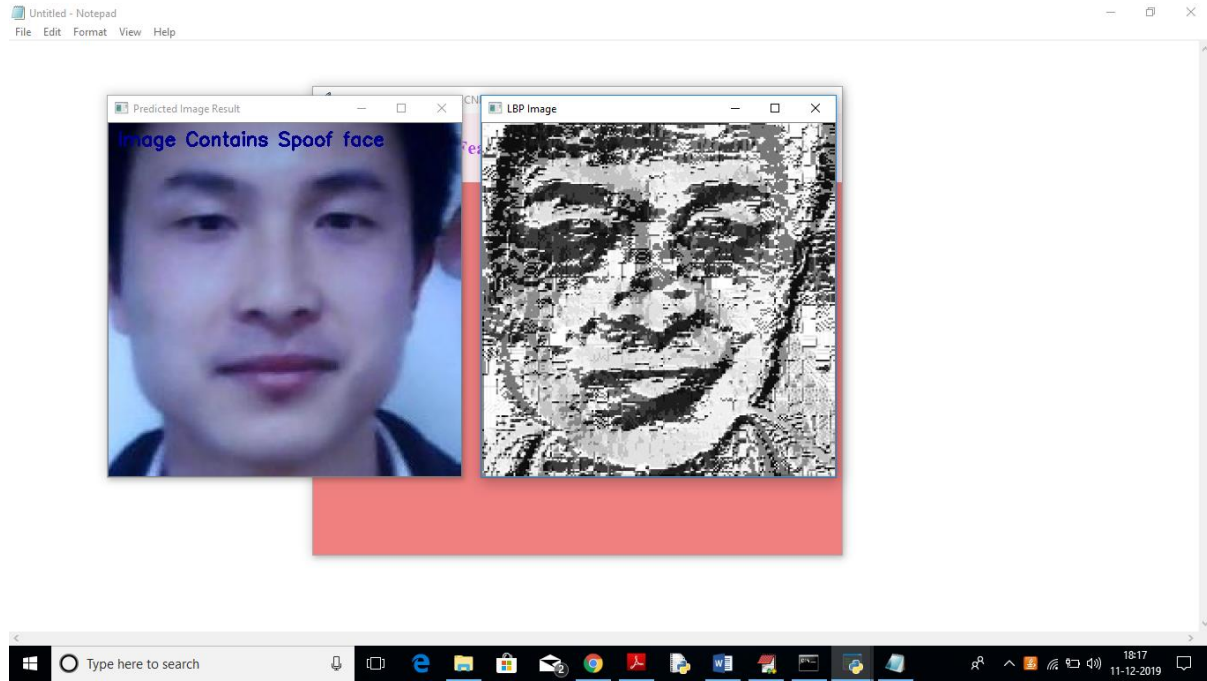
In above screen click on 'Generate NLBPNet Train & Test Model' button to generate CNN model using LBP images contains inside LBP folder.



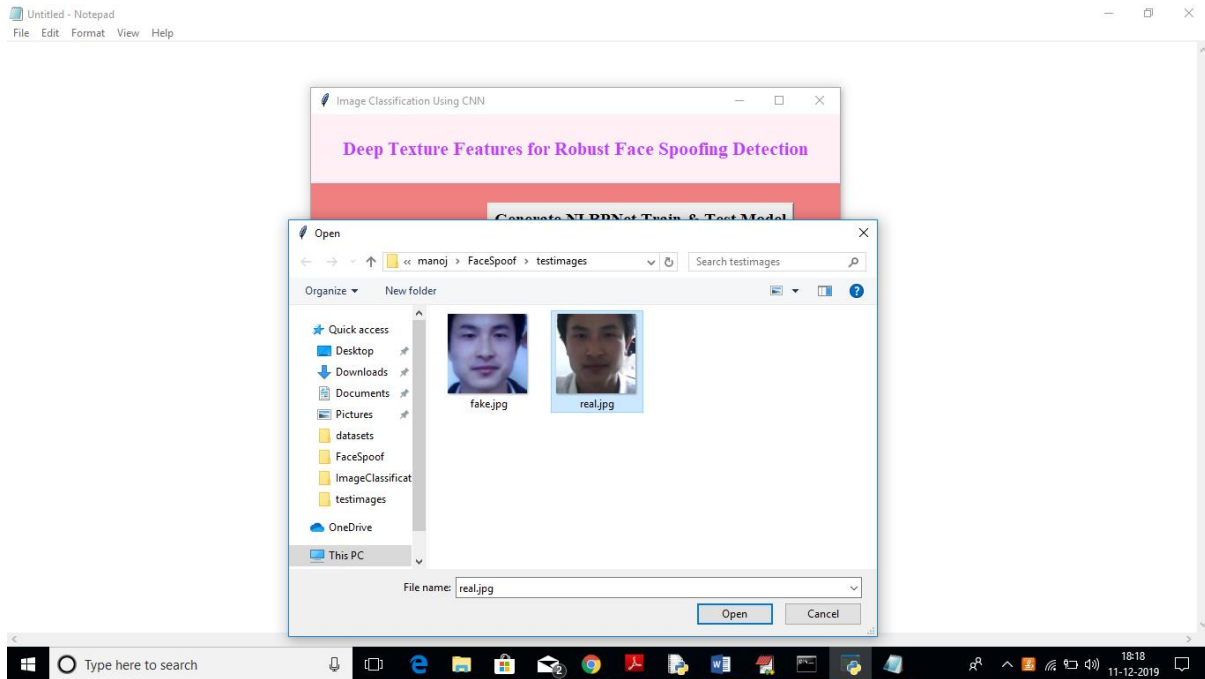
In above screen we can see CNN LBPNET model generated. Now click on ‘Upload Test Image’ button to upload test image



In above screen we can see two faces are there from same person but in different appearances. For simplicity I gave image name as fake and real to test whether application can detect it or not. In above screen I am uploading fake image and then click on 'Classify Picture In Image' button to get below result



In above screen application display message on image as it contains spoof face and I am displaying LBP format image also. Now we I will upload real image



6. CONCLUSION

In this brief, two LBP-based Convolutional Neural Networks, LBPnet and n-LBPnet, are proposed for spoofing detection in face recognition systems, which presented great results on the NUAA spoofing dataset, outperforming other assessed state-of-the-art techniques. With the highest ROC curves, low EER as well as high accuracy, the proposed LBPnet and n-LBPnet networks configure effective alternatives for spoofing detection in real face recognition applications of nowadays. Besides of presenting great results, the proposed methods are more efficient than other state-of-the-art techniques that combine lots of handcrafted information to detect attacks. Our approaches use the LBP descriptor with a single neighborhood, a forward bottom-up pass and simple softmax neurons at the top for detecting spoofing attempts quickly, being more suitable for real time applications. Based on all this it is possible to conclude that deep texture features are rich sources of information for face spoofing detection, propiciating better results than handcrafted ones (or even combination of them, which may become impractical). The integration of the LBP descriptor in a deep learning architecture is a suitable and robust alternative to prevent such criminal activities.

7. REFERENCES

- [1] A. K. Jain et al., “Biometrics: A grand challenge,” in Proc. Int. Conf. Pattern Recognit., Cambridge, U.K., 2004, pp. 935–942.
- [2] M. Fons, F. Fons, and E. Cantó, “Fingerprint image processing acceleration through run-time reconfigurable hardware,” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 57, no. 12, pp. 991–995, Dec. 2010.
- [3] D. Menotti et al., “Deep representations for iris, face, and fingerprint spoofing detection,” *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 864–879, Apr. 2015.
- [4] N. K. Ratha, J. H. Connell, and R. M. Bolle, “An analysis of minutiae matching strength,” in Proc. Int. Conf. Audio Video Based Biometric Person Authentication, Halmstad, Sweden, 2001, pp. 223–228.
- [5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [6] T. Ojala, M. Pietikäinen, and T. Mäenpää, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
- [7] J. Määttä, A. Hadid, and M. Pietikäinen, “Face spoofing detection from single images using micro-texture analysis,” in Proc. Int. Joint Conf. Biometrics, Washington, DC, USA, 2011, pp. 1–7.
- [8] S. Parveen et al., “Face liveness detection using dynamic local ternary pattern (DLTP),” *Computers*, vol. 5, no. 2, p. 10, 2016.
- [9] S. R. Arashloo, J. Kittler, and W. Christmas, “Face spoofing detection based on multiple descriptor fusion using multiscale dynamic binarized statistical image features,” *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 11, pp. 2396–2407, Nov. 2015.
- [10] J. Yang, Z. Lei, S. Liao, and S. Z. Li, “Face liveness detection with component dependent descriptor,” in Proc. Int. Conf. Biometrics, 2013, pp. 1–6.

A System for Real Time Botnet Command and Control Traffic Detection

YANDAPALLI NAGA MAHALAKSHMI

PG Scholar, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

S.K.ALISHA

Associate Professor, Department of M.C.A,
B.V.Raju College,
Bhimavaram, W.G.Dt., A.P, India

Abstract

Over the past decade, the digitization of services transformed the healthcare sector leading to a sharp rise in cyber security threats. Poor cyber security in the healthcare sector, coupled with high value of patient records attracted the attention of hackers. Sophisticated advanced persistent threats and malware have significantly contributed to increasing risks to the health sector. Many recent attacks are attributed to the spread of malicious software, e.g., ransom ware or bot malware. Machines infected with bot malware can be used as tools for remote attack or even crypto mining. This paper presents a novel approach, called BotDet, for botnet Command and Control (C&C) traffic detection to defend against malware attacks in critical ultra structure systems. There are two stages in the development of the proposed system: 1) we have developed four detection modules to detect different possible techniques used in botnet C&C communications and 2) we have designed a correlation framework to reduce the rate of false alarms raised by individual detection modules. Evaluation results show that BotDet balances the true positive rate and the false positive rate with 82.3% and 13.6%, respectively. Furthermore, it proves BotDet capability of real time detection.

