Volume 11. No. 3
October - December 2023

ISSN-2347-2227
E-ISSN-2347-6141
Subscribers copy
Not for sale

i-manager's

Journal on Computer Science

Disseminating new ideas in Information and Computation

i-manager® Publications

ISSN: 2347-2227
Volume 11. No. 3
October - December 2023

i-manager's

Journal on Computer Science

About the Journal

i-manager's Journal on Computer Science deals with all aspects of computer science and contributes theoretical results and offers a compilation of high quality articles to encompass a wide spectrum of advancements in the actively developed domain. i-manager's Journal on Computer Science covers a great deal of what has been done in the field recently and intends to bring together the most recent advances and applications in all branches of the academic computer science community with new knowledge and technology for the benefit of students, professionals and industrial practitioners.

i-manager's Journal on Computer Science is presently in its 11th Year. The first issue was launched in 2013.

i-manager's Journal on Computer Science is published by i-manager Publications, one of India's leading Academic Journal Publisher, publishing 38 Academic Journals in diverse fields of Engineering, Education, Management and Science.

Why Publish with us

i-manager Publications currently publishes academic Journals in Education, Engineering, Scientific and Management streams. All of i-manager's Journals are supported by highly qualified Editorial Board members who help in presenting high quality content issue after issue. We follow stringent Double Blind Peer Review process to maintain the high quality of our Journals. Our Journals target both Indian as well as International researchers and serve as a medium for knowledge transfer between the developed and developing countries. The Journals have a good mix of International and Indian academic contributions, with the peer-review committee set up with International Educators.

Submission Procedure

Researchers and practitioners are invited to submit an abstract of maximum 200 words on or before the stipulated deadline, along with a one page proposal, including Title of the paper, author name, job title, organization/institution and biographical note.

Authors of accepted proposals will be notified about the status of their proposals before the stipulated deadline. All submitted articles in full text are expected to be submitted before the stipulated deadline, along with an acknowledgement stating that it is an original contribution.

Review Procedure

All submissions will undergo an abstract review and a double blind review on the full papers. The abstracts would be reviewed initially and the acceptance and rejection of the abstracts would be notified to the corresponding authors. Once the authors submit the full papers in accordance to the suggestions in the abstract review report, the papers would be forwarded for final review. The final selection of the papers would be based on the report of the review panel members.

Format for Citing Papers

Author surname, initials (s.) (2023). Title of paper. i-manager's Journal on Computer Science, 11(3), xx-xx.

Copyright

Contact e-mails

editor@imanagerjournals.com
submissions@imanagerpublications.com

ISSN: 2347-2227
Volume 11. No. 3
October - December 2023

i-manager's
Journal on Computer Science

ISSN: 2347-2227
Volume 11. No. 3
October - December 2023

i-manager's
Journal on Computer Science

OUR OFFICES

Registered Office

3/343,Hill view,
Town Railway Nager,
Nagercoil, Kanyakumari District - 629001
Ph : 91-4652- 277675
E-mail : info@imanagerpublications.com

Editorial Office

13-B, Popular Building,
Mead Street, College Road,
Nagercoil, Kanyakumari District - 629001
Ph : (91-4652) 231675, 232675, 276675
E-mail : editor@imanagerjournals.com

Join with us

https://www.facebook.com/imanJCOM

https://twitter.com/imanagerpub

Abstracting / Indexing

CONTENTS

	RESEARCH PAPERS
1	EVALUATING HYBRID MODEL PERFORMANCE IN PREDICTING THE IMPACT OF LOCKDOWN ON CONTROLLING COVID-19 CASES PROGRESSION IN MIDDLE INCOME SETTINGS By Yahaya Mohammed Sani, Benjamin Davou Pam
12	A METHOD FOR THE IDENTIFICATION OF DENIAL OF SERVICE (DoS) ATTACK IN NETWORK TRAFFIC USING MACHINE LEARNING TECHNIQUES By Gottapu Sankara Rao, P. Krishna Subbarao
25	AGRICULTURE MANAGEMENT SYSTEM By Takondwa Kaiya, Chipatso Medi, Fanny Chatola
36	UNVEILING SENTIMENT ANALYSIS: A COMPARATIVE STUDY OF LSTM AND LOGISTIC REGRESSION MODELS WITH XAI INSIGHTS By Chandu D. Vaidya, Mayuri Botre, Yash Rokde, Sagar Kumbhalkar, Soham Linge, Soham Pitale, Shreyash Bawne
47	RESUME SCREENER SYSTEM By Muhammad Savad N., T. Preethi

EDITORIAL

i-manager's Journal on Computer Science (JCOM), (October-December 2023: Volume 11 Issue 3), features five peer-reviewed research papers covering various topics within the realm of computer science. The Journal is dedicated to publishing research articles on information and computation, as well as practical techniques for their implementation and application in computer systems.

Yahaya and Benjamin evaluated the performance of a hybrid model in predicting the impact of lockdowns on controlling the progression of COVID-19 cases in middle-income settings. This study developed various models, including the Hybrid ANN-CNN, to predict COVID-19 cases based on human mobility data. The hybrid model demonstrated superior performance compared to others, indicating its effectiveness in assessing the impact of lockdown measures. This research aims to assist policymakers in making informed decisions during pandemics by highlighting the factors influencing case numbers.

Gottapu and Krishna propose a method for identifying Denial-of-Service (DoS) attacks in network traffic using machine learning techniques. This article presents classification models employing algorithms such as KNN, Logistic Regression, and Random Forest. Data from Wireshark were used for evaluation, employing metrics such as Accuracy, Precision, Recall, and F-1 score. The study demonstrates how machine learning enhances DoS attack prediction, thereby aiding network security analysis.

Takondwa et al. introduce the Agriculture Management System (AMS), a web platform for farmers. AMS offers vital agricultural data such as weather forecasts, soil or crop statistics, and market prices, aiding informed decision-making. It enhances productivity and profitability by enabling efficient planning, pricing, and marketing decisions. This feature allows farmers to share information, seek assistance, and ask questions about the agricultural management system, including crops, soil, market trends, and weather management. An AI chatbot promotes community-driven agricultural management, fostering farmer success.

Chandu et al. explore Sentiment Analysis and Explainable AI (XAI), focusing on techniques like Lime for transparency in Logistic Regression and LSTM models. The paper highlights how XAI uncovers biases in ML predictions and aids in understanding DL models. Identifying a gap in training and interpretation, they aim to bridge it by applying XAI to both ML and DL models on the same dataset. Their research reveals that it outperforms ML models, a distinction only discernible with XAI techniques, particularly Lime.

Muhammad and Preethi introduce a cutting-edge "Resume Screener System" designed to automate resume analysis for recruitment. With a dataset of a thousand resumes from Kaggle, the system undergoes rigorous model training and evaluation, employing methods like SVC and Neighbours Classifier. The goal is to optimize resume screening by matching job requirements with candidates' skills objectively. By automating screening, it saves time, ensures consistency, and advances machine learning in HR, transforming traditional hiring practices.

We extend our profound thanks to the authors for their contribution towards this issue and we are grateful to the reviewers for spending their quality time in reviewing these papers. Our special thanks to the Editor-in-Chief Dr. Kamal Kumar Mehta for his constant support and efforts in further enhancing the quality of the Journal.

Hope this issue imparts an enlightening reading experience! Enjoy Reading!

Warm regards,

Cindhiya Jislin R.
Senior Editor,
i-manager Publications

EVALUATING HYBRID MODEL PERFORMANCE IN PREDICTING THE IMPACT OF LOCKDOWN ON CONTROLLING COVID-19 CASES PROGRESSION IN MIDDLE INCOME SETTINGS

By

YAHAYA MOHAMMED SANI * BENJAMIN DAVOU PAM **

*-** Makerere University, Kampala, Uganda.

Date Received: 19/09/2023 Date Revised: 21/10/2023 Date Accepted: 25/11/2023

ABSTRACT

The rapidly spreading COVID-19 pandemic in 2020-2021 affected more than 190 countries, including Nigeria. Following this scenario, countries around the world were monitoring confirmed cases, recovering and dying. In order to reduce the impact of the pandemic, most countries implemented several measures to control the spread of the virus. These include closing schools and borders, shutting down public transport and workplaces, and restricting public gatherings until herd immunity is achieved through vaccination. The breakdown of the health system and the unpredictable nature of human behaviour makes it difficult to predict and evaluate the impact of lockdown on the COVID-19 pandemic in the long term. In light of the above, this study developed Hybrid ANN-CNN and four other models, namely LASSO, ANN, CNN and LSTM, to predict and evaluate the effect of human mobility on COVID-19 confirmed cases. To evaluate the effectiveness of non-pharmaceutical interventions (NPIs) and to predict the spread of COVID-19 confirmed cases, publicly available data on human mobility collected by Google and air passenger data were used. In this study, our motivation was to evaluate effect of lockdown on COVID-19 and models that predict the impact of human mobility on COVID-19 confirmed cases based on MSLE, Huber loss, and Log Cosh performance measures. At the end of the experiment, the developed hybrid ANN-CNN outperformed the other four models with MSLE of (0.0022), Huber Loss (0.0014) and Log Cosh (0.0013) respectively. This study serves as an alarm system to provide policy makers with the human mobility factors that can trigger large numbers of cases during a pandemic. This will allow for urgent public action.

Keywords: Mobility, COVID-19, Prediction, Confirmed, Cases, Metrics and Lockdown.

INTRODUCTION

The world has experienced the dramatic consequences of the COVID-19 pandemic in 2020-2021, which unfortunately go beyond the impact of the healthcare sector (García-Cremades et al., 2021). More so, COVID-19 has caused one of the most serious crises in our recent history (García-Cremades et al., 2021). This is as, more than 190 countries, including Nigeria, were infected by the rapidly spreading COVID-19 disease. In the wake of this scenario, countries around the world were monitoring confirmed cases, recoveries and deaths (Khalifa et al., 2023). Even more so, policy makers everywhere have been working to determine the set of restrictions that will effectively contain the spread of COVID-19 without unduly stifling economic activity and save lives (Khalifa et al., 2023; Ilin et al., 2021).

Consequently, while eliminating random noise, the perfect model should adopt an effective evaluation technique, with lowest loss values after training and testing the model. The model should be able to fit well as much as the complexity of the data demands due to the ease

of computing power, but there is still some challenge regarding the choice of evaluation techniques that gives low loss functions, in the context of predicting human mobility factors against COVID- 19 confirmed cases, in Nigeria (Jadon et al., 2022). This is as, the aim of every machine learning model's primary objective is to decrease the loss connected with the model and enhance the metrics that are chosen. The loss function is an essential part of any machine learning or deep learning model used for time series forecasting; the model's learning parameters are determined by the function's minimization, and its performance is evaluated against the loss function (Jadon et al., 2022). In order to reduce the impact of the pandemic, most countries have implemented several measures to control the spread of the virus, including the closure of schools and borders, public transport and workplaces, and the limitation of public gatherings (Said et al., 2020; García-Cremades et al., 2021). Until herd immunity is reached through vaccination, these are the only mechanisms available to control the outbreak, and were designed to reduce movement (García-Cremades et al., 2021; Ilin et al., 2021). Since these measures involve the closure of economic activities such as tourism, cultural activities, among others, governments apply them for a limited period of time. These measures and the allocation of social and economic subsidies were based on the development of the epidemic. The break- down of the health system and the unpredictable nature of human behaviour makes it difficult to predict this evolution in the long run (García-Cremades et al., 2021).

In light of the above, publicly available data on human mobility collected by Google and other providers can be used to evaluate the effectiveness of these non-pharmaceutical interventions (NPIs) and to predict the spread of COVID- 19 (García-Cremades et al., 2021; Khalifa et al., 2023). Most countries have developed surveillance systems based on pandemic evolution indicators to trigger social distancing measures when significant increases in infections are detected. Data analysis can help predict the long-term evolution of the pandemic and thus assist policy makers in their decision-making (García-Cremades et al., 2021). Several computational methods have been used to predict epidemic time series. Deep learning was one of the most promising methods for time series forecasting (Khalifa et al., 2023). In this study, we are motivated to address the problem associated with adopting evaluation techniques with minimum loss functions. More so, this study evaluated models that predict the effect of human mobility on COVID-19 confirmed cases based on MSLE, Huber Loss and Log Cosh performance comparison evaluation metrics with achieved decreased loss functions. The models used in this study include but are not limited to: the developed Hybrid ANN-CNN, LASSO, ANN, CNN and LSTM. At the end of the experiment, the developed Hybrid ANN-CNN outperformed the other models based on MSLE, Huber Loss and Log Cosh.

The contributions in this work are:

- We developed a Hybrid ANN-CNN that predict the impact of human mobility in long term and also takes in to account daily COVID-19 confirmed cases under lockdown with small amount of data and with extraction of relevant features.
- The trend of confirmed cases versus human mobility variables under lock- down period was analyzed.
- Architecture for the model was designed.
- The developed model was evaluated based on MSLE, Huber Loss and Log Cosh models comparison performance evaluation metrics.

The reminder of the sections are: literature review, methodology, results, discussions and conclusion.

1. Related Work

Said et al. 2020 proposed a data-driven approach to predict daily cases of COVID-19, which also allows the testing of different scenarios in terms of lockdown policy. The proposed model took into account both the lockdown information and the daily cases of countries with a similar lockdown policy and showed the same response to the outbreak of the virus. The authors focused their experiments on Qatar as a use case and showed that the proposed model achieved better prediction by including lockdown information and training the model

on data from countries with similar policies. Their analysis also showed that a sudden increase in the number of cases in Qatar would be caused by the complete lifting of restrictions on schools and borders, but regrettably the model could not predict under small amount of data.

Considering human mobility in Hubei province, Wang et al. 2021 proposed a simple but effective short-term predictive multiple linear regression model for COVID-19 cases. Several large cities in Guangdong province are used to validate the performance of the proposed model. The values of R2 of daily prediction reach 0.988 and 0.985 for cities such as Shenzhen and Guangzhou with frequent daily population flows. The proposed model has served as a reference for decision support in the prevention and control of the pandemic in Shenzhen, unfortunately, this model can only predict in the short term.

In this study, Khalifa et al. (2023) proposed a model to predict the spread of COVID-19 in Egypt based on deep learning sequential regression using data on population mobility reports. In the presented model, a new combined data set from two different sources was used. One is population mobility data from Google, and one is daily infection figures from " world in data". With the lowest prediction error, the proposed model was able to predict new cases of COVID-19 infection within 3-7 days. For a 3-day prediction, the proposed model achieved an accuracy of 96.69%. This study is remarkable because it is one of the first to estimate the daily inflow of new COVID-19 infections using data on population mobility rather than the daily infection rate. In this regard, 3 days prediction is insufficient to provide trend of the pandemic in order to give insight on how to curtail it.

In this approach Ilin et al. (2021), a simple statistical model was used to estimate the impact of non-pharmaceutical interventions (NPIs) on mobility, and basic machine learning methods were used to generate a 10-day forecast of COVID-19 cases. The approach has the advantage of making minimal assumptions about disease dynamics, using only publicly available data. The method has been assessed using local and regional data from China, France, Italy, Korea and the US, and national data from 80 countries worldwide. For the countries in our sample, the MPE is 6.35% (5 days) and 15.24% (10 days) including mobility, and 11.46% and 31.12% without mobility. The researchers found that NPIs are associated with significant reductions in human mobility and that changes in mobility can be used to predict COVID-19 infections.

In order to create a decision support system for policy makers, García-Cremades et al. (2021) evaluated different machine learning models for early prediction of the evolution of the COVID-19 pandemic. The authors considered a broad range of models, including artificial neural networks like LSTMs and GRUs, and statistical models like AR or ARIMA. Several consensus strategies were proposed to ensemble all the models into one system. This was done to obtain better results in this uncertain environment. Finally, in order to better predict trend changes in the 14-day CI, a multivariate model incorporating mobility data from Google was proposed. A real case study in Spain has been evaluated, and accurate results have been obtained for the prediction of the 14-day CI in scenarios with and without trend changes, reaching an R2 of 0.93, an RMSE of 4.16 and an MAE of 1.08. This model was also based on short term prediction.

From the available literature reviewed, most previous studies do not dwell on resolving the challenge regarding the choice of evaluation techniques with a lower loss functions, in the context of predicting human mobility factors against COVID-19 cases. This provides solid foundation for this research to close the identified gaps.

2. Methodology

2.1 Research Methodology

In this study, COVID-19 confirmed cases and human mobility data were collected. The human mobility change trend includes six categories: food and pharmacy, parks, retail and recreation, transit stations, workplaces and homes, and in addition, air passenger arrival and departure data. The data is then normalised by rescaling it based on the proposed 2-step feature scaling method, which involves the use of Max Absolute Values and Min Max feature scaling techniques sequentially. The

data is then divided into training and test in the ratio of 80% for training and 20% for testing. The models are now trained one at a time. After training, each model that is fit is now moved to the testing phase, or retrained if it was not fit before the testing phase. After this phase, the model is validated and evaluated based on the performance comparison evaluation metrics of Huber Loss, Mean Square Logarithmic Error and Log Cosh. The results are presented in tables and analysed, discussed and conclusions are drawn based on the results obtained to select the best model. Figure 1 presents the methodology.

2.2 Data Collection and Description

In order to combat COVID-19, Google released a data of human mobility trend to places from around the world,

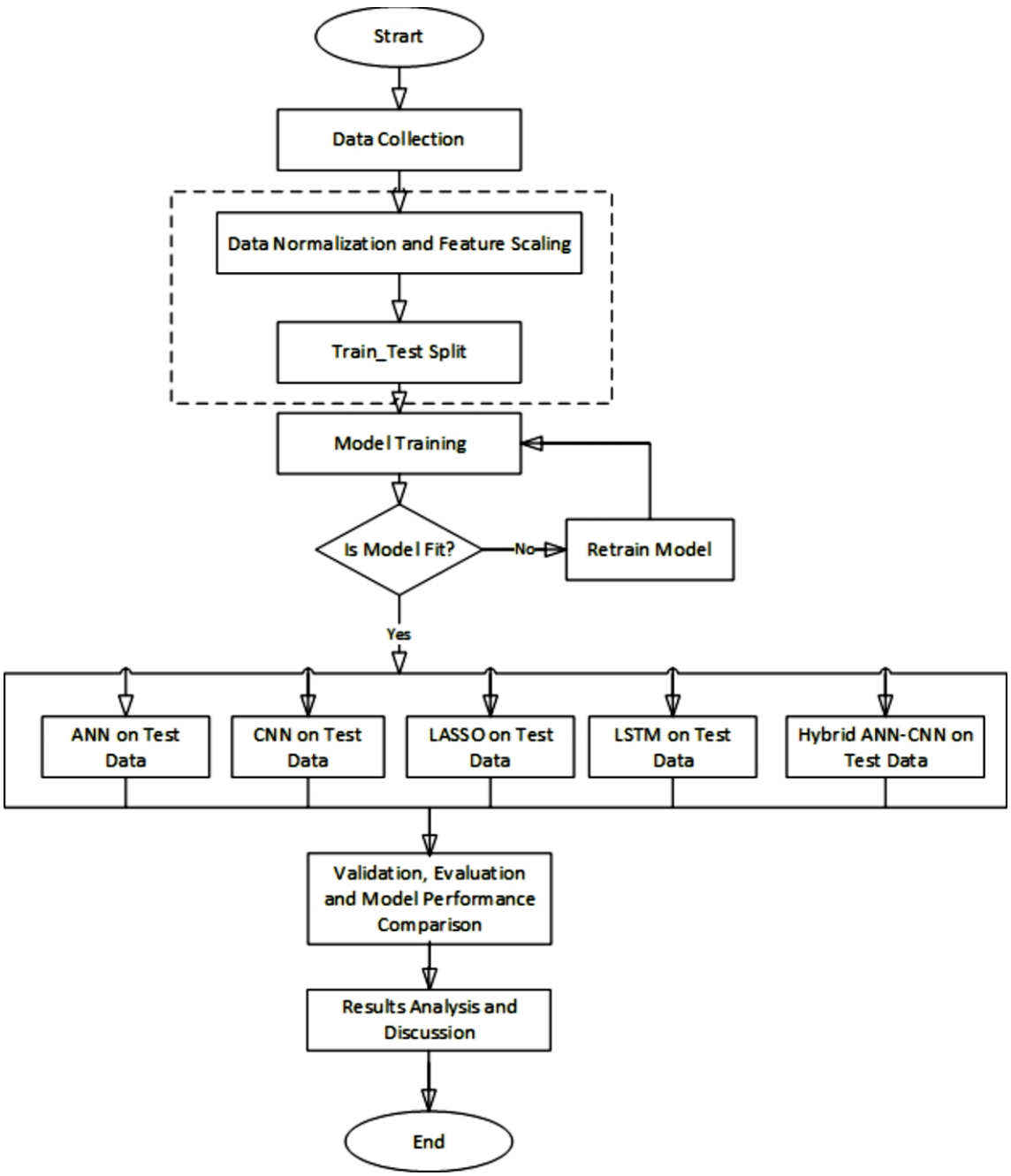


Figure 1. Methodology Adopted in this Study

represented by summary statistics. The freely available Google COVID-19 Mobility Reports (GMR) were extracted, approximating mobility changes due to different social distancing measures (Athanasios et al., 2021; Lawal & Nwegbu, 2022).

The GCMR is the percentage change in the mobility of Google Maps users compared to a prepandemic baseline. The mobility change trend includes six categories: food and pharmacy, parks, retail and recreation, transit stations, work- places and homes, and air passenger arrival and departure data from Erim et al. (2021) and Air Transport full year 2021.xlsx (live.com). This is complemented by COVID-19 confirmed cases collected from the Nigeria Centre for Diseases Control microsite. The period considered for this study is the interdiction period. This is between 1 May, 2020 and 30 April, 2021.

2.3 Normalization and Feature Scaling

Normalising is used to smooth nonlinear data. Each sample is divided by the highest value of the sample data to normalise the data set. The data values have been normalised to aid the training process using a proposed 2-step feature scaling technique. This involves the Max Absolute Value and Min Max scaler. Simply put, the MaxAbs scaling takes the absolute maximum value from each column, dividing each value within the column by the maximum value on the [-1, 1] range. While the Min-Max scaling scales the values on the range (0, 1), where 0 represents the minimum values and 1 represents the maximum values (Hota et al., 2021; Pandey & Jain (2017).

2.4 Training Test Split

The dataset used in this study was divided into a training subset and a test subset, at a ratio of 80% and 20% for the training and test validation, respectively.

2.5 Mobility Measuring for the Hybrid ANN-CNN Model

These data are provided by local mobility reports, which provide valuable information on changes in people's mobility patterns as a result of government responses to the COVID-19 pandemic. The information contained in these reports is of particular interest to us because it shows the movement of the population over time. This information is broken down by geographical area and by different categories of place, such as workplaces, shops, supermarkets, leisure areas, pharmacies, parks, transport stations and residential areas (García-Cremades et al., 2021). The main variables provided by GMD are as follows:

- *Retail and Leisure:* This variable shows mobility trends for places such as restaurants, cafes, museums, shopping centres, cinemas and libraries.
- *Supermarket and Pharmacy:* Locations such as supermarkets, grocery stores and pharmacies are represented by this variable.
- *Parks:* Shows mobility trends for places like national parks, public beaches, squares and gardens.
- *Public Transport:* This variable shows mobility trends for places that are hubs for public transport, such as railway stations, underground stations or bus stations.
- *Workplaces:* Mobility trends for places of work are shown in this variable.
- *Residential Areas:* This variable shows mobility trends for places of residence.
- *Air Passengers Arriving:* This shows the data for air passengers arriving at these locations.
- *Passengers Departing:* This shows the data for passengers departing from the places concerned.