Keyword: Botnet, Malware, Medical services, Real-time systems, Correlation, Monitoring, Command and control systems

1. INTRODUCTION

1.1 Introduction:

Country's national security, economic vitality and daily life rely on a safe, stable, and resilient cyberspace. This work depends on this vast array of networks to provide healthcare services, transport and communication, power our homes and run our economy [1]. Over the last decade, cyber attacks and intrusions have increased substantially, disrupting critical operations, resulting in business downtime and exposing sensitive personal and business information. Statistics draw a grim picture about the cyber security challenges and digital risks in the healthcare industry.

A report by the US Department of Health and Human Services [2] reveals that the healthcare sector has suffered from approximately four data breaches a week in 2016. To put this into perspective, one in every three American citizens was a victim of a breach in the healthcare sector. One of the primary reasons behind targeting healthcare organizations is that these organizations do not set protecting patient data as a priority, hence they under invest in qualified IT security personnel. The lack of solid information security infrastructure makes healthcare organizations an easy target. For instance, the recent attack on the National Health Service (NHS) in the UK

showed that some hospitals and care providers systems were obsolete or has not been patched against well-known vulnerabilities. Additionally, patient records contain a wealth of information that can be used for identifying theft, financial/insurance fraud and even blackmailing. In 2017, 15,000 medical records have been stolen from Beverly Hills plastic surgery clinic to bully several high-profile celebrities. Today, intelligence agencies and governments military are actively preparing for cyber warfare.

Global activities against software, hardware, or data are referred to as cyber attack in the field of computer networks or systems. These activities lead to degrading, disrupting, destroying or denying access to network/system services or resources. Activities that target gathering intelligence are referred to as cyber exploitation [3]. The main objective of these activities is to gain unauthorized access to information and data. Over the last decade, malicious software or malware has increased, particularly in the healthcare industry. They have become one of the main reasons for the majority of the (distributed) denial-of-service (Dos) activities [4], direct and scanning attacks [5]. Noticeably, the motivation from fame seeking and curiosity has been shifted to unlawful financial attainment, which resulted in the sophistication of malicious software. Moreover, the availability of easy to-use toolkits to build malware will probably keep these malwares a threat to individuals, business and governments in the foreseeable future. Generally, there are two classes of malware: (a) malware that targets the general population and (b) customized information-stealing malware that targets particular organizations such as healthcare providers.

Zombies, which refer to those machines infected with bot malware, can be

used as tools for remote attack or can be part of a botnet, which is completely controlled by the botnet master. Bots are "enslaved" host computers in botnets (networks formed by bots). One or more botmasters control bots in botnets and the intention is to perform malicious activities. The essential goal of botnets is to control organized crime syndicate, criminal, or group of criminals to use compromised machines for performing illegal activities. Experts mention that about 16–25% of the machines connected to the Internet are parts of botnets. Bots are different from the other malware. They are capable to create Command and Control (C&C) channels. Bots recognize themselves by their C&C channels through which they can be controlled, updated and instructed. The C&C servers are usually machines that have been exploited and sorted in a distributed form to limit traceability.

The detection of botnet C&C traffic is challenging for current Intrusion Detection Systems (IDS) for several reasons: (1) it is a benign traffic and follows normal protocol usage; (2) their volume of traffic is small; (3) the number of bots may be very small in the monitored network; and (4) Bots' communications may be encrypted. This work aims to contribute to IDS research, particularly to botnet C&C traffic detection. The proposed approach, called BotDet, undergoes two main phases. The first phase runs various modules to detect different possible techniques used in botnet C&C communications. The second phase uses a framework for alert correlation to reduce the number of false positives.

1.2 Purpose:

Intelligence agencies and governments military are actively preparing for cyber warfare. Global activities against software, hardware, or data are referred to as cyber attack in the field of computer networks or systems. These activities lead to degrading,

disrupting, destroying or denying access to network/system service so sources. Activities that target gathering intelligent are referred to as cyber exploitation. The main objective of these activities is to gain unauthorized access to information and data.

1.3 Scope:

Statistics draw a grim picture about the cyber security challenges and digital risks in the healthcare industry. A report by the US Department of Health and Human Services [2] reveals that the healthcare sector has suffered from approximately four data breaches a week in 2016. To put this into perspective, one in every three American citizens was a victim of a breach in the healthcare sector. One of the primary reasons behind targeting healthcare organizations is that these organizations do not set protecting patient data as a priority, hence they under invest in quailed IT security personnel.

1.4 Motivation:

The motivation from fame seeking and curiosity has been shifted to unlawful financial attainment, which resulted in the sophistication of malicious software [7]. Moreover, the availability of easy to-use toolkits to build malware will probably keep these malwares a threat to individuals, business and governments in the foreseeable future.

1.5 Overview:

An approach for bot-infected machines detection which requires no previous knowledge of the way a bot spreads. It depends on the characteristic behavior of a bot, particularly: (a) receiving commands from the botmaster, and (b) responding to these commands by carrying out some activities. Both commands and responses can be monitored in the network traffic and detection models can be built. The authors ran a bot in a controlled

network to record its traffic and then they examine the received commands and responses activities. For this purpose, they proposed techniques to determine points in the network that were involved in the response activity. Afterwards, the traffic had been observed before this response is analyzed to find the corresponding command. By these detection models the network traffic is scanned for similar actions aiming to detect bot-infected machines.

2. RELATED WORKS

In [6], Balram and Wilscoy propose a host-based approach for botnet C&C communication detection. This approach analyses suspicious flows produced by filtering out benign traffic from the traffic created by a host. A normal profile of the host traffic is used for the filtering. The behavioral pattern of flows to all destinations is examined in a bid to generate the host profile. This approach achieved a detection rate of 100% and false positives of 8%.

In [7], Fedynyshyn et al. present a host-based detection method able to detect the existence of botnet C&C traffic on the observed machine, and also categorize the type of C&C communication used by the bot, e.g., peer-to-peer (P2P) based, HTTP-based or IRC-based. As it does not examine the packets payloads, their detection method is independent of the content of the C&C messages. Their method for detecting and categorizing botnet C&C connections is based on three hypotheses: (1) it is possible to distinguish between botnet C&C communication and botnet non-C&C communication, (2) it is possible to distinguish between botnet C&C communication and valid communication and (3) there are shared characteristics between different styles of C&C and different botnet families.

An approach for bot-infected machines detection was presented by Wurzinger et al. [8], which require no previous knowledge of the way a bot spreads. It depends on the characteristic behavior of a bot, particularly: (a) receiving commands from the botmaster, and (b) responding to these commands by carrying out some activities. Both commands and responses can be monitored in the network traffic and detection models can be built. The authors ran a bot in a controlled network to record its traffic and then they examine the received commands and responses activities. For this purpose, they proposed techniques to determine points in the network that were involved in the response activity. Afterwards, the traffic had been observed before this response is analyzed to find the corresponding command. By these detection models the network traffic is scanned for similar actions aiming to detect bot-infected machines.

Giroire et al. [9] presented another host-based detection method for botnet C&C traffic detection. This method is based on the fact that the infected machines should stay in contact with C&C servers to be instructed and controlled by the botmaster. It is assumed that those connections are persistent and established regularly. A white-list of benign destinations that the user regularly contacts is built and all the user outbound traffic is monitored. When a connection is persistent enough and the destination is not white-listed, an alert is generated and the user is informed and asked to decide. If the destination is legitimate, the user can easily add it to the white-list; otherwise the connection is deemed as C&C communication and blocked.

A network-based botnet detection system, BotSniffer, was proposed in [10]. This system is based on anomalybased detection

algorithms to detect both HTTP and IRC based C&Cs with no previous knowledge of C&C server addresses or signatures. The main goal in BotSniffer is to identify spatial-temporal similarity patterns and correlation in network traffic that are generated between the infected hosts and botnet C&C servers. They study two common styles usually used for botnet control, "push" and "pull". An example for the push style is IRC-based C&C is where the commands are sent or pushed to the infected hosts. In the pull style, the commands are downloaded (or pulled) by the infected hosts, as in HTTP-based C&C. When a set of hosts is found to carry out the same actions in response to similar messages from the same server, it is considered to be part of a botnet

3. EXISTING SYSTEM

There are two main approaches for botnet C&C traffic detection in the existing systems. The first one is based on setting up honey nets in the network. This approach is often used to understand and analyze a botnet technology and characteristics. However, honeynets are not always capable of detecting bot infection. The second approach is based on passive traffic monitoring. These approaches can be classified into signature-based and anomaly-based methods, respectively. Signature-based detection methods make use of known signatures and behavior of existing botnets, therefore it can be used for detecting only known botnets. Anomaly-based detection methods are able to detect unknown botnets as they try to detect botnets based on network traffic anomalies like traffic on unusual ports, high volumes of traffic, unusual system behavior and high network latency. Balram and Wilscy propose a host-based approach for botnet C&C communication detection. Fedynyshyn et al. present a host-based detection method able to detect the existence of botnet C&C traffic on the observed machine, and also

categorize the type of C&C communication used by the bot, e.g., peer-to-peer (P2P) based, HTTP-based or IRC-based.

3.1 Disadvantages:

These existing works cannot analyze and detects hidden botnet C&C. Botnet C&C traffic is cannot detect by the observation of direct causes of traffic flows. These works cannot reduce the rate of false alarms raised by individual detection modules.