Mobility on the date of the report is compared with mobility on the date of the reference value using the figure provided by GMD. The data is calculated for the date of the report (if the information is available). A positive or negative percentage is shown. The data shows how visiting (or time spent at) the categorised sites changes compared to our baseline. A baseline represents a normal value for the day of the week in question. The baseline is the average value for the 52-week period from the 1st of May, 2020 to the 30th of April 2020 in each state category, which includes, but is not limited to, Adamawa, Enugu, FCT, Kano, Lagos and Rivers states, as shown in Figures 2 to 7. The baseline is not a single value, but 7 individual values. Different percentage changes result from the same number of visitors on two different days of the week. Importantly, the baseline days never change. A hybrid ANN-CNN model including these

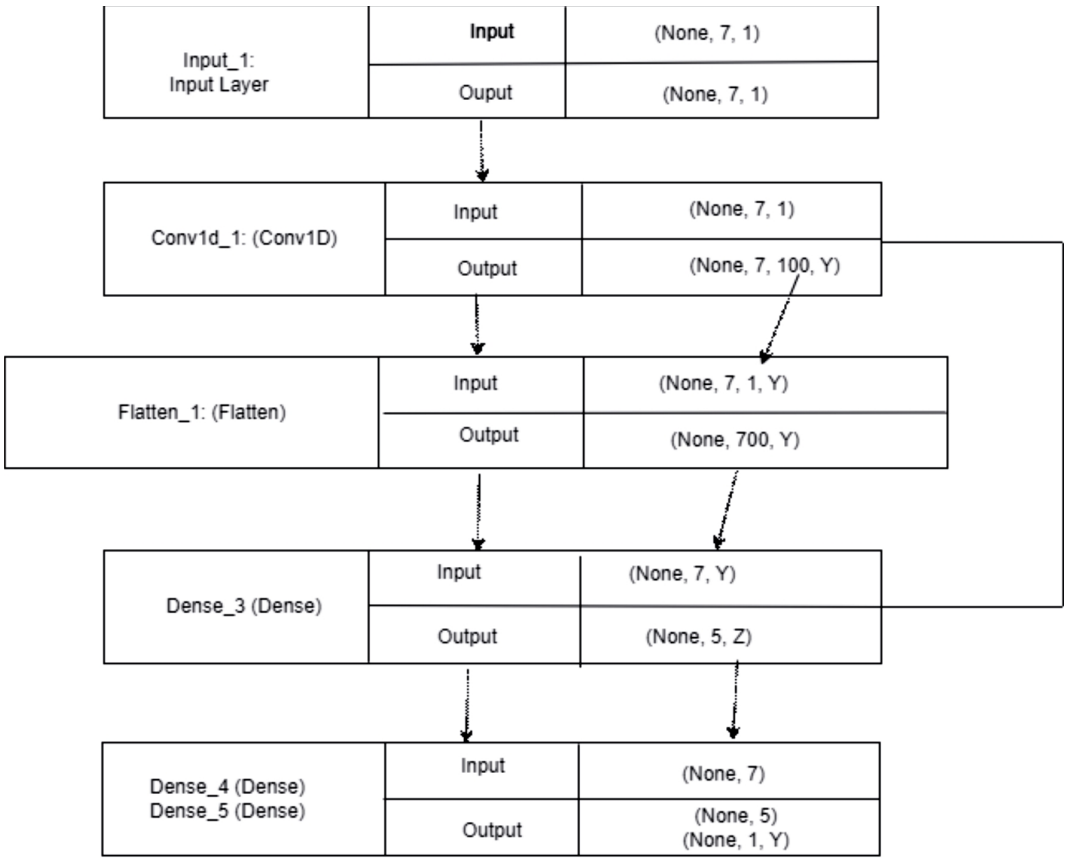


Figure 2. Architecture of the Hybrid ANN-CNN

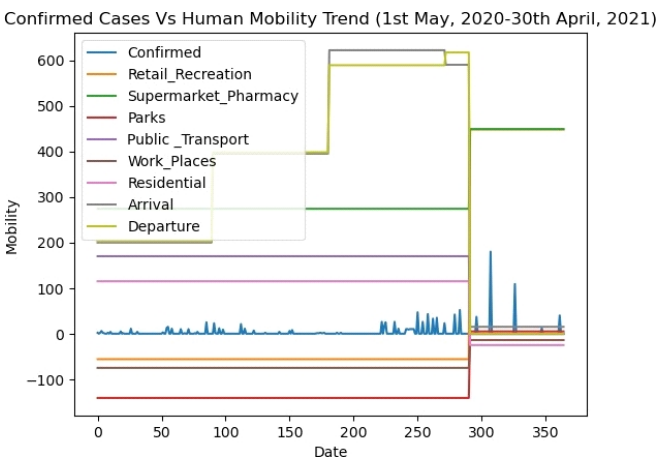


Figure 3. Adamawa

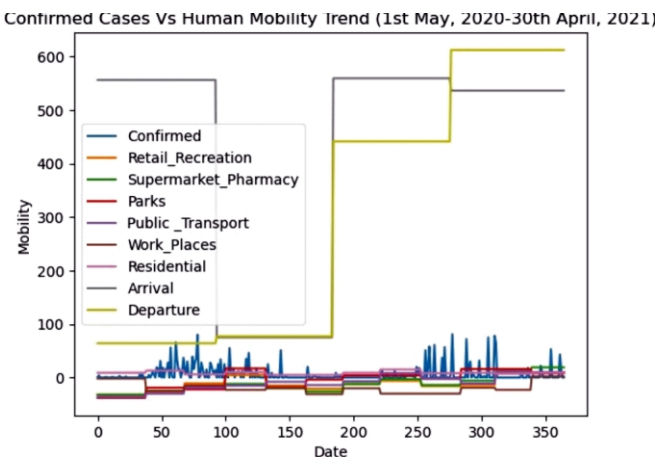


Figure 4. Enugu

variables was developed to predict human mobility against COVID-19. Equations (1) and (2) shows the equation for the developed hybrid ANN-CNN model. Let,

$$X = (x1t, x2t, \ldots..xnt) \rightarrow Y = (y1) \qquad (1)$$

Where X and Y are the Hybrid ANN-CNN inputs and outputs variables pairs at time t, respectively.

Suppose:

$$X(t) = D(tm),\ 1 \leq t \leq T \qquad (2)$$

D(tm), is the mobility change trend and includes six categories: food and pharmacy, parks, retail and recreation, transit stations, workplaces and homes, and air passenger arrival and departure on time t as

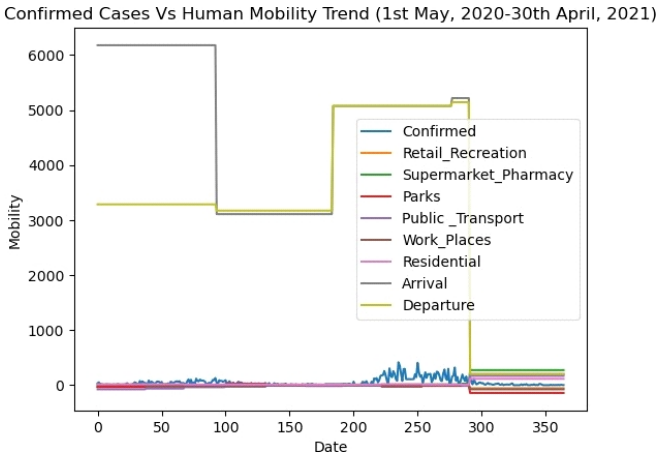


Figure 5. FCT

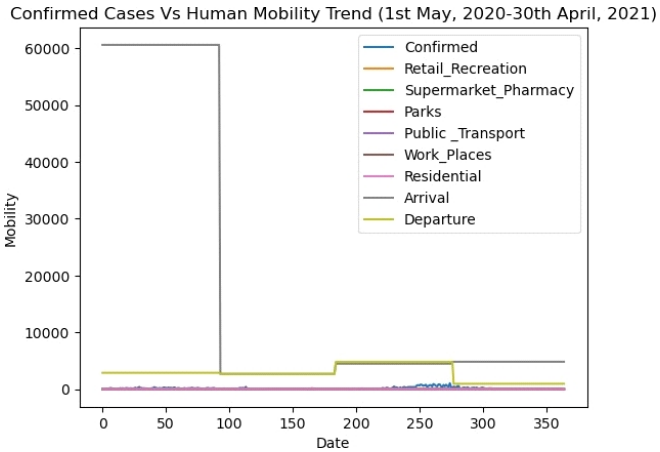


Figure 7. Lagos

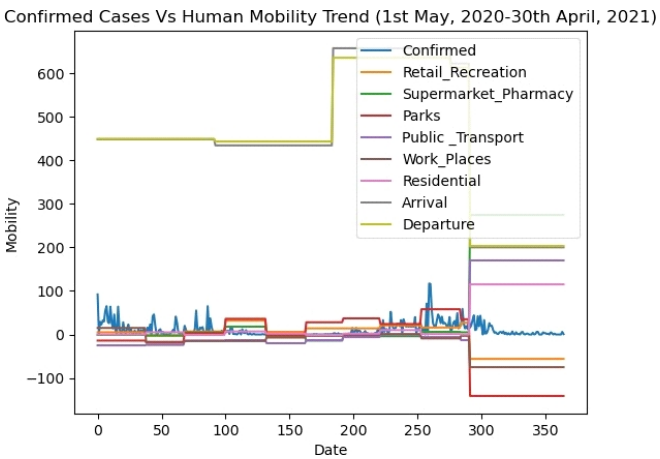


Figure 6. Kano

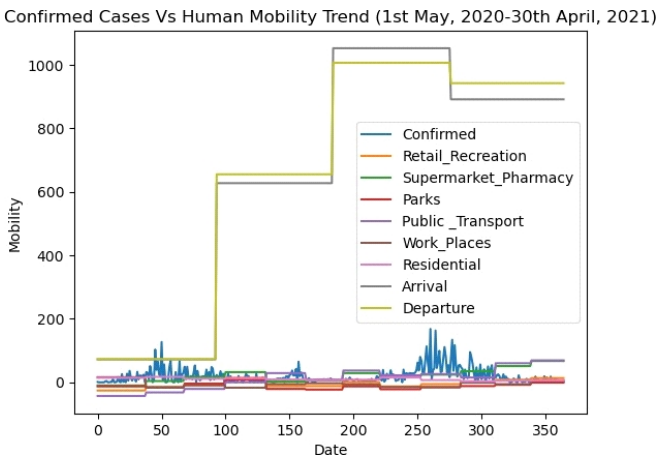


Figure 8. Rivers

independent variables against the corresponding dependent variables which is: number of daily confirmed cases and T is the number of days for the training datasets for both independent and dependent variables. Figures 3 to 8 showing one year confirmed cases distribution and human mobility trend.

2.6 Performance Evaluation Metrics

2.6.1 Mean Square Logarithmic Error (MSLE)

The mean squared logarithmic error (MSLE) measures the difference between the actual and expected values. With the addition of the logarithm, the MSLE is a measure of the percentage difference between the percentage difference between the actual value and the predicted value and as well as the relative difference between the two. As a result, MSLE will roughly handle tiny differences between small actual and expected values and large differences between large actual values and prediction values. The loss is the average over the observation data of the squared differences between the actual and of the squared differences between the actual value and the predicted value after logarithmic transformation, and is written as:

$$MSLE = \frac{1}{n}\sum_{i=1}^{N}\left(\log\left(1+\hat{y}i\right)-\log\left(1+yi\right)\right)^{2} \quad (3)$$

Where N is the number of data samples, yi is the true value and $\hat{y}_i$ is the predicted value.

2.6.2 Log Cosh

On the other hand, Log cosh works similarly to mean squared error, but is less affected by the occasional wildly incorrect prediction. It has all the advantages of Huber loss; unlike Huber loss, it is doubly differentiable (Jadon et al., 2022). The formula is written as:

$$LogCosh(x) = \sum_{i=1}^{N} (\log(Cosh(\hat{y}i - yi))) \quad (4)$$

2.6.3 Huber Loss

Huber loss is like 'patching' the square of the loss, which is more robust to anomalies. It behaves like squared loss for small errors, but behaves like absolute loss for large errors:

$$Huber(x) = \begin{cases} \frac{1}{2}x^2 \ for|x| \le \delta, \delta|x| - \frac{1}{2}\delta^2 \ otherwise. \end{cases} \quad (5)$$

where δ is a parameter that can be adjusted to control where the change takes place.

3. Results and Discussion

Table 1 shows the result of evaluation of the performance of five machine learning models involving ANN, CNN, LSTM, LASSO and Hybrid ANN- CNN in the prediction of human mobility against COVID-19 confirmed cases during lockdown period in Nigeria, covering the period from 1st May, 2020 to 30th April, 2021. The models were evaluated based on Huber loss as shown in Table 1, in Adamawa state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0143), followed by ANN (0.0105), LASSO (0.0156), LSTM (0.0174) and CNN (0.0696). In Enugu state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0014), followed by ANN (0.0015), LASSO (0.0021), LSTM (0.0040) and CNN (0.0555). Similarly in FCT, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0095), followed by ANN (0.0097), LASSO (0.0098), LSTM (0.0122) and CNN (0.0650).