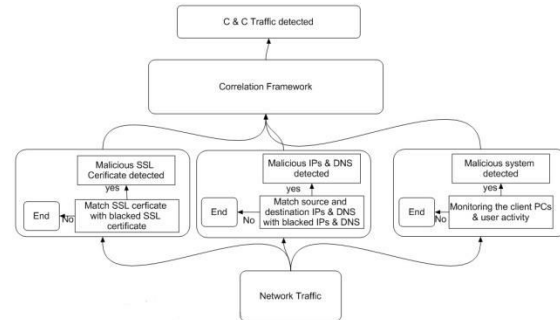
4. PROPOSED SYSTEM

In this project focused on network-based detection and host based detection of bots in Internet-connected networks with regard to the botnet threat and botnet detection. Invisibility is an important factor in botnet survivability; fortunately the invisibility of a botnet has practical limitations. Important causes that limit the invisibility are attack traffic, malware installation, limited resources and other survivability measures. The proposed work based on three major detection methods such as Untrusted Destination by Identifier (UDI), malicious SSL certificate, Traffic Flow Causality (TFC) and it is used to analyze and detect the malicious network traffic in real time.

4.1 Advantages

The framework that analyzes and detects hidden botnet C&C. Botnet C&C traffic is detected by the observation of direct causes of traffic flows. To reduce the rate of false alarms raised by individual detection modules

5. ARCHITECTURE

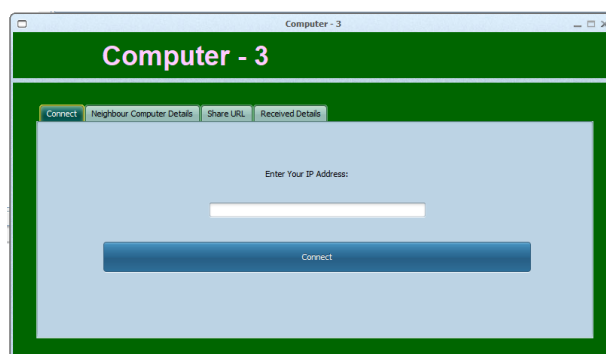
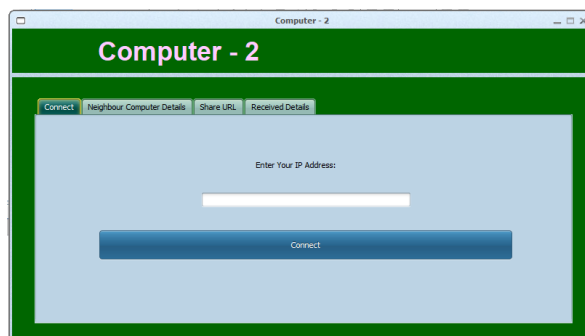
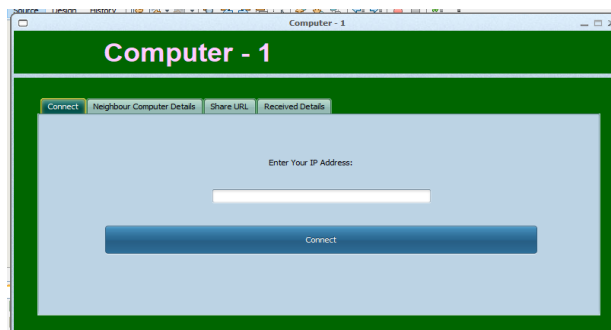
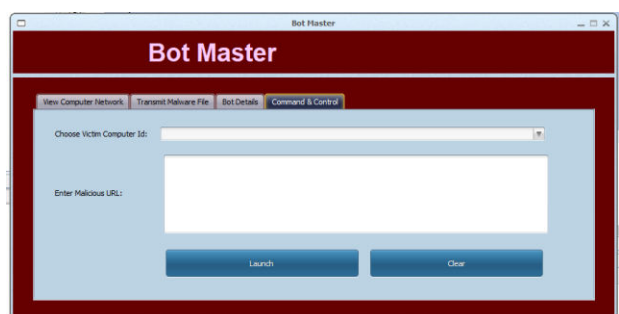
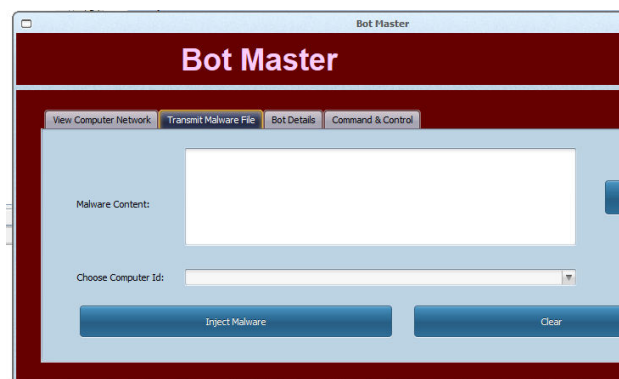
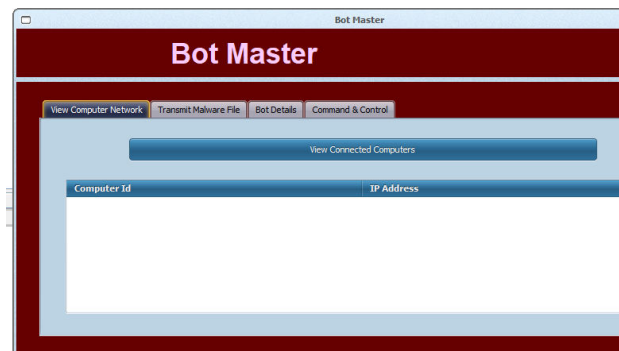
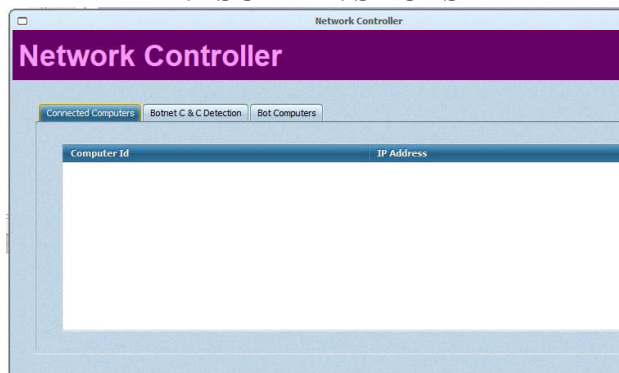


6. Modules Description:

Botnet C & C detection by SSL certificate (https □ is valid or not). Botnet C & C detection by IP Address.. Botnet C & C detection by DNS. (Detect Malicious URL or not). Botnet C & C detection by causal analysis of traffic flows.

Our proposed approach for botnet C&C traffic detection is outlined. This approach is based on the correlation between the events, which are the outputs of the detection modules. The proposed approach consists of two main phase's communication. To this end, three detection modules have been proposed: Botnet C & C by SSL certificate detection module, Botnet C & C by untrusted destinations detection module and Botnet C & C by causal analysis of traffic flows detection module each detection module is independent of the other modules and aims to detect one technique that can be used in C&C communication. The outputs of these detection modules should be submitted to the second phase where they are correlated to raise an alert and block on botnet C&C traffic detection. In the second phase, the correlation framework takes events (the outputs of our detection modules) as an input and correlates them to raise an alert and block on botnet C&C traffic detection. The correlation method is based on voting between the detection methods to make the final decision about the detection.

7. SCREEN SHOTS



8 CONCLUSION AND FUTURE ENHANCEMENTS

This work presents a novel approach called BotDet for botnet C&C traffic detection. The developed system (BotDet) runs through two main phases, the first one includes developed modules to detect possible techniques used in botnet C&C communications. The second phase uses a framework for alert correlation, based on voting between the detection modules. BotDet achieves detection rate and false alarm of 82:3% and 13:6% respectively. Additionally, the blacklists used in some of the detection modules are automatically updated based on different intelligent feeds,

which gives BotDet the capability of real time detection

9. BIBLIOGRAPHY

[1] S. Belguith, N. Kaaniche, A. Jemai, M. Laurent, and R. Attia, "PAbAC: A privacy preserving attribute based framework for fine grained access control in clouds," in Proc. 13th Int. Joint Conf. e-Bus. Telecommun., 2016, pp. 133–146.

[2] US Department of Health and Human Services Report. Accessed: Jan. 7, 2018.

[Online]. Available:
https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf

[3] P. J. Denning and D. E. Denning, "Discussing cyber attack," Commun. ACM, vol. 53, no. 9, pp. 29–31, 2010.

[4] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring internet denial-of-service activity," ACM Trans. Comput. Syst., vol. 24, no. 2, pp. 115–139, 2006.

[5] K. G. Anagnostakis, S. Sidiroglou, P. Akritidis, K. Xinidis, E. P. Markatos, and A. D. Keromytis, "Detecting targeted attacks using shadow honeypots," in Usenix Secur., 2005.

[6] S. Balram and M. Wilscy, "User traffic profile for traffic reduction and effective botnet C&C detection," IJ Netw. Secur., vol. 16, no. 1, pp. 46–52, 2014.

[7] G. Fedynyshyn, M. C. Chuah, and G. Tan, "Detection and classification of different botnet C&C channels," in Autonomic and Trusted Computing (Lecture Notes in Computer Science), vol. 6906, J. M. A. Calero, L. T. Yang, F. G. MÆrmol, L. J. G. Villalba, A. X. Li, and Y. Wang, Eds. Berlin, Germany: Springer, 2011

[8] P. Wurzinger, L. Bilge, T. Holz, J. Goebel, C. Kruegel, and E. Kirda, "Automatically generating models for botnet detection," in Computer Security—ESORICS (Lecture Notes in Computer Science), vol. 5789, M. Backes and P. Ning, Eds. Berlin, Germany: Springer, 2009.