Also in Kano state, the developed Hybrid ANN-CNN still outperformed the remaining four models with the Huber loss of (0.0087), followed by ANN (0.0092), LASSO (0.0099), LSTM (0.0116) and CNN (0.0634). In Lagos state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0116), followed by LASSO (0.0117), ANN (0.0120), LSTM (0.0145) and CNN (0.0668). In Rivers state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0077), followed by ANN (0.0080), LASSO (0.0089), LSTM (0.0104) and CNN (0.0629).

From Table 1, it can be deduced that, the developed Hybrid ANN-CNN outperformed the remaining four models with respect to Huber loss performance evaluation metrics, followed by ANN, LASSO, LSTM and CNN respectively.

Table 2 revealed the results of evaluation of the performance of five machine learning models involving ANN, CNN, LSTM, LASSO and Hybrid ANN-CNN in the prediction of human mobility against COVID-19 confirmed cases during lockdown period in Nigeria, covering from 1st May, 2020 to 30th April, 2021. The models were evaluated based on Mean Square Logarithmic Error (MSLE) as shown in Table 2, in Adamawa state, the developed Hybrid ANN-CNN outperformed the remaining four models with the MSLE of (0.0022), followed by ANN (0.0027), LASSO (0.0029), LSTM (0.0049) and CNN (0.0571). In Enugu state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Huber loss of (0.0171), followed by ANN (0.0179), LASSO (0.0186), LSTM (0.0203) and CNN (0.0706). Similarly in FCT, the developed Hybrid ANN-CNN outperformed the remaining four models with the MSLE of (0.0086), followed by LASSO (0.0119), ANN (0.0128), LSTM (0.0130) and CNN (0.0685). Also in Kano state, the developed Hybrid ANN-CNN still outperformed the remaining four models with the MSLE of (0.0116), followed by ANN (0.0122), LASSO (0.0125), LSTM (0.0145) and CNN (0.0674). In Lagos state,

State	ANN	CNN	LSTM	LASSO	Hybrid
Adamawa	0.0015	0.0555	0.0040	0.0021	0.0014
Enugu	0.0150	0.0696	0.0174	0.0156	0.0143
FCT	0.0097	0.0650	0.0122	0.0098	0.0095
Kano	0.0092	0.0634	0.0116	0.0099	0.0087
Lagos	0.0120	0.0668	0.0145	0.0117	0.0116
Rivers	0.0080	0.0629	0.0104	0.0089	0.0077

Table 1. Mobility Factors Vs Confirmed Cases (Huber Loss)

State	ANN	CNN	LSTM	LASSO	Hybrid
Adamawa	0.0027	0.0571	0.0049	0.0029	0.0022
Enugu	0.0179	0.0716	0.0203	0.0186	0.0171
FCT	0.0128	0.0685	0.0130	0.0119	0.0086
Kano	0.0122	0.0674	0.0145	0.0125	0.0116
Lagos	0.0142	0.0696	0.0166	0.0139	0.0135
Rivers	0.0112	0.0661	0.0135	0.0119	0.0108

Table 2. Mobility Factors Vs Confirmed Cases (MSLE)

the developed Hybrid ANN-CNN outperformed the remaining four models with the MSLE of (0.0135), followed by LASSO (0.0139), ANN (0.0142), LSTM (0.0166) and CNN (0.0696). In Rivers state, the developed Hybrid ANN-CNN outperformed the remaining four models with the MSLE of (0.0108), followed by ANN (0.0112), LASSO (0.0119), LSTM (0.0135) and CNN (0.0661). From Table 2, it can be deduced that, the developed Hybrid ANN-CNN outperformed the remaining four models with with respect to MSLE performance evaluation metrics, followed by ANN, LASSO, LSTM and CNN respectively.

Table 3 shows the results of evaluation of the performance of five machine learning models involving ANN, CNN, LSTM, LASSO and Hybrid ANN-CNN in the prediction of human mobility against COVID-19 confirmed cases during lockdown period in Nigeria, covering from 1st May, 2020 to 30th April, 2021. The models were evaluated based on Log Cosh as shown in Table 3, in Adamawa state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Log Cosh of (0.0013), followed by ANN (0.0015), LASSO (0.0024), LSTM (0.0040) and CNN (0.0562). In Enugu state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Log Cosh of (0.0135), followed by ANN (0.0141), LASSO (0.0603), LSTM (0.0623) and CNN (0.0685). Similarly in FCT, the developed Hybrid ANN-CNN out- performed the remaining four models with the Log Cosh of (0.0060), followed by LASSO (0.0063), LSTM (0.0064), ANN (0.0095) and CNN (0.0650). Also in Kano state, the developed Hybrid ANN-CNN still outperformed the remaining four models with the Log Cosh of (0.0084), followed by ANN (0.0089), LASSO (0.0097), LSTM (0.0114) and CNN (0.0639). In Lagos state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Log Cosh of (0.0112), followed by ANN (0.0116), CNN (0.0666), LASSO (0.0730) and LSTM (0.0754). In Rivers state, the developed Hybrid ANN-CNN outperformed the remaining four models with the Log Cosh of (0.0075), followed by ANN (0.0079), LASSO (0.0109), LSTM (0.0130) and CNN (0.0629).

State	ANN	CNN	LSTM	LASSO	Hybrid
Adamawa	0.0015	0.0562	0.0040	0.0024	0.0013
Enugu	0.0141	0.0685	0.0623	0.0603	0.0135
FCT	0.0095	0.0650	0.0064	0.0063	0.0060
Kano	0.0089	0.0639	0.0114	0.0097	0.0084
Lagos	0.0116	0.0666	0.0754	0.0730	0.0112
Rivers	0.0079	0.0629	0.0130	0.0109	0.0075

Table 3. Mobility Factors Vs Confirmed Cases (Log Cosh)

3.1 Line plots Showing Performance Comparison Evaluation of Models Based on Huber Loss, MSLE and Log Cosh

The line plots in Figures 9 to 14 show the performance comparison evaluation of five machine learning models including but not limited to ANN, CNN, LSTM, LASSO and Hybrid based on Huber Loss MSLE and Log Cosh in predicting human mobility factors against COVID-19 confirmed cases in six states of Nigeria that is, Adamawa, Enugu, FCT, Kano, Lagos and Rivers states. From the results as shown in the line plots, the developed hybrid ANN-CNN

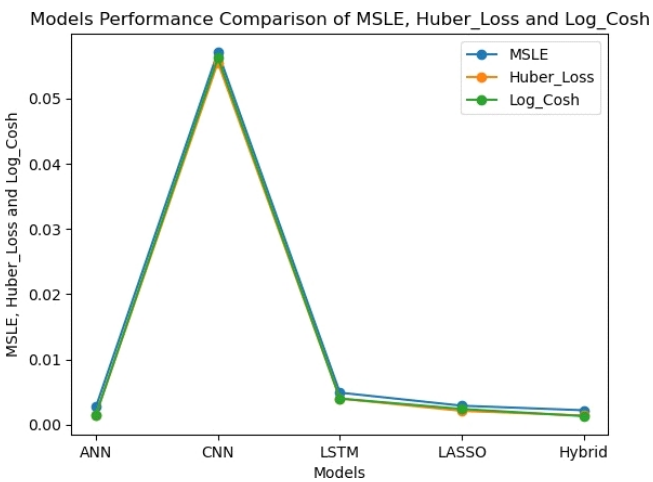


Figure 10. Adamawa

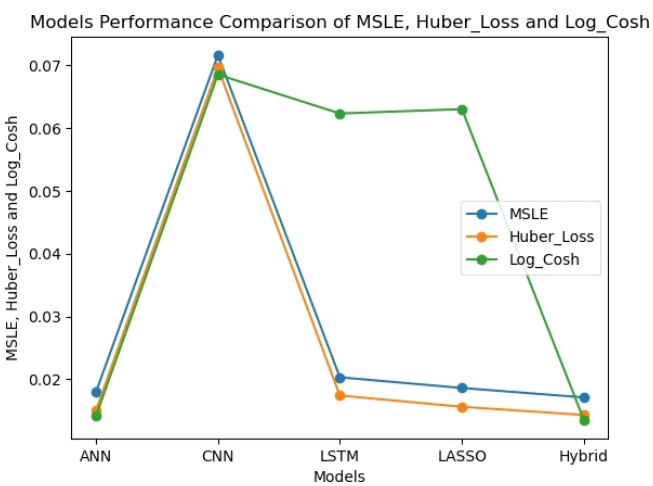


Figure 11. Enugu

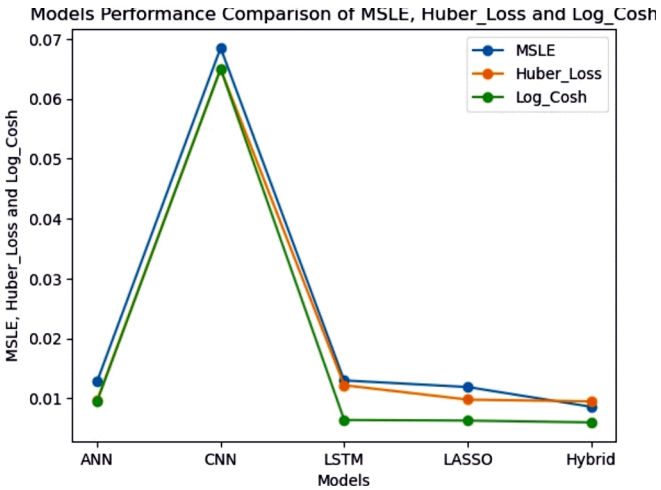


Figure 12. FCT

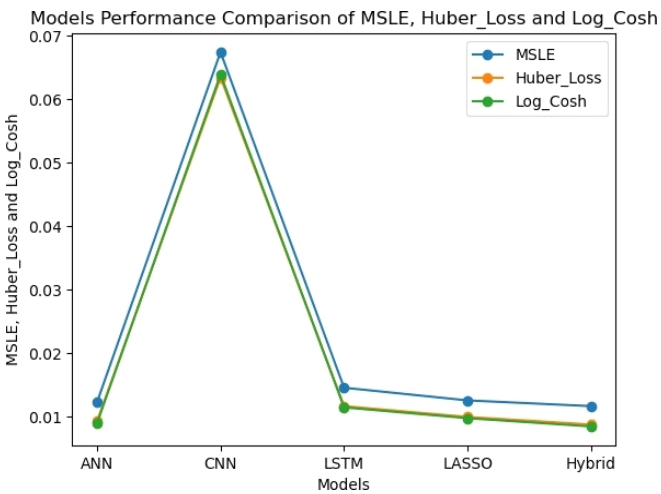


Figure 13. Kano

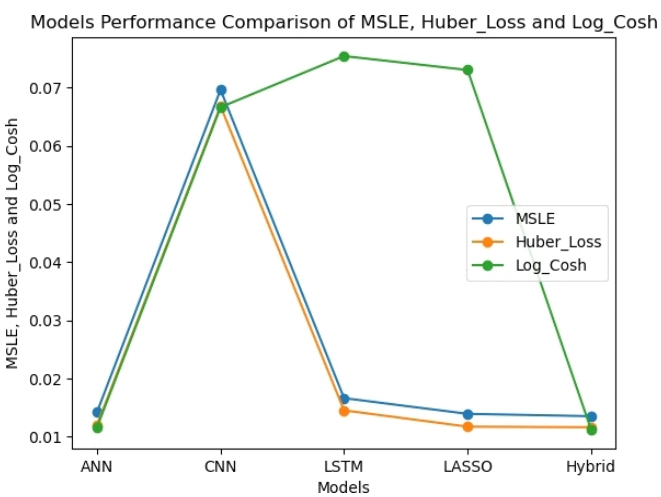


Figure 14. Lagos

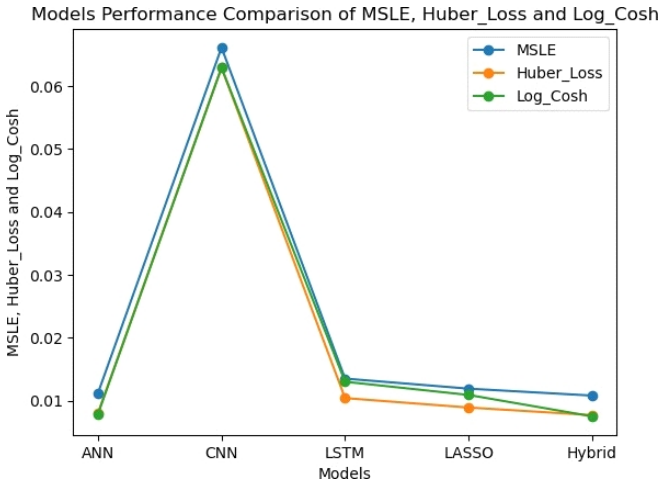


Figure 15. Rivers

outperformed all the four models followed by ANN, LASSO, LSTM and CNN respectively.