[9] F. Giroire, J. Chandrashekar, N. Taft, E. Schooler, and D. Papagiannaki, "Exploiting temporal persistence to detect covert botnet channels," in Recent Advances Intrusion Detection (Lecture Notes in Computer Science), vol. 5758, E. Kirda, S. Jha, and D. Balzarotti, Eds. Berlin, Germany: Springer, 2009.

[10] G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting botnet command and control channels in network traffic," in Proc. 15th Annu. Netw. Distrib. Syst. Secur. Symp. Dayton, OH, USA: Wright State Univ., 2008.

MITIGATING DDOS ATTACK IN IOT NETWORK ENVIRONMENT

Yandrapu Krishnamouli (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram,
West Godavari District, Andhra Pradesh, India, 534202.

S. K. Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra
Pradesh, India, 534202.

Abstract Nowadays, there are lots of Internet of Things (IoT) devices are used in Industry, Home appliances, Automobile industry and many more places. Issues regarding security of the IoT are the primary reason why it fails to attract more people. Each day, beside the new technology comes a millions of vulnerabilities waiting to be exploited. IoT is that the latest trend and like all technology, it's open for exploitation. In IoT environment, Distributed Denial of service attack (DDoS) could be a major issue, because of the limited computing and power resources of standard IoT devices are prioritized in implementing functionality instead of security features. DDoS attack is the most common attack which is used to bring down the whole network without having any loophole in the network security. The main purpose of this work is to mitigate DDoS attacks against the IoT using honeypot model in Raspberry Pi. Honeypots are the network setup with intentional loopholes. The purpose of honeypot is to invite attackers, so the activities and methods used to attack can be studied and it can also help to increase network security. Conclusion came out after reviewing some research papers that there are lots of DDoS attack mitigation techniques are proposed by many researchers, but very few are proposed for IoT environment. Primary focus of this dissertation work is to proposed mitigation techniques of DDoS attack on IoT device using honeypot and IDS with low cost and high performance resources.

Keyword: Internet of Things, Denial of Service Attack, Distributed Denial of Service Attack, IoT security, Security Breach

1. INDRODUCTION

1.1 Internet of things

Internet of Things is not another word now-a-days for anybody in light of the fact that everything now going to be accessed by means of Internet. The word IoT defined by Wikipedia as, "The Internet of things (IoT) is the network of physical devices, vehicles, and other items embedded with electronics, software, sensors, actuators, and network connectivity which enable these objects to collect and exchange data." [1] The "thing" in the internet of thing can be a person

with a smart watch, a farm with some sensors, car that has built-in sensors to notify the driver when any object near the car or any other devices that has IP address for connecting to the network for the transfer of the data. Internet of Things (IoT) speaks to a general idea for the adaptability of system gadgets to detect and gather data from the globe around us, at that point share that information over the web where it will be handled and used for various interesting purposes. These days some utilization the term Industrial Internet conversely with IoT (IIoT). This

alludes basically to business uses of IoT innovation in the realm of manufacturing.

1.2 Dos/DDos Attack

A denial of service (DoS) attack take effect once a service that might usually work is inaccessible. There may be many reasons for inaccessibility, however it always refers to infrastructure that can't cope because of capability overload. During a Distributed Denial of Service (DDoS) attack, an oversized range of systems maliciously attack on one target system or network. This attack can be often perform through a botnet, where there are lots of devices are preprogrammed to request a particular service at same time.

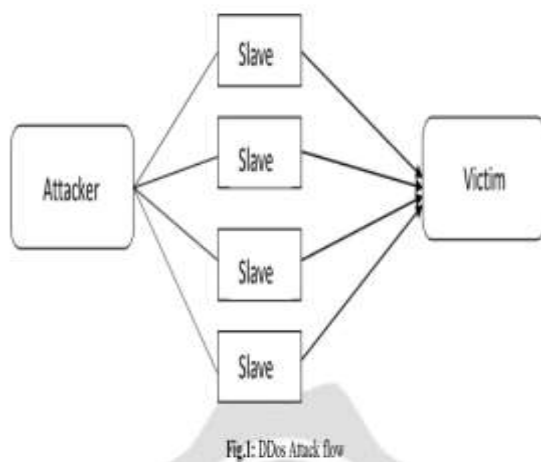


Fig.1: DDoS Attack flow

In fig.1 general DDoS attack flow is shown where attacker use slave systems as botnet to perform attacks like send flood packets into victim system to consume resources and network bandwidth. Nowadays people are getting used to IoT devices i.e. smart watches, smart phone, smart refrigerator, etc. As increasing usage of IoT in their daily life number of devices are increased day by day, so attacks related to IoT became major concern. Above described attacks are common attacks happened in IoT environments. Among them DDoS

attack would be a very danger attack because of its characteristic take benefits of limited computing power of IoT device. DDoS attack made device unavailable or irresponsible. On the cusp of 2017, one thing's clear: distributed denial-of-service (DDoS) attacks created their mark in 2016. Arbor Networks half-track 124,000 DDoS attacks every week between Jan 2015 and June 2016. Moreover, 274 of the attacks determined within the first half of 2016 reached over 100 Gbps (as compared to 223 in all of 2015), whereas 46 attacks registered higher than 200 Gbps (as compared to 16 in 2015). Together, those campaigns' peak attack size inflated by 73 % to 579 Gbps. 1.3

CLASSIFICATION OF DDOS ATTACK

- 1) UDP flood
- 2) ICMP/PING flood
- 3) SYN flood
- 4) Ping of Death
- 5) DNS amplification

2. RELATED WORK

Authors in paper [2] gave an equipment based watermarking checking framework technology to shield organizations from these attacks. This techniques utilizes trust investigation of the incoming packets using trace-back methods. In this procedure just the trusted packets are permitted inside the network. Authors in paper [3] introduced an intrusion detection system that uses a layered model integrated with neural network. They proposed two models in particular A and B where model A considers all features of the practice dataset and B considers features adding to the order procedure.



This proposed framework detects four regular types of attacks like DOS, Remote to local (R2L), User to root (U2R) and ordinary records. This framework used the KDD 1999 database with a specific end goal to accomplish accurate results. Further, this approach other than detecting wide variety of attacks additionally has less false alarm rate. Paper [4] gives solution for DDoS mitigation using software defined network. This paper gives solution free from the limitation of proprietary software of routers. Here author presents approach for anomaly detection using SDN infrastructure in which collection of traffic data flow information which is maintained on all the SDN enabled switches placed on network. This method successfully achieve high detection accuracy. This mitigation technique need some future work of sharing in-line sampling based ADSs in an efficient way to overcome burden of growing IP traffic and limited computational resources. Author in [5] presents detection and mitigation of DDoS attack methods which are distinguish by various stages. All the stages are capable to filter malicious users of DDoS attack. Stages are named as restriction of user access, limitation of traffic rate and CAPTCHA verification technique. In Restriction of access, Blacklisting of IP address is used as concept. In Limitation of rate and Captcha verification stage, reducing the rate of http connection bound the same IPs accessing with the same object in the server. In paper [6] author proposed system in which the Dendric Cell Algorithm (DCA) continuous check the traffic and compare the SYN packet and SYN-ACK packet ratio. If the ratio is higher than median value, it means there is

lot of SYN packets are incoming and very little SYN_ACK packet. Like that TCP SYN flood attack is detected. This proposed system is could be used with IDS system and it is implemented in python language. Authors in [7] proposes an event detection system which can be embedded into IoT devices. The proposed module able to focuses on the system behavior under DDoS attacks and detects it by information obtained from NTP (Network Time Protocol) used in time synchronization service. The advantage of this solution is that, it is different from the existing ones, it does not require any expensive equipment or tools (e.g. monitoring server) nor periodic maintenance involving technical knowledge.

3. PROPOSED SYSTEM

Based on conducted survey about mitigation techniques of DDoS attack on IoT devices as it has low computing power, providing security in every device is not possible. However at the present juncture this research remains theoretic with only theoretic models being proposed. The practical implementation of Rule based detection and mitigation in IoT systems for the purpose of DDoS attack prevention is an area that remains unexplored. With keeping this in mind, we would plan to deploy a rule based security system for an IoT environment in this research work. Use of rule based detection system would be helpful to mitigate DDoS attack as well as collect information of the attacker so that information could be used for future attacks prevention. Use of rule based security system in Raspberry Pi is might be cost effective solution. After reviewed some research paper about DDoS

detection techniques, using rule based intrusion detection system is still unexplored area for IoT environment. To mitigate Dos attack in IoT environment by using rule based security system which is protect like pillared with a verification system to maintain the efficiency (data received/data transmitted).

interface, Backup Raspberry Pi, Prepare testing bed, Search for various tool for DDoS attack, Install them on 5-6 machine, Set bandwidth according to test case and Connect all the machine with raspberry pi network. Testing: Perform DDoS attack, Check Resource (i.e. Memory, cache) Monitor of Raspberry Pi and Check Network bandwidth of Raspberry Pi and Prepare test report. Following are some screenshots of the implementation of proposed work.