Conclusion

This study evaluated the impact of human mobility during COVID-19 lockdown and developed Hybrid ANN-CNN machine learning model that predict the impact of human mobility on COVID-19 confirmed cases for six Nigerian states, viz: Adamawa, Enugu, FCT, Kano, Lagos and Rivers states based on MSLE, Huber Loss and Log Cosh performance evaluation metrics. The models used in this study include but are not limited to: the developed Hybrid ANN-CNN, LASSO, ANN, CNN and LSTM. At the end of the experiment, the developed Hybrid ANN-CNN outperformed the other models based on MSLE, Huber Loss and Log Cosh performance evaluation metrics.

References

[1]. Athanasios, A., Irini, F., Tasioulis, T., & Konstantinos, K. (2021). Prediction of the effective reproduction number of COVID-19 in Greece: A machine learning approach using Google mobility data. *medRxiv.* https://doi.org/10.1101/2021.05.14.21257209

[2]. Erim, D. O., Oke, G. A., Adisa, A. O., Odukoya, O., Ayo-Yusuf, O. A., Erim, T. N., ... & Agaku, I. T. (2021). Associations of government-mandated closures and restrictions with aggregate mobility trends and SARS-CoV-2 infections in Nigeria. *JAMA Network Open,* 4(1), 1-11. https://doi.org/10.1001/jamanetworkopen.2020.32101

[3]. García-Cremades, S., Morales-García, J., Hernández-Sanjaime, R., Martínez-España, R., Bueno-Crespo, A., Hernández-Orallo, E., & Cecilia, J. M. (2021). Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Scientific Reports,* 11(1), 15173. https://doi.org/10.1038/s41598-021-94696-2

[4]. Hota, H. S., Handa, R., & Shrivas, A. K. (2021). COVID-19 pandemic in India: Forecasting using machine learning techniques. In Data Science for COVID-19 (pp. 503-525). *Academic Press.* https://doi.org/10.1016/B978-0-12-824536-1.00030-7

[5]. Ilin, C., Annan-Phan, S., Tai, X. H., Mehra, S., Hsiang, S., & Blumenstock, J. E. (2021). Public mobility data enables COVID-19 forecasting and management at local and global scales. *Scientific Reports,* 11(1), 13531. https://doi.org/10.1038/s41598-021-92892-8

[6]. Jadon, A., Patil, A., & Jadon, S. (2022). A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting. *arXiv.*

[7]. Khalifa, N. E., Mawgoud, A. A., Abu-Talleb, A., Taha, M. H. N., & Zhang, Y. D. (2023). A COVID-19 Infection Prediction Model in Egypt Based on Deep Learning Using Population Mobility Reports. *International Journal of Computational Intelligence Systems,* 16(1), 96. https://doi.org/10.1007/s44196-023-00272-z

[8]. Lawal, O., & Nwegbu, C. (2022). Movement and risk perception: evidence from spatial analysis of mobile phone-based mobility during the COVID-19 lockdown, Nigeria. *Geo Journal,* 87(3), 1543-1558. https://doi.org/10.1007/s10708-020-10331-z

[9]. Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security,* 11(11), 36. https://doi.org/10.5815/ijcnis.2017.11.04

[10]. Said, A. B., Erradi, A., Aly, H., & Mohamed, A. (2020). A Deep-Learning Model for Evaluating and Predicting the Impact of Lockdown Policies on COVID-19 Cases. *arXiv.*

[11]. Wang, R., Ji, C., Jiang, Z., Wu, Y., Yin, L., & Li, Y. (2021). A short-term prediction model at the early stage of the COVID-19 pandemic based on multisource urban data. *IEEE Transactions on Computational Social Systems,* 8(4), 938-945. https://doi.org/10.1109/TCSS.2021.3060952

ABOUT THE AUTHORS

Yahaya Mohammed Sani, Makerere University, Kampala, Uganda.

Benjamin Davou Pam, Makerere University, Kampala, Uganda.

A METHOD FOR THE IDENTIFICATION OF DENIAL OF SERVICE (DoS) ATTACK IN NETWORK TRAFFIC USING MACHINE LEARNING TECHNIQUES

By

GOTTAPU SANKARA RAO * P. KRISHNA SUBBARAO **

* Department of CSE, University College of Engineering, JNTUK, India.
** Department of CSE, GVPCE (A), Visakhapatnam, India.

Date Received: 06/11/2023 Date Revised: 15/11/2023 Date Accepted: 27/11/2023

ABSTRACT

Computer Networks and the internet are essential to our daily lives and enterprises. DoS assaults threaten computer networks and network security. The world is evolving toward online businesses and services. This has increased network traffic over time. We need NIDS and DoS attack detection since there are more network risks and attacks. DoS attacks now threaten computer network servers. This threat must be detected automatically to protect corporate assets. Anomaly-based intrusion detection was developed because signature-based DoS attack and intrusion detection methods are inadequate. Many studies employ Machine Learning and Deep Learning to detect network anamolies. This article describes classification models constructed with the aid of machine learning algorithms. On the own dataset, this research was performed utilizing machine learning algorithms including K-Nearest Neighbor (KNN), Logistic Regression, and Random Forest. Random Forest outperforms other supervised machine learning algorithms, as demonstrated by this study's findings. It achieved an accuracy rate of 99.62% when nine features were selected utilizing Pearson's correlation coefficient method. The own dataset file (myNetworkGenerateTraffic.csv) which was captured through wireshark tool were utilized to accurately evaluate machine learning algorithms. We utilized the following performance metrics in this investigation: Accuracy, Precision, Recall, and F-1 score. In this paper, we examine how machine learning techniques can improve DoS attack prediction in network traffic to better analyze network traffic and help improve network security.

Keywords: Network Security, Dataset, DoS Attack, Machine Learning.

INTRODUCTION

In recent years, Denial of Service (DoS) attacks have become a major network security issue. Today's linked world requires network security. Denial of Service (DoS) attacks threaten network availability and can harm businesses and individuals. Researchers have used machine learning, deep learning, and innovative methods to detect and mitigate DoS assaults. Maintaining network service availability and integrity requires detecting and mitigating DoS and DDoS assaults on network servers. Researchers have investigated deep learning, machine learning, and block chain-based systems to detect and mitigate these assaults. DoS/DDoS assaults flood a communication network with malicious data traffic and requests, making it unstable. Computer networks have a complex chain of nodes linked together. In this circumstance, providing a safe and efficient consumer environment is difficult. Many approaches have been used to discover and block DoS/DDoS attacks but none are fast or reliable. There is still plenty to do to improve DoS/DDoS protection. Over the past decade,

i-manager Publications
Member of:

This paper has objectives related to SDG

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

more companies have digitized their confidential data. With massive data creation, network traffic has skyrocketed. Computer networks have grown rapidly in the previous decade, notably with cloud computing and IoT. This data is difficult to secure. Also, network attacks have increased, and DoS attack, network intrusion is the biggest security risk (Wankhede & Kshirsagar, 2018; Zekri et al., 2017). DoS, Zero-day assaults are major issues in the IT world today. In this work DoS attack detection system detects network traffic classification. Signature-based and anomaly-based categorization can detect intrusion. Signature-based or pattern-based abnormalities are compared to database patterns. Signature-based attack or intrusion detection cannot learn unusual patterns and invasions in raw data. The anomaly-based attack or intrusion detection can identify normal or benign activity and abnormality (Aslan, 2022; Barati et al., 2014; Bhuyan et al., 2014; Cetinkaya et al., 2019; Kim et al., 2020; Kowski and Ksiezopolski, 2021; Loukas & Oke , 2010; Lima Filho et al., 2019; Masdari & Jalali, 2016; Nandi et al., 2020; Perez-Diaz et al., 2020; Rahman et al., 2019; Rao & Subbarao, 2023; Shamsolmoali & Zareapoor ,2014; Shah et al., 2023; Smys et al., 2020; Tayyab et al., 2020; Ulemale ,2022; Wankhede & Kshirsagar, 2018; Wang et al., 2019; Wu et al., 2018; Yuan et al., 2017; Yuso et al., 2017; Zekri et al., 2017).

It can learn anomalous patterns using Machine Learning and Deep Learning. DoS attack or IDS inputs include traffic logs, application logs, file system modifications, and packets, whereas output is the label for attack type or normal type common security threats which include unauthorized denial of service, failure to authenticate, and violation of confidentiality, which is depicted in Figure 1. Many embranchment names explain different DoS kinds. DDoS implies several, unaffiliated attackers. The subclass of DoS attack is DDoS.

Categories of DDoS attacks are explained below:

ICMP Flood

An adversary transmits to a target a large number of ICMP packets (including Echo Requests and other types) in an ICMP flood attack, thereby depleting the target's

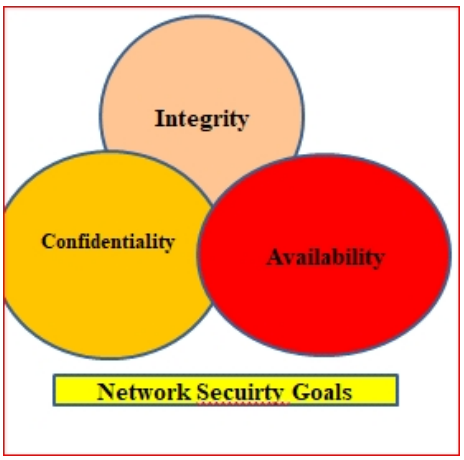


Figure 1. Network Security Goals

resources and potentially inducing a denial of service. (Rao & Subbarao, 2023).

TCP-SYN Flood

Any internet-connected device can be targeted by this DoS attack. A SYN flood occurs when the sender makes multiple SYN requests from a bogus IP address notwithstanding the host's SYN-ACK. It makes so many half connections. Requests coming from trustworthy clients causes denial of service (Figure 2).

Ping of Death

In this attack, someone sends a ping message that isn't formatted correctly or is too big to fit on a target system. This makes the target system crash or stop responding.

HTTP Flood

This flaw operates via HTTP requests. Upon transmitting an HTTP request to the recipient, the assailant proceeds with the execution of the attack (Aslan, 2022; Barati et al., 2014; Bhuyan et al., 2014; Cetinkaya et al., 2019; Kim et al., 2020; Kowski and Ksiezopolski, 2021; Loukas & Oke , 2010; Lima Filho et al., 2019; Masdari & Jalali, 2016; Nandi

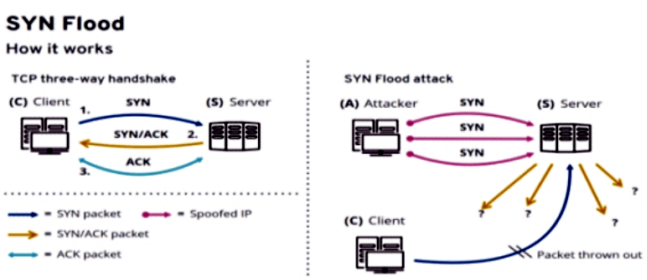


Figure 2. TCP- SYN Flood Attack

et al., 2020; Perez-Diaz et al., 2020; Rahman et al., 2019; Rao & Subbarao, 2023; Shamsolmoali & Zareapoor ,2014; Shah et al., 2023; Smys et al., 2020; Tayyab et al., 2020; Ulemale ,2022; Wankhede & Kshirsagar, 2018; Wang et al., 2019; Wu et al., 2018; Yuan et al., 2017; Yuso et al., 2017; Zekri et al., 2017).

1. Literature Review

Zekri et al. (2017) predicted DDoS attacks more accurately with decision tree C4.5 than with conventional machine learning algorithms. ML, neural network, RF, and MLP classifiers were utilized by Wankhede and Kshirsagar (2018) to categorize the dataset as benign or dos attack. Yuan et al. (2017) propose a deep learning-based DDoS detection system and found Deep Defense reduced error rate by 39.69%. ML-based IDS by Tayyab et al. (2020) uses ensemble learning and collaborative architecture to detect DoS and DDoS attacks. GA and Artificial ANN helped Barati et al. (2014) detect DDoS attacks accurately and without false alarms. A machine learning-based internet traffic monitoring system that detects real-time DDoS attacks via spark streaming. This system was evaluated in comparison to the Naïve Bayes, Logistic Regression, and Decision Tree methodologies by the authors. Yuso et al. (2017) developed a superior intrusion detection feature selection system. Rahman et al. (2019) found that J48 outperformed JRF, SVM, and KNN in SDN network DDoS detection and prevention training and testing. Shamsolmoali and Zareapoor (2014) discovered that C2DF was more precise and quicker when employing a statistical method to detect and filter DDoS attacks. Aslan (2022) observed that the suggested classifier system efficiently differentiated DDoS traffic from normal traffic. A Hybrid technique that selects the most important attributes provides the highest DDoS detection rate, according to Nandi et al. (2020). This class visualizes Wu et al. (2018) multi-dimensional DDoS data (Wu et al., 2018). Identification of attack type helps detect DDoS attacks. Lima Filho et al. (2019) introduced the clever detection system, an online detection system for dos/ddos attacks that classified network traffic using a random forest tree algorithm and increased DR, FAR, PRECT raditional testing. Kozlowski & Ksiezopolski (2021) showed machine learning models were accurate against UDP DDoS attacks. In the study, a comparison was made between various DDoS attack detection methods, including Linear Regression, Random Forest, Support Vector Machine, Gaussian, and Naïve Bayes, by Ulemale (2022). Loukas and Oke (2010) review denial of service protection research extensively. Recent research and the hardest defense tasks are emphasized in the survey. It discusses DoS attacks and defenses.

Cetinkaya et al. (2019) explore estimating the uncertainty in denial-of-service attack techniques. DoS assaults are dynamic and adaptive, therefore this is crucial. Understanding assault models helps create successful defenses. Bhuyan et al. (2014) explore DDoS detection methodologies, tools, and future approaches. DDoS assaults target services or network resources with huge coordination. Understanding DDoS attack tactics helps improve detection and mitigation strategies. The cloud is also subject to DoS assaults. Masdari and Jalali (2016) analyze cloud computing DoS attacks and examine state-of-the-art literature methods to prevent, detect, and mitigate each type of attack. Smys et al. (2020) study denial of service attack tactics in the IoT. Understanding attack techniques can help design IoT-specific protection measures as attackers target IoT devices. SDN is another area where machine learning and deep learning can identify and prevent DoS attacks. Wang et al. (2019) present a safe system that periodically collects forwarding element network statistics and uses machine learning classification methods to identify misbehavior and new flow attacks. Kim et al. (2020) offer a machine learning and deep learning-based IoT botnet detection approach. IoT device security requires detecting botnets, which initiate DDoS assaults. Block chain-based solutions may minimize IoTDDoS attacks. Shah et al. (2023) review block chain-based IoTDDoS mitigation strategies in the literature. These solutions use block chain's decentralization and immutability to protect IoT networks from DDoS attacks (Shah et al., 2023). Other DoS/DDoS detection and mitigation methods have been investigated besides deep learning, machine learning, and block chain. A flexible SDN-based architecture can

detect and mitigate low-rate DDoS attacks, (Perez-Diaz et al., 2020). This architecture uses SDN principles to adjust network configurations and implement mitigation measures in real-time to counter attacks (Perez-Diaz et al., 2020). The literature analysis found that deep learning, machine learning, and block chain-based solutions are being studied to detect and mitigate network server DoS/DDoS assaults. These strategies may boost network infrastructure resilience and security. More research is needed to develop more robust and effective detection and mitigation strategies for developing and complex DoS/DDoS threats.

2. Experiment # 1

2.1 Proposed Methodology

2.1.1 Dataset Preparation

Wireshark captured streaming packets are sniffed for making this own dataset. A well-known open-source network analyzer is Wireshark. Packets are captured by Wireshark via Pcap, Ethernet, Wi-Fi, Bluetooth, and additional network media. My network Generated traffic dataset has 504576 traffic records. Table 1 lists 9 dataset attributes used for model training. The features of the dataset consist of the following: source IP address, destination IP address, protocol type (ICMP, TCP, UDP, DNS), packet length, total length, packet information. destination port number, label, and source port number. These attributes are utilized to distinguish between regular type ("0") and attack type ("1") traffic. The dataset consists of 1983 DNS instances, 35409 TCP, 10264 ICMP, and 1152 UDP instances. A assault traffic label in the dataset is 1, while normal traffic is labeled 0.

2.1.2 Proposed Model Architecture

The recommended work methodology step by step is drawn in Figure 1 and it is implemented using python in jupyter notebook. This suggested system classifies the captured Network traffic as benign network traffic type or DoS attack type network traffic. The main parts of this suggested system are pre-processing (selecting

Time	Source IP Address	Destination IP Address	Protocol	Length	Source port	Destination Port	Label length	Packet information

Table 1. Own Dataset 9- Attributes

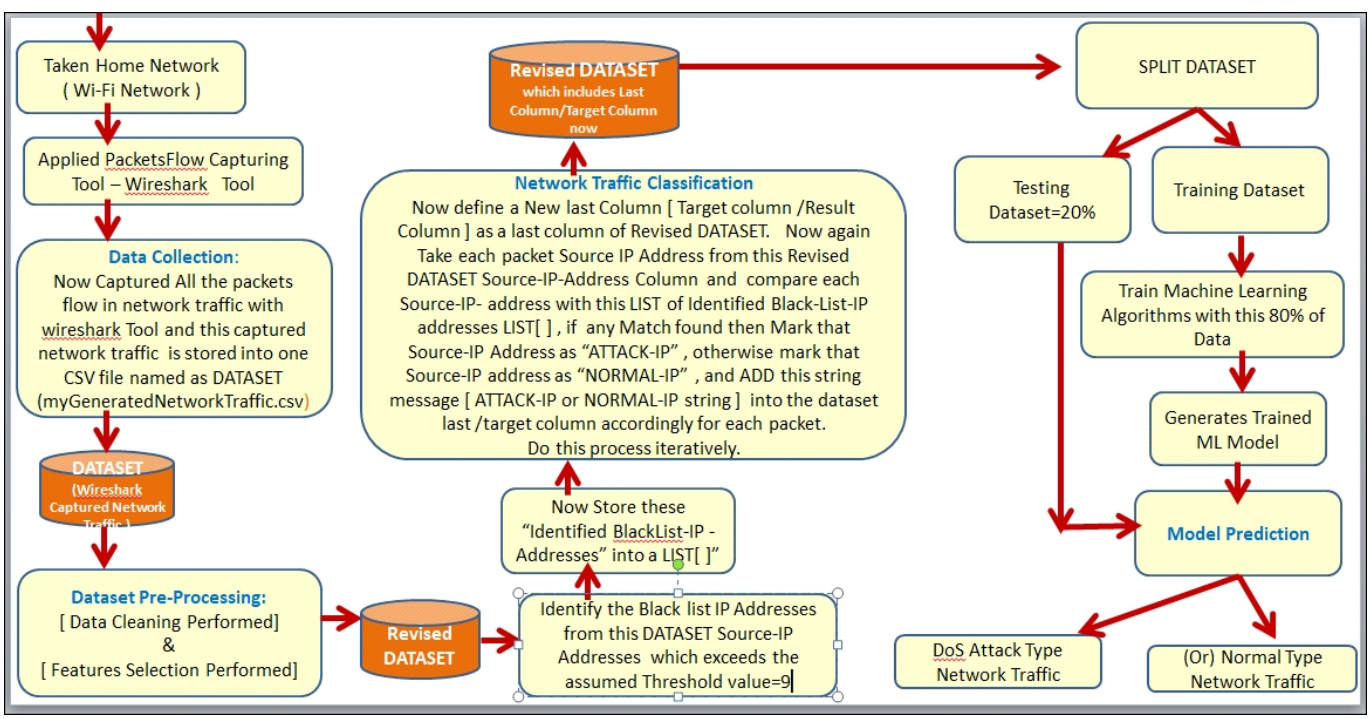


Figure 1. Architecture of Proposed Model

attributes) in a way that can be changed, and figuring out the type of traffic, which can be either dos type malicious traffic or normal network traffic.

2.1.3 Proposed Methodology Working Steps

Experiment#1 steps:

Step 1: Data Collection step: In the first step we applied Wireshark and Winpcap Tool and started wireshark on the required network and captured the full live network traffic (i.e. all the packets flow) on the network up to certain period of time. Packets flow capture done by wireshark tool is shown in Figure 2.

Step 2: This resultant captured packets traffic is stored into one log file (named as myNetworkGeneratedTraffic. csv file) to implement my experiment.

Step 3: This Generated CSV file is considered as network traffic DATASET (own dataset) for conducting the experiment. This dataset is depicted below in Figure 3 & Figure 4.

```
In [7]: df_r.shape
Out[7]: (504576, 10)
```

The number of instances of each protocol wise in this own dataset is shown here in Figure 5.

Step 4: Dataset Pre-Processing step: Next step is, this GeneratedNetworkTraffic.csv file [DATASET] will be pre-processed. After this step we get REVISED DATASET.

Step 5: Here in preprocessing Dataset step, clean-up and feature extraction are done. Feature selection is crucial in machine learning processing. We need to find a dataset with few attributes that may accurately identify which type of traffic it is. Essential Features selected from this dataset use Pearson correlation. The selected features are = ["Time", "Protocol", "Length", "Source_port", "Destination_port"].

Step 6: Missing values, inappropriate features, categorical data, and other faults in dataset may prohibit a machine learning algorithm from interpreting a dataset. So here in Data Cleaning step we filled null values with mean. Some methods remove rows or columns with a certain amount of NaN values, while other methods replace missing values with the mean, median, mode, or