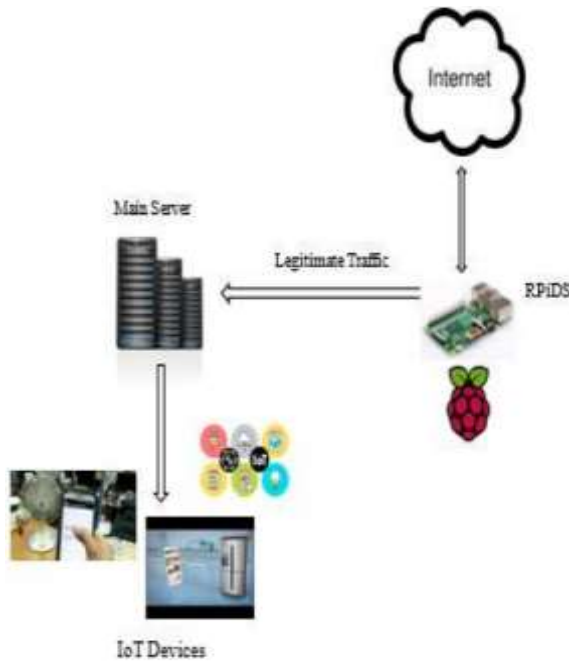
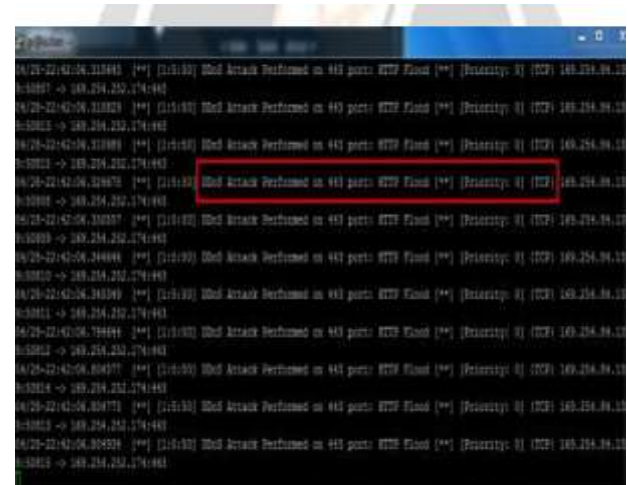


Fig.2: Proposed system architecture

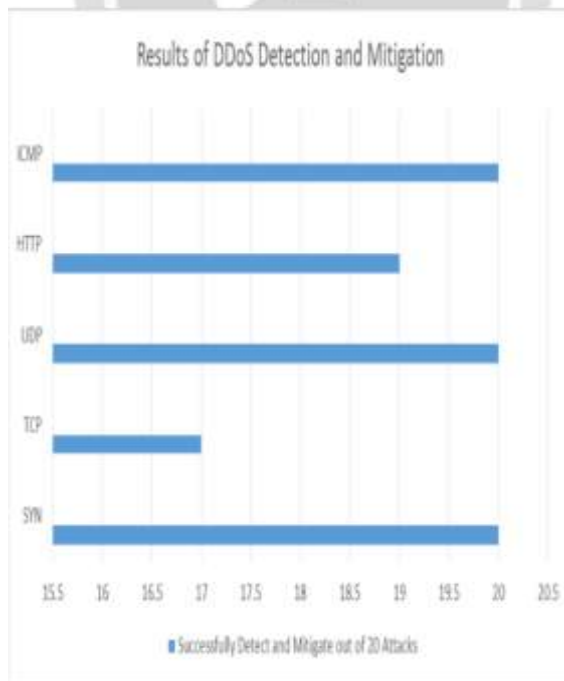
4. IMPLEMENTATION

Flow of the implementation would be described here. Implantation work divided into three parts as following: Installation: Raspbian OS on Raspberry Pi device, Snort IDS in raspberry PI, DDoS Attack tools in windows OS, Kali Linux for more attacking tools, Putty and Windows and Linux OS on 5-6 machine for DDoS attack. Configuration: Raspberry Pi access using SSH via Putty, Snort configuration for Raspberry Pi network interface, Create and Set Rules for Detection of DoS and DDoS attack and Create and Set Rules for Iptables in raspberry pi, Setup attack tool on Kali and Windows OS, Network



5. RESULT AND ANALYSIS

The series of attack on the IoT server is performed here. The different types of DDoS attack are successfully detected as well as mitigated by the proposed system solution. All the testing results are enlisted below. Testing of the proposed system based on manually test. All the attacks are performed 20 times. As shown in below graph, Detection and Mitigation of SYN flood attack is 100%, TCP flood attack 85%, UDP flood attack 100%, HTTP flood attack 95% and ICMP flood attack 100%. After analysis of testing results, overall efficiency of the proposed work is 96%.



6. CONCLUSIONS

With emergence of IoT technology there is a requirement to secure the IoT environment from the DDoS attack. Finding out the DDoS attack and mitigate those attacks are the most challenging task. The proposed system detects and mitigate the DDoS attack using rule based approach and implemented in Raspberry Pi. The proposed system precisely detects and mitigates the DDoS attack in IoT environment. This thesis gives the detailed idea about IoT Environment, DDoS attack, Detection and Mitigation techniques of the DDoS attack. 7.

7. REFERENCES

[1] M. Ahmed and H. Kim, "DDoS Attack Mitigation in Internet of Things Using Software Defined Networking", 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), 2017.

[2] K. Singh and T. De, "DDoS Attack Detection and Mitigation Technique Based on Http Count and Verification Using CAPTCHA", 2015 International Conference on Computational Intelligence and Networks, 2015.

[3] G. Ramadhan, Y. Kurniawan and Chang-Soo Kim, "Design of TCP SYN Flood DDoS attack detection using artificial immune systems", 2016 6th International Conference on System Engineering and Technology (ICSET), 2016.

[4] T. Kawamura, M. Fukushi, Y. Hirano, Y. Fujita and Y. Hamamoto, "An NTP-based detection module for DDoS attacks on IoT", 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2017.

[5] S. Dowling, M. Schukat and H. Melvin, "A ZigBee honeypot to assess IoT cyberattack behaviour", 2017 28th Irish Signals and Systems Conference (ISSC), 2017.

[6] S. Misra, P. Krishna, H. Agarwal, A. Saxena and M. Obaidat, "A Learning Automata Based Solution for Preventing Distributed Denial of Service in Internet of Things", 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing, 2011.

[7] S. Khattab, C. Sangpachatanaruk, D. Mosse, R. Melhem and T. Znati, "Roaming honeypots for mitigating service-level denial-of-service attacks", 24th International Conference on Distributed Computing Systems, 2004. Proceedings, 2004.

Detect DUI An In Car Detection System for Drink Driving and BACs

Yarlagadda Vikram (MCA Scholar), B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

S.K.Alisha, B V Raju College, Vishnupur, Bhimavaram, West Godavari District, Andhra Pradesh, India, 534202.

ABSTRACT

As one of the biggest contributors to road accidents and fatalities, drink driving is worthy of significant research attention. However, most existing systems on detecting or preventing drink driving either require special hardware or require much effort from the user, making these systems inapplicable to continuous drink driving monitoring in a real driving environment. In this paper, we present *DetectDUI*, a contactless, non-invasive, real-time system that yields a relatively highly accurate drink driving monitoring by combining vital signs (heart rate and respiration rate) extracted from in-car WiFi system and driver's psychomotor coordination through steering wheel operations. The framework consists of a series of signal processing algorithms for extracting clean and informative vital signs and psychomotor coordination, and integrate the two data streams using a self-attention convolutional neural network (i.e., C-Attention). In safe laboratory experiments with 15 participants, *DetectDUI* achieves drink driving detection accuracy of 96.6% and BAC predictions with an average mean error of 2 _ 5mg/dl. These promising results provide a highly encouraging case for continued development.

1. INTRODUCTION

IN 2018, The US suffered 10, 511 deaths from drunk driving crashes [1]. The WHO reports that, in high-income countries, as many as 20% of fatally injured drivers have excess alcohol in their blood.1 COVID-related deaths may dwarf these numbers, but it is important not to lose our pre pandemic perspective. Tens of thousands of deaths per year due to drink driving is a staggering loss of life. At an estimated cost of \$44 billion2 for just one year in the US alone [1], the economic impact is not insignificant either. According to the newly-enacted Halt Act (H.R.

2138) [2] and Ride Act (S.B. 1331) [3], drunk-driving prevention technology will be a safety standard for all new cars in the future. There is an urgent demand for in-vehicle drunk-driving detection systems to help prevent drunk-related accidents.

Most traditional methods for detecting drunk drivers need to interrupt the driving process. To administer a breathalyzer, the police must hail the driver to pull over, giving the driver time to implement means of avoiding detection. Blood tests are invasive, and require the driver to stop. Similarly, urine tests and pupil measuring tests require special operations and expert examiners. While it is desirable to detect whether the driver is drunk before driving and prevent potential risks, it is possible that alcohol consumption takes time to take effect and the driver may consume alcohol during driving. Therefore, the most reliable way is to have a continuous monitoring of drunkenness during driving without interfering the driving process.

In this regard, there are many existing studies on leveraging sensing technologies to determine inebriation levels and Blood Alcohol Content (BAC). You *et al.* [4] devised a transdermal sensing wristband with an accompanying smart phone app connected via Bluetooth that calculates and tracks a person's BAC, while Jung *et al.* [5] developed a smart phone attachment that performs a colorimetric analysis on saliva. Nonetheless, these systems require extra devices, some of which are expensive. Researchers have found others ways to detect drunkenness with only smart phones. Kao *et al.* [6], for example, devised an app that measures a person's walking patterns with an inference model that learns to detect abnormal gaits associated with inebriation. Bae *et al.* [7] use the sensor data a person's phone collects to train a machine learning model to distinguish between drinking and non-drinking episodes. Similarly, Markakis *et al.* [8] also focus on changes in a person's unique patterns of coordination to determine inebriation. These solutions discern psychomotor and cognition skills under the influence of alcohol. However, they require users to perform certain activities that interfere with the driving process.

Our solution is a passive continuous drunk-driving detection system called *Detect DUI*. *Detect DUI* measures a person's vital signs through WiFi signals and their psychomotor coordination through steering wheel operations. However, the complicated driving conditions make it challenging to extract clear vital signs and it is essential to find a non-disturbing way of measuring psychomotor coordination. We manage the interference through a multi-step process.