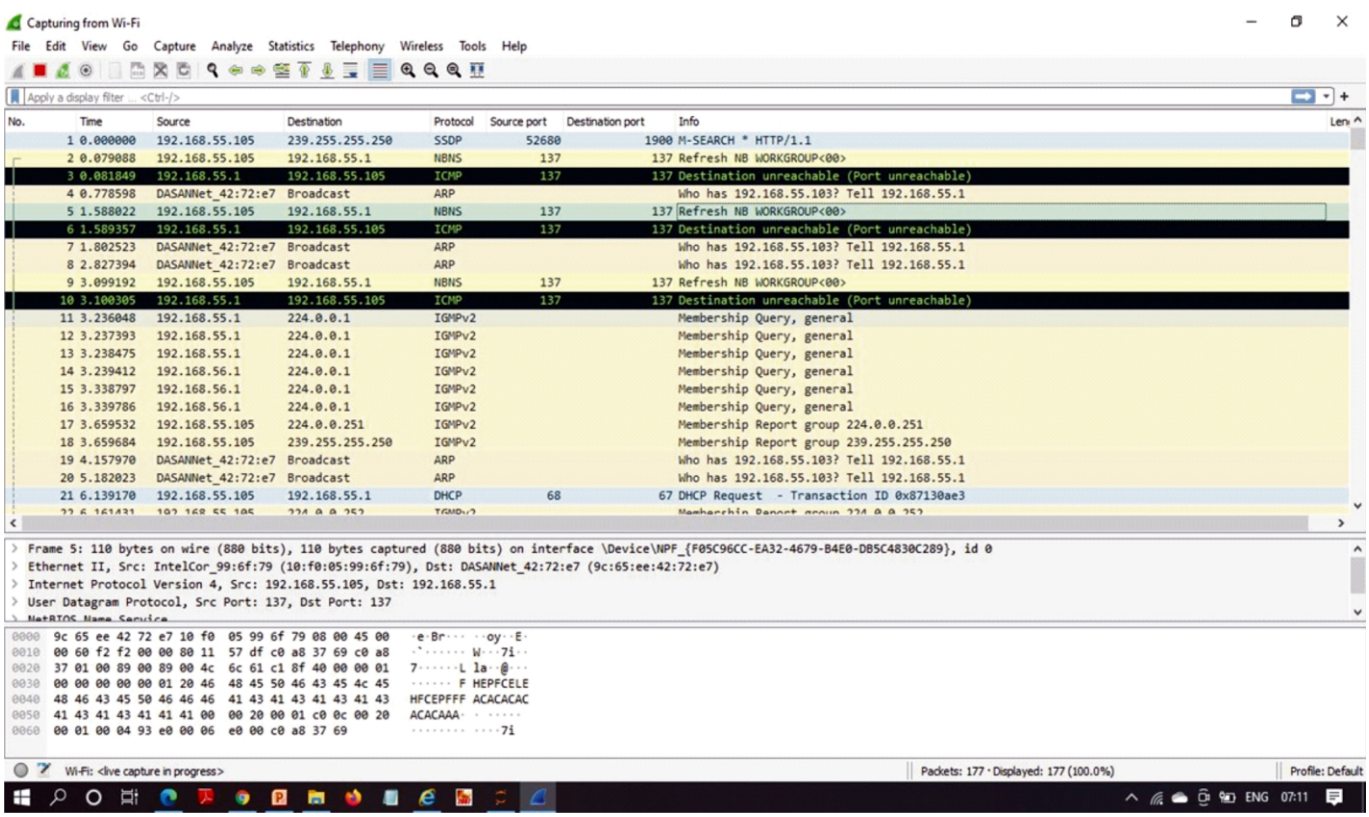
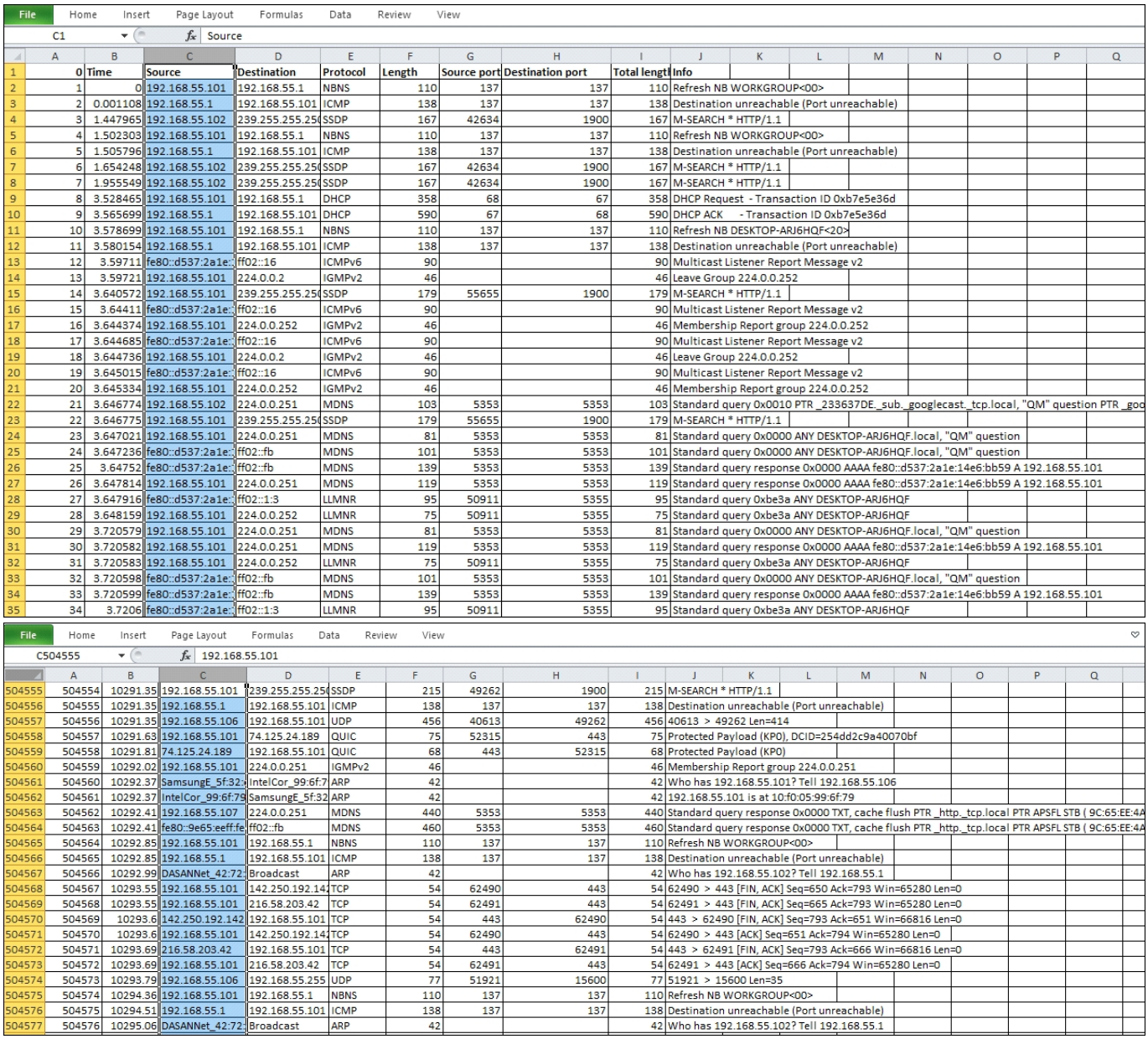


Figure 2. Wire Shark Captured Network Traffic Result – GUI Interface



	0	Time	Source	Destination	Protocol	Length	Source port	Destination port	Total length	Info
1	0		192.168.55.101	192.168.55.1	NBNS	110	137	137	110	Refresh NB WORKGROUP<00>
2	1	0.001108	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
3	2	1.447965	192.168.55.102	239.255.255.25(	SSDP	167	42634	1900	167	M-SEARCH * HTTP/1.1
4	3	1.502303	192.168.55.101	192.168.55.1	NBNS	110	137	137	110	Refresh NB WORKGROUP<00>
5	4	1.505796	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
6	5	1.654248	192.168.55.102	239.255.255.25(	SSDP	167	42634	1900	167	M-SEARCH * HTTP/1.1
7	6	1.955549	192.168.55.102	239.255.255.25(	SSDP	167	42634	1900	167	M-SEARCH * HTTP/1.1
8	7	3.528465	192.168.55.101	192.168.55.1	DHCP	358	68	67	358	DHCP Request - Transaction ID 0xb7e5e36d
9	8	3.565699	192.168.55.1	192.168.55.101	DHCP	590	67	68	590	DHCP ACK - Transaction ID 0xb7e5e36d
10	9	3.578699	192.168.55.101	192.168.55.1	NBNS	110	137	137	110	Refresh NB DESKTOP-ARJ6HQF<20>
11	10	3.580154	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
12	11	3.59711	fe80::d537:2a1e:	ff02::16	ICMPv6	90			90	Multicast Listener Report Message v2
13	12	3.59721	192.168.55.101	224.0.0.2	IGMPv2	46			46	Leave Group 224.0.0.252
14	13	3.640572	192.168.55.101	239.255.255.25(	SSDP	179	55655	1900	179	M-SEARCH * HTTP/1.1
15	14	3.64411	fe80::d537:2a1e:	ff02::16	ICMPv6	90			90	Multicast Listener Report Message v2
16	15	3.644374	192.168.55.101	224.0.0.252	IGMPv2	46			46	Membership Report group 224.0.0.252
17	16	3.644685	fe80::d537:2a1e:	ff02::16	ICMPv6	90			90	Multicast Listener Report Message v2
18	17	3.644736	192.168.55.101	224.0.0.2	IGMPv2	46			46	Leave Group 224.0.0.252
19	18	3.645015	fe80::d537:2a1e:	ff02::16	ICMPv6	90			90	Multicast Listener Report Message v2
20	19	3.645334	192.168.55.101	224.0.0.252	IGMPv2	46			46	Membership Report group 224.0.0.252
21	20	3.646774	192.168.55.102	224.0.0.251	MDNS	103	5353	5353	103	Standard query 0x0010 PTR _233637DE._sub._googlecast._tcp.local, "QM" question PTR _goo
22	21	3.646775	192.168.55.101	239.255.255.25(	SSDP	179	55655	1900	179	M-SEARCH * HTTP/1.1
23	22	3.647021	192.168.55.101	224.0.0.251	MDNS	81	5353	5353	81	Standard query 0x0000 ANY DESKTOP-ARJ6HQF.local, "QM" question
24	23	3.647236	fe80::d537:2a1e:	ff02::fb	MDNS	101	5353	5353	101	Standard query 0x0000 ANY DESKTOP-ARJ6HQF.local, "QM" question
25	24	3.64752	fe80::d537:2a1e:	ff02::fb	MDNS	139	5353	5353	139	Standard query response 0x0000 AAAA fe80::d537:2a1e:14e6:bb59 A 192.168.55.101
26	25	3.647814	192.168.55.101	224.0.0.251	MDNS	119	5353	5353	119	Standard query response 0x0000 AAAA fe80::d537:2a1e:14e6:bb59 A 192.168.55.101
27	26	3.647916	fe80::d537:2a1e:	ff02::1:3	LLMNR	95	50911	5355	95	Standard query 0xbe3a ANY DESKTOP-ARJ6HQF
28	27	3.648159	192.168.55.101	224.0.0.252	LLMNR	75	50911	5355	75	Standard query 0xbe3a ANY DESKTOP-ARJ6HQF
29	28	3.720579	192.168.55.101	224.0.0.251	MDNS	81	5353	5353	81	Standard query 0x0000 ANY DESKTOP-ARJ6HQF.local, "QM" question
30	29	3.720582	192.168.55.101	224.0.0.251	MDNS	119	5353	5353	119	Standard query response 0x0000 AAAA fe80::d537:2a1e:14e6:bb59 A 192.168.55.101
31	30	3.720583	192.168.55.101	224.0.0.252	LLMNR	75	50911	5355	75	Standard query 0xbe3a ANY DESKTOP-ARJ6HQF
32	31	3.720598	fe80::d537:2a1e:	ff02::fb	MDNS	101	5353	5353	101	Standard query 0x0000 ANY DESKTOP-ARJ6HQF.local, "QM" question
33	32	3.720599	fe80::d537:2a1e:	ff02::fb	MDNS	139	5353	5353	139	Standard query response 0x0000 AAAA fe80::d537:2a1e:14e6:bb59 A 192.168.55.101
34	33	3.7206	fe80::d537:2a1e:	ff02::1:3	LLMNR	95	50911	5355	95	Standard query 0xbe3a ANY DESKTOP-ARJ6HQF