To eliminate reflections from other passengers and from car interiors (e.g., seats, windows), we leverage power delay profile to separate the direct reflection from the chest of the driver from multipath interference. WiFi signals are carried by multiple subcarriers. Different subcarriers have different sensitivities to subtle chest motions. To take full advantage of the diverse information from all subcarriers and avoid interference, we adopt principal component analysis (PCA) to sift noises and preserve the first principal component. Due to bumpy driving conditions, the received signals contain sudden changes with increased amplitude. We remove sudden changes and preserve only signals during relatively stable driving periods, which show a clear cyclic pattern that corresponds to breathing cycles, but the heartbeat pattern is drowned due to its much weaker amplitude. To address this problem, we propose a novel adaptive variational mode decomposition (AVMD) method to separate the mixed signal into multiple modes, and then keep the modes that relate to breathing and heartbeat respectively. Previous works usually measure psychomotor coordination of a person using interactive games or operations with a smart phone or computer, which is not applicable to the driving environment. We find a natural way to gauge the psychomotor coordination of the driver by monitoring their steering wheel operations. In particular, we use IMU to record the acceleration and gyroscope data during operation. In this way, we obtain a continuous monitoring of psychomotor coordination of the driver without interfering with their driving. Integrating the vital signs and the psychomotor signals is done with neural networks and an attention mechanism. Random Forest (RF) is then used to predict concrete BAC values. Trials with 15 volunteers show that *Detect DUI* was able to detect a drunken driver with 96.6% accuracy. Further, it was able to predict a person's BAC within a mean absolute error (MAE) of 0.002% to 0.005%.

Detect DUI can be supported by in-car IMU and WiFi systems. The lightweight learning-based detection model can be deployed locally, with data collected to fine-tune the model locally without privacy leakage.

In summary, the contributions of this research include:

- As far as we are concerned, *Detect DUI* is the first contactless method of detecting drink driving, including measuring the driver's BAC that can be administered while driving.

- We have proposed a series of signal processing algorithms for extracting human vital signs from Wi-Fi signals given chest motions with high levels of accuracy.
- We have proposed to use C-Attention to combine the information of vital signs and psychomotor coordination to reach a well-rounded drunk driving prediction.
- Extensive experiments on 15 individuals show *DetectDUI* is able to distinguish normal driving from drunk driving in real-time with a 96.6%-accurate estimation and the driver's BAC to within an MAE of 0.002% to 0.005%

2. EXISTING SYSTEM

A. Drunkenness Detection

Hardware-Based Detection: First used in the United Kingdom in the 1970s [17], breathalyzers are the world's most commonly used tools for testing inebriated drivers. Over its years of usage, researchers have connected breathalyzers, as well as other types of breath alcohol sensors, to smartphones via Bluetooth to improve BAC tracking, especially for self-monitoring by drivers themselves. Example systems include: BACtrack Mobile Pro [18], Breathmeter [19]. One major disadvantage of breathalyzers is that the results are highly susceptible to the oral environment [20] and certain diseases (e.g., diabetes, liver and kidney diseases [20]), which may lead to false detection. Alternatives to breathalyzers include SCRAM, a transdermal sensor that measures the wearer's BAC through their sweat every 30 minutes [21]. The same kind of system is available in a tight wristband that fits closely to the skin [4]. However, SCRAM-based systems require a close contact between the skin and the sensor. Any space or anything between the skin and the sensor will affect the detection accuracy. Moreover, these systems require users to purchase extra devices or sensors, which may be expensive.

Camera-Based Detection: Camera-based drunk driving systems have also been developed [22], [23]. In [22], facial landmarks and motions are recognized in images to detect whether the driver is drunk driving or not. In [23], an audiovisual database is utilized to realize bimodal intoxication detection. However, camera-based approaches are sensitive to lighting conditions and there is potential risk of privacy violation [24].

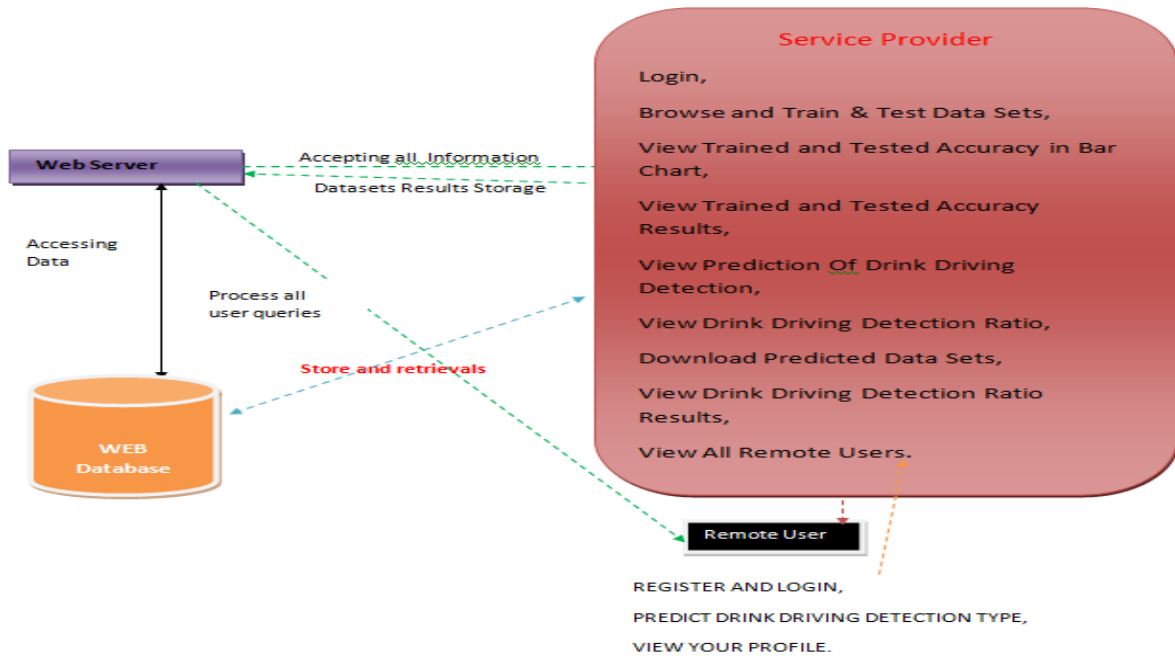
Behavior-Based Detection: The side effects of alcohol consumption include arrhythmia [14], slowed respiratory rates [15], impaired psychomotor performance [8], and unsteady gait [6]. This

abnormality in vital signs and behaviors can be leveraged to detect whether the user is under the influence of alcohol. Bae *et al.* [7] developed a smart phone based system to track the drinking episodes of users based on built-in sensors (e.g., accelerometer) and the smartphone status (e.g., battery and network usage). Leveraging alcohol's influence on motor coordination and cognition, Markakis *et al.* [8] designed five human-computer interactions to detect BACs (such as swiping or touching the screen in particular ways), akin to the finger-to-nose DUI tests. However, these works require users to interact with their phones (swipe the phone or engage in games), which interrupts the driving task and cannot offer a continuous drunk driving detection.

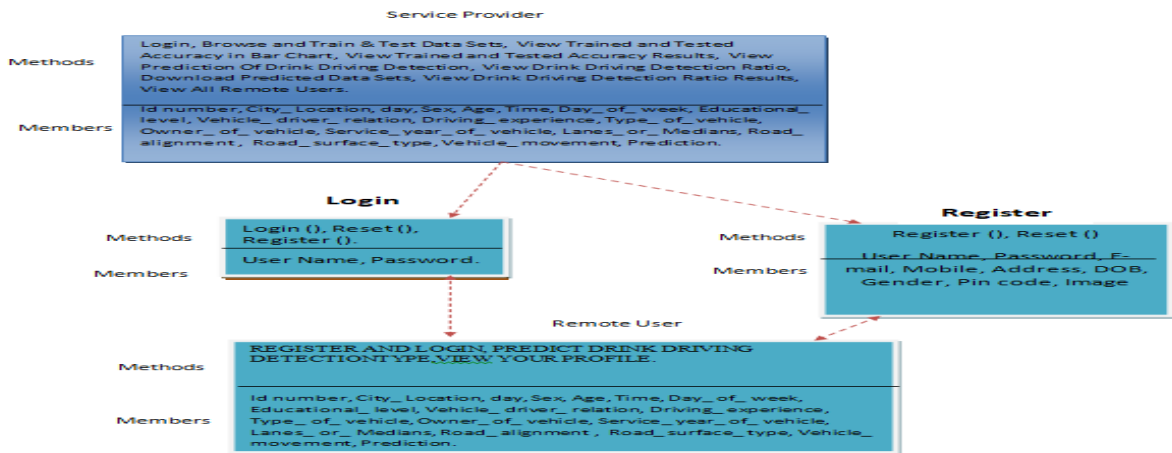
Proposed System

- As far as we are concerned, *DetectDUI* is the first contactless method of detecting drink driving, including measuring the driver's BAC that can be administered while driving.
- We have proposed a series of signal processing algorithms for extracting human vital signs from WiFi signals given chest motions with high levels of accuracy.
- We have proposed to use C-Attention to combine the information of vital signs and psychomotor coordination to reach a well-round drink driving prediction.
- Extensive experiments on 15 individuals show *DetectDUI* is able to distinguish normal driving from drink driving in real-time with a 96.6%-accurate estimation and the driver's BAC to within an MAE of 0.002% to 0.005%.

Architecture Diagram



Class Diagram :



3. SYSTEM STUDY

FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed

system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

ECONOMICAL FEASIBILITY

TECHNICAL FEASIBILITY

SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and

to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

4. CONCLUSION

In this paper, we presented *Detect DUI*, a non-intrusive, contactless, and continuous system of measuring and monitoring the side effects of alcohol on drivers. To develop *Detect DUI* to this stage, we have overcome two main challenges. The first is to eliminate interference in the WiFi signals caused by the motions of a moving vehicle. This problem was solved with a series of signal processing algorithms. The second is determining which specific features of alcohol's side effects best reflect driving under the influence of alcohol. We have addressed this challenge with a C-Attention network. The results of extensive experiments confirm that *Detect DUI* provides highly accurate drink driving detection and BAC prediction.

Apart from drinking alcohols, other factors may also affect vital signs and psychomotor coordination, e.g., catching a cold or other respiratory diseases. Respiratory diseases will change breathing patterns, which are expected to be different from the breathing patterns of drinking. However, it is difficult to collect training samples to help differentiate the breathing patterns under the two conditions. In the future, we intend to refine our drink driving detection model by considering other impact factors.

5. REFERENCES

- [1] *Alcohol Impaired Driving: 2018 Data (Traffic Safety Facts. Report No. DOT HS 812 864)*. Accessed: Sep. 2019. [Online]. Available: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812864>
- [2] *H.R. 2138: Honoring Abbas Family Legacy to Terminate Drunk Driving Act of 2021*. Accessed: Mar. 2021. [Online]. Available: <https://www.govtrack.us/congress/bills/117/hr2138>
- [3] *S.1331—RIDE Act of 2021*. Accessed: 2021. [Online]. Available: <https://www.congress.gov/bill/117th-congress/senate-bill/1331/text>
- [4] C.-W. You *et al.*, “Enabling personal alcohol tracking using transdermal sensing wristbands: Benefits and challenges,” in *Proc. 21st Int. Conf. Hum.-Comput. Interact. Mobile Devices Services*, Oct. 2019, pp. 1–6.
- [5] Y. Jung, J. Kim, O. Awofeso, H. Kim, F. Regnier, and E. Bae, “Smartphone-based colorimetric analysis for detection of saliva alcohol concentration,” *Appl. Opt.*, vol. 54, no. 31, pp. 9183–9189, 2015.

- [6] H.-L. Kao, B.-J. Ho, A. C. Lin, and H.-H. Chu, "Phone-based gait analysis to detect alcohol usage," in *Proc. ACM Conf. Ubiquitous Comput. (UbiComp)*, 2012, pp. 661–662.
- [7] S. Bae *et al.*, "Detecting drinking episodes in young adults using smartphone-based sensors," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 2, pp. 1–36, Jun. 2017.
- [8] A. Mariakakis, S. Parsi, S. N. Patel, and J. O. Wobbrock, "Drunk user interfaces: Determining blood alcohol level through everyday smartphone tasks," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2018, pp. 1–13.
- [9] *Alcohol*. Accessed: 2021. [Online]. Available: <https://en.wikipedia.org/wiki/Alcohol>
- [10] *Alcohol's Effects on the Body*. Accessed: 2021. [Online]. Available: <https://www.niaaa.nih.gov/alcohols-effects-health/alcohols-effects-body>
- [11] *Hangovers*. Accessed: 2021. [Online]. Available: <https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/hangovers>

Forecasting Big Mart Sales: A Machine Learning Approach

Y. Jyothisna Vinitha¹ and S. K. Alisha²

Student, MCA, B. V Raju College, Bhimavaram, India¹

Assistant Professor, MCA, B. V Raju College, Bhimavaram, India²

Abstract: In modern supermarkets like Big Marts, meticulous tracking of sales data for each item is pivotal for projecting latent consumer demand and streamlining operational strategies. This entails anticipating product demand for inventory management, logistics, and optimal resource allocation. By strategically dissecting the vast reservoir of sales data, insights are consistently root out, revealing anomalies and overarching trends. Employing a deliberate entanglement of data warehousing, the data store continually exposes nuanced patterns. Establishments like Big Mart harness this trove of information to predict forthcoming transaction volumes through diverse machine learning methodologies, akin to the practices observed in prominent retail giants such as big bazaars. The prevailing machine learning algorithms have reached a pinnacle of sophistication, furnishing tools to predict or comprehend transactions of various natures. This capability proves invaluable in shaping and honing marketing strategies, extremely through more precise and informed forecasting. This study introduces a predictive model, leveraging advanced methods involves linear regression and Ridge regression, for dissecting the transactional dynamics of an enterprise like Big Mart. Notably, this model's performance surpassed standalone methodologies.

Keywords: Polynomial Regression, Linear Regression, Mean Absolute Error, XgBoost Regression, Ridge Regression

I. INTRODUCTION

In the dynamic outlook of retail, staying ahead of market trends and consumer preferences is pivotal for businesses' sustained success. As the competition intensifies, retailers are progressively twisting to advanced analytical techniques to gain insights into their sales patterns and optimize their strategies. One such approach is predictive analysis, which leverages machine learning algorithms to forecast future sales based on historical data. This study on applying predictive analysis to Big Mart, a prominent retail chain, to unlock valuable insights for informed decision-making.

Table

Variable	Description	Relation to Hypothesis
Outlet_Identifier	Unique store ID	ID variable
Outlet_Establishment_Year	The year in which store was established	Not considered in hypothesis
Outlet_Size	The size of the store in terms of ground area covered	Linked to 'store capacity' hypothesis
Outlet_Location_Type	The type of the city in which the store is located	Linked to 'city type' hypothesis
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket	Linked to 'store capacity' hypothesis again
Item_Outlet_Sales	Sales of the product in the particular store. This is outcome variable to be predicted	Outcome Variable
Item_MRP	Maximum retailer price (list price) of the product	Not considered in hypothesis

Item_Type	The category to which the product belongs	More interfaces about 'utility' can be derived from this
Item_Visibility	The % of the total display area of all products in a store allocated to the particular product	Linked to 'display area' hypothesis
Item_Fat_Content	Whether the product is low fat or not	Linked to 'utility' Hypothesis. Low fat items are generally used more than others
Item_Weight	Weight of product	Not considered in hypothesis
Item_Identifier	Unique product ID	ID variable

II. FIGURES AND TABLES

A dataset comprising a group of data points obtained from the internet serves as a unified entity for computer analysis and predictive purposes. This dataset is taken from the Kaggle.com platform. The testing dataset involved in this investigation encompasses 8542 rows distributed across 12 distinct classes. Rigorous training has applied to optimize predictive outcomes, striving for the utmost accuracy in predictions.

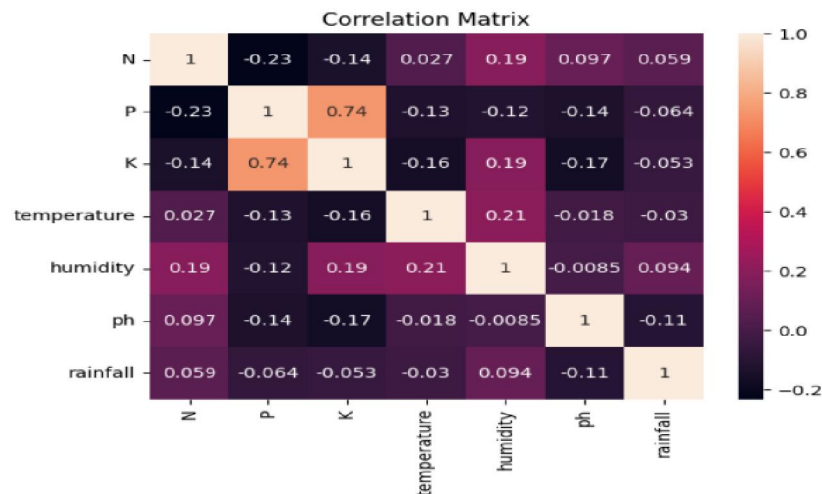


Fig.1 Matrix Graph

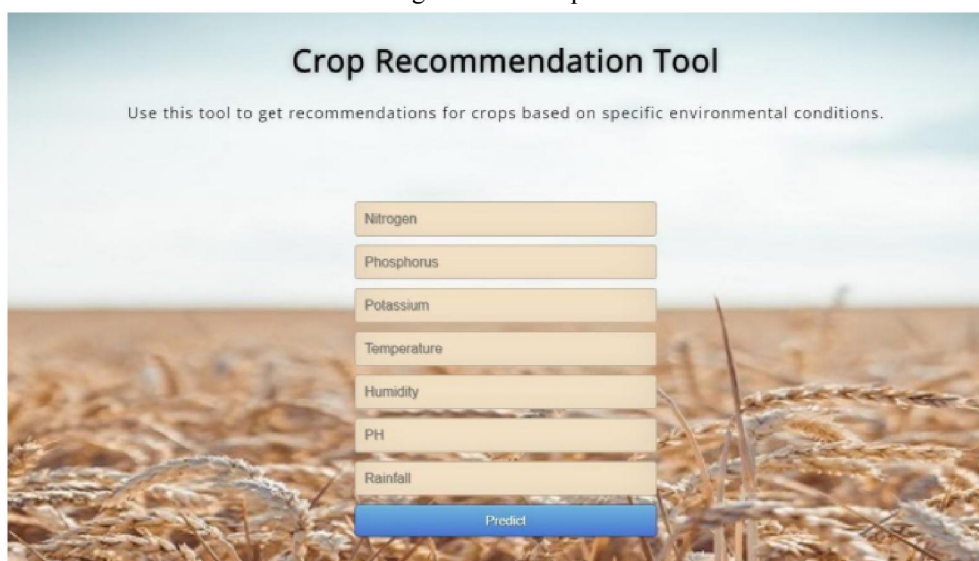


Fig.2 Home Page

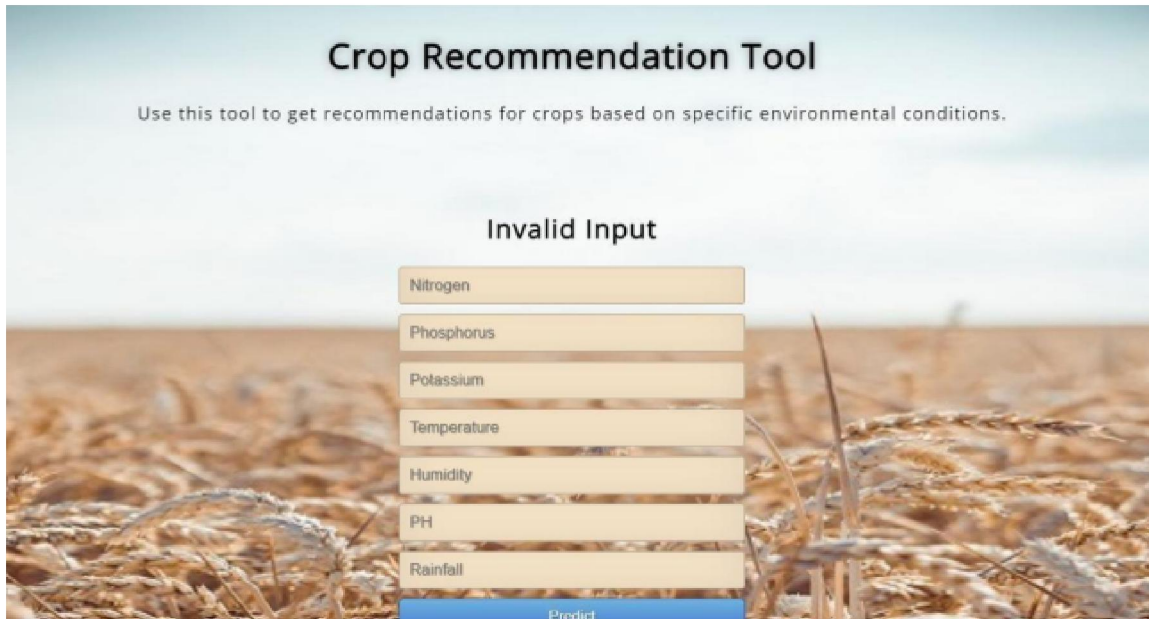


FIG.3 Alert for Invalid

III. LINKS AND BOOKMARKS

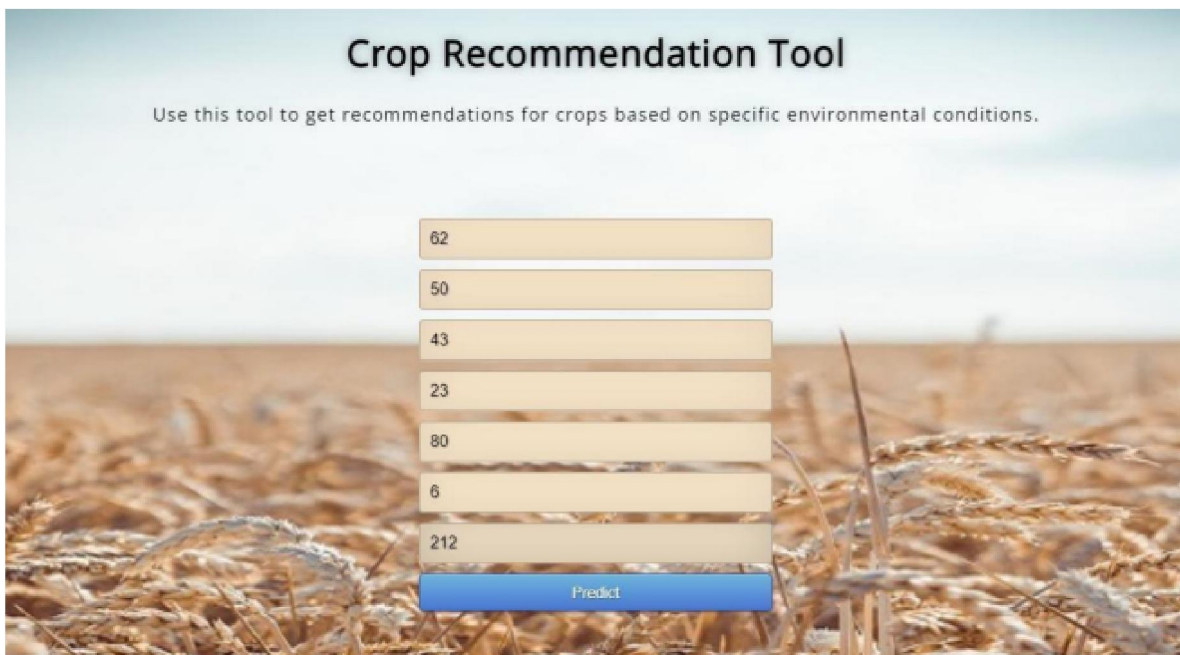


FIG.4 OUTPUT

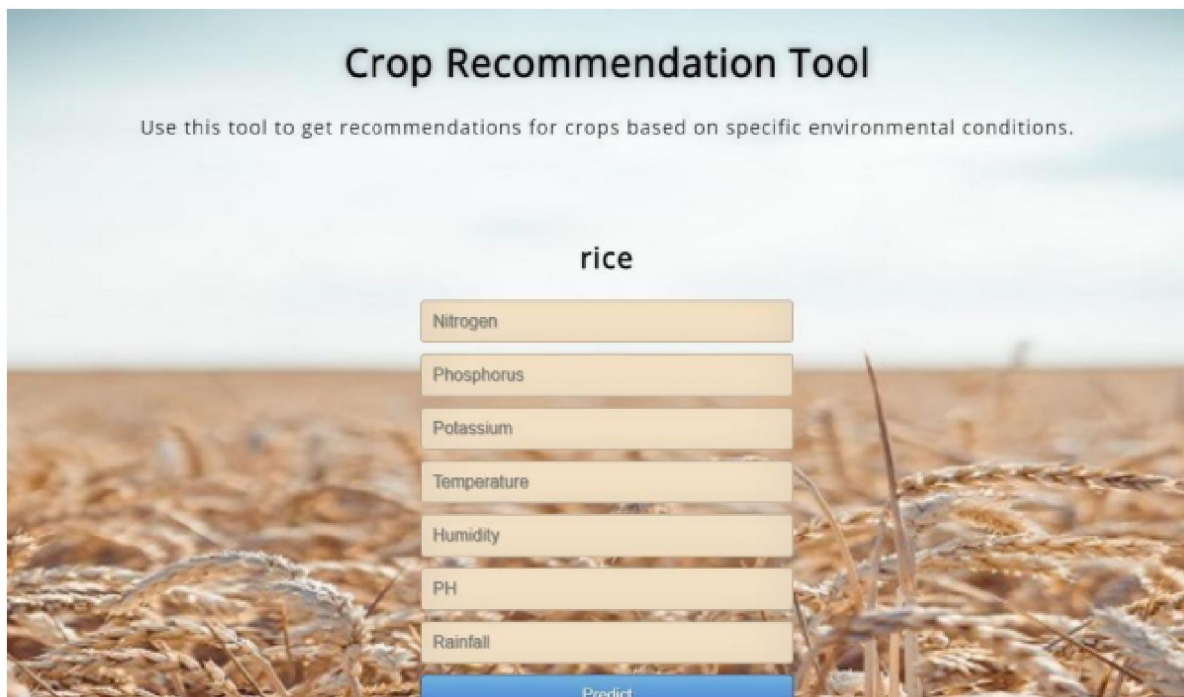


Fig.5 Giving Output

IV. CONCLUSION

In conclusion, the appeal of predictive analysis through machine learning techniques has proven to be a valuable tool for understanding and forecasting sales trends at Big Mart. Through this study, we have demonstrated that utilizing historical sales data, together relevant features such as store demographics, item attributes, and promotional strategies, can lead to accurate predictions of future sales. Combination of machine learning models, such as regression, decision trees, and ensemble methods, has allowed us to capture complex relationships within the data, enabling better decision-making and resource allocation. By leveraging predictive insights, Big Mart can enhance its inventory management, optimize pricing strategies, and tailor marketing efforts to specific segments. This data-driven approach excluding improve customer satisfaction by ensuring product availability but also contributes to cost savings by minimizing overstock and reducing instances of understock. Moreover, the ability to expect demand variations enables Big Mart to make proactive adjustments, ensuring efficient resource utilization and overall operational improvement. However, it's important to note that predictive analysis is not a one-time effort; it requires continuous monitoring, model refinement, and adaptation to evolving market dynamics. Additionally, the success of predictive analysis heavily relies on data quality, feature engineering, and model selection. As the retail landscape and consumer behaviour carry on with to evolve, the model must be regular updated to maintain their accuracy and relevance. In conclusion, the marriage of predictive analysis and machine learning offers Big Mart a powerful means to stay competitive in an ever-changing market. By harnessing the potential of data, the company can make informed decisions, drive growth, and foster innovation, ultimately leading to a more agile and customer-centric retail operation.

V. ACKNOWLEDGMENT

I would like to thank Assistant Professor S. K. Alisha for his valuable suggestion, expert advice and moral support in the process of preparing this paper.

REFERENCES

- [1]. PYTHON - REFERENCE
- [2]. TUTORIAL - CODEGNAN IT SOLUTIONS

- [3]. <https://academy.codegnan.com/learn/home/Python-Fullstack/Python/section/177592/lesson/967216>
- [4]. HTML AND CSS - REFERENCE
- [5]. TUTORIAL - CODEGNAN IT SOLUTIONS
- [6]. <https://academy.codegnan.com/learn/home/Python-Fullstack/Introduction-to-HTML/section/317116/lesson/1937841>
- [7]. FLASK,MACHINE LEARNING - REFERENCE
- [8]. TUTORIAL - CODEGNAN IT SOLUTIONS
- [9]. <https://academy.codegnan.com/learn/home/Python->