504555	504554	10291.35	192.168.55.101	239.255.255.25(	SSDP	215	49262	1900	215	M-SEARCH * HTTP/1.1
504556	504555	10291.35	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
504557	504556	10291.35	192.168.55.106	192.168.55.101	UDP	456	40613	49262	456	40613 > 49262 Len=414
504558	504557	10291.63	192.168.55.101	74.125.24.189	QUIC	75	52315	443	75	Protected Payload (KP0), DCID=254dd2c9a40070bf
504559	504558	10291.81	74.125.24.189	192.168.55.101	QUIC	68	443	52315	68	Protected Payload (KP0)
504560	504559	10292.02	192.168.55.101	224.0.0.251	IGMPv2	46			46	Membership Report group 224.0.0.251
504561	504560	10292.37	SamsungE_5f:32:	IntelCor_99:6f:7	ARP	42			42	Who has 192.168.55.101? Tell 192.168.55.106
504562	504561	10292.37	IntelCor_99:6f:79	SamsungE_5f:32	ARP	42			42	192.168.55.101 is at 10:f0:05:99:6f:79
504563	504562	10292.41	192.168.55.107	224.0.0.251	MDNS	440	5353	5353	440	Standard query response 0x0000 TXT, cache flush PTR _http._tcp.local PTR APSFL STB (9C:65:EE:4A
504564	504563	10292.41	fe80::9e65:eeff:fe	ff02::fb	MDNS	460	5353	5353	460	Standard query response 0x0000 TXT, cache flush PTR _http._tcp.local PTR APSFL STB (9C:65:EE:4A
504565	504564	10292.85	192.168.55.101	192.168.55.1	NBNS	110	137	137	110	Refresh NB WORKGROUP<00>
504566	504565	10292.85	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
504567	504566	10292.99	DASANNet_42:72:	Broadcast	ARP	42			42	Who has 192.168.55.102? Tell 192.168.55.1
504568	504567	10293.55	192.168.55.101	142.250.192.142	TCP	54	62490	443	54	62490 > 443 [FIN, ACK] Seq=650 Ack=793 Win=65280 Len=0
504569	504568	10293.55	192.168.55.101	216.58.203.42	TCP	54	62491	443	54	62491 > 443 [FIN, ACK] Seq=665 Ack=793 Win=65280 Len=0
504570	504569	10293.6	142.250.192.142	192.168.55.101	TCP	54	443	62490	54	443 > 62490 [FIN, ACK] Seq=793 Ack=651 Win=66816 Len=0
504571	504570	10293.6	192.168.55.101	142.250.192.142	TCP	54	62490	443	54	62490 > 443 [ACK] Seq=651 Ack=794 Win=65280 Len=0
504572	504571	10293.69	216.58.203.42	192.168.55.101	TCP	54	443	62491	54	443 > 62491 [FIN, ACK] Seq=793 Ack=666 Win=66816 Len=0
504573	504572	10293.69	192.168.55.101	216.58.203.42	TCP	54	62491	443	54	62491 > 443 [ACK] Seq=666 Ack=794 Win=65280 Len=0
504574	504573	10293.79	192.168.55.106	192.168.55.255	UDP	77	51921	15600	77	51921 > 15600 Len=35
504575	504574	10294.36	192.168.55.101	192.168.55.1	NBNS	110	137	137	110	Refresh NB WORKGROUP<00>
504576	504575	10294.51	192.168.55.1	192.168.55.101	ICMP	138	137	137	138	Destination unreachable (Port unreachable)
504577	504576	10295.06	DASANNet_42:72:	Broadcast	ARP	42			42	Who has 192.168.55.102? Tell 192.168.55.1

Figure 3. Dataset Values View (.csv file view) in Excel Sheet

```
In [2]: df_r=pd.read_csv("D:\\my_network3.csv") #network traffic file
In [3]: df_r.head()
Out[3]:
```

	No.	Time	Source	Destination	Protocol	Length	Source port	Destination port	Total length	Info
0	1	0.000000	192.168.55.101	192.168.55.1	NBNS	110	137.0	137.0	110	Refresh NB WORKGROUP<00>
1	2	0.001108	192.168.55.1	192.168.55.101	ICMP	138	137.0	137.0	138	Destination unreachable (Port unreachable)
2	3	1.447965	192.168.55.102	239.255.255.250	SSDP	167	42634.0	1900.0	167	M-SEARCH * HTTP/1.1
3	4	1.502303	192.168.55.101	192.168.55.1	NBNS	110	137.0	137.0	110	Refresh NB WORKGROUP<00>
4	5	1.505796	192.168.55.1	192.168.55.101	ICMP	138	137.0	137.0	138	Destination unreachable (Port unreachable)

```
In [6]: df_r.tail()
Out[6]:
```

	No.	Time	Source	Destination	Protocol	Length	Source port	Destination port	Total length	Info
504571	504572	10293.692061	192.168.55.101	216.58.203.42	TCP	54	62491.0	443.0	54	62491 > 443 [ACK] Seq=666 Ack=794 Win=65280 ...
504572	504573	10293.794011	192.168.55.106	192.168.55.255	UDP	77	51921.0	15600.0	77	51921 > 15600 Len=35
504573	504574	10294.357831	192.168.55.101	192.168.55.1	NBNS	110	137.0	137.0	110	Refresh NB WORKGROUP<00>
504574	504575	10294.509261	192.168.55.1	192.168.55.101	ICMP	138	137.0	137.0	138	Destination unreachable (Port unreachable)
504575	504576	10295.060886	DASANNet_42:72:e7	Broadcast	ARP	42	NaN	NaN	42	Who has 192.168.55.102? Tell 192.168.55.1

Figure 4. Dataset Top View and Bottom View in jupyter Notebook

other statistical measures of a column, row, or set of data.

Step 7: Do this step iteratively i.e. identifying the Blacklist IP Addresses from this revised dataset column named "source-IP addresses column". Means identify the list of Blacklist IP-Addresses from the full list of existing IP values in that column, further in this step for each IP "maximum Requests Count" assumed threshold value is =9 , So take soure-IPs column each IP-Address value ,and determine

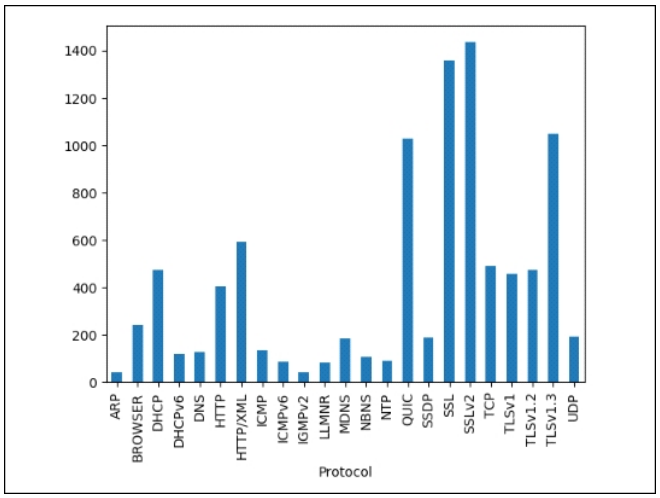


Figure 5. Protocol Wise Packets Instances Count

the number of times a selected source IP makes a request to a unique destination IP address. If the count of "same packet requests made" surpasses the presumptive threshold value of nine, that source IP address is marked as a blacklisted IP (one that is considered suspect or malicious). otherwise marked that IP as normal ip address. Such manner the black listed-IPaddresses list is identified. Some of the identified blacklist (suspicious) IP addresses are depicted in Figure 6.

Step 8: Traffic Classification step: Next we appended a new column [Target/Result column] to existing dataset features at the last column of DATASET, now take each packet "source IP Address" from full list of "source-IP" column values of the DATASET, and verify whether any of

```
Each Source IP ,        its Requests Made count
192.168.55.101 7154
192.168.55.1 6888
192.168.55.102 645
192.168.55.101 3458
192.168.55.1 1355
192.168.55.102 4055
192.168.55.102 4290
192.168.55.101 188
192.168.55.1 4846
192.168.55.101 4480
192.168.55.1 1604
fe80::d537:2a1e:14e6:bb59 3583
192.168.55.101 365
192.168.55.101 345
fe80::d537:2a1e:14e6:bb59 324
192.168.55.101 320
fe80::d537:2a1e:14e6:bb59 3
192.168.55.101 2
```

Figure 6. Some of the Identified Blacklist IP Addresses + Count [No of Times Requests Made by that IP].

these source-IP addresses match the Identified Blacklist-IPs above; if a match is found, label the address as "ATTACK-IP"; otherwise, indicate it as "Normal-IP" in the final column (i.e., add the string message "ATTACK IP" or "Normal IP" to the newly defined dataset's final column/Result column, as illustrated in Figure 7. After this step we get RE-Revised dataset depicted in Figure 8 which includes result column or Target column as a last column along with previous 9 columns of dataset.

Step 9: Split Re-Revised dataset: This re-revised dataset is divided into train and test data with ratio of 80%, 20% respectively for training ML models.

Step 10: Model Training step: The model was trained using Logistic Regression(), KNeighborsClassifier (n_neighbors= 5), RandomForestClassifier() with the 80% of whole data.

Step 11: Model Prediction step: At the last step the trained ML model predicts based on the input test data i.e. this model predicts whether input traffic is DoS Type (malicious) traffic or Normal type network traffic. And then accuracy of dos kind network traffic detection is calculated. Finally k-fold cross validation is used to validate this accuracy.

The code snippet for the Model training & Model Evaluating for accuracy detection is shown in Figure 9. The drive link is given here to access and test my experiment

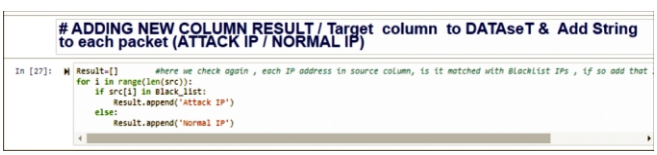


Figure 7. Preparing Last Column of the Dataset

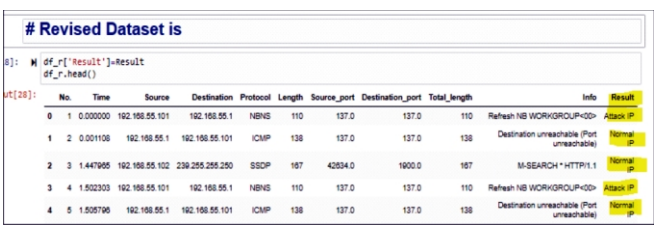


Figure 8. RE-Revised DATASET with target column

https://drive.google.com/drive/folders/1XiC4mKovW0P-aWZgKZD9WI_YqW1fVBA7?usp=sharing

Figure 9. Drive Link For Accessing the & Testing My Code for this Experiment

code. [just copy and paste this link into the browser to access the own dataset + implemented code]

3. Experiment # 1 Results

The implementation of this Experiment #1 using different Machine Learning Algorithms Logistic Regression, K Nearest Neighbour, Random Forest Algorithms yields the following results. The models were evaluated using performance metrics like accuracy, precision, recall, and F1 score after being trained on 80% of the data. The Receiver Operating Characteristic (ROC) curve is employed here for graphical depiction in binary classification for the purpose of evaluating a classification model's performance. Figures 10 to 15 show the results achieved in this experiment. Figure 16 shows the Accuracy comparison of the models.

```
MODEL :  LogisticRegression()
Accuracy of the model is:  90.62382327099284
Confusion Matrix:
 [[  8816   7812]
 [  6381 128364]]
Report:
              precision    recall  f1-score   support

           0       0.58      0.53      0.55     16628
           1       0.94      0.95      0.95    134745

    accuracy                           0.91    151373
   macro avg       0.76      0.74      0.75    151373
weighted avg       0.90      0.91      0.90    151373


==============***===============
```

Logistic Regression

Figure 10. Logistic Regression Results

```
MODEL :  KNeighborsClassifier(n_neighbors=3)
Accuracy of the model is:  98.63383826706216
Confusion Matrix:
 [[ 15530   1098]
 [   970 133775]]
Report:
              precision    recall  f1-score   support

           0       0.94      0.93      0.94     16628
           1       0.99      0.99      0.99    134745

    accuracy                           0.99    151373
   macro avg       0.97      0.96      0.96    151373
weighted avg       0.99      0.99      0.99    151373


==============***===============
```

K Nearest Neighbour

Figure 11. K Nearest Neighbour Results

```
==============***===============
MODEL :  RandomForestClassifier(random_state=42)
Accuracy of the model is:  99.62278609791707
Confusion Matrix:
 [[ 16335    293]
 [   278 134467]]
Report:
              precision    recall  f1-score   support

           0       0.98      0.98      0.98     16628
           1       1.00      1.00      1.00    134745

    accuracy                           1.00    151373
   macro avg       0.99      0.99      0.99    151373
weighted avg       1.00      1.00      1.00    151373


==============***===============
```

Figure 12. Random Forest Classifier Results

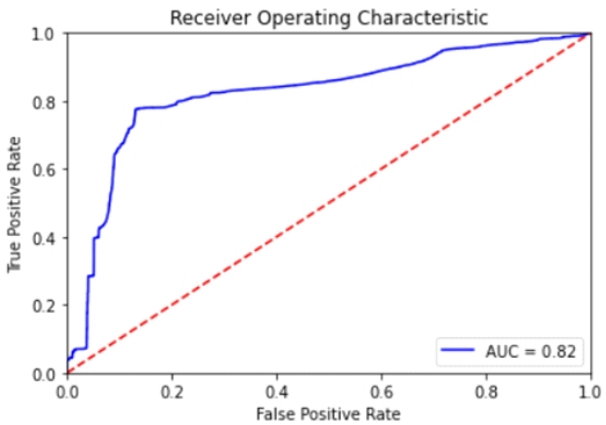


Figure 13. ROC Curve for- Logistic Regression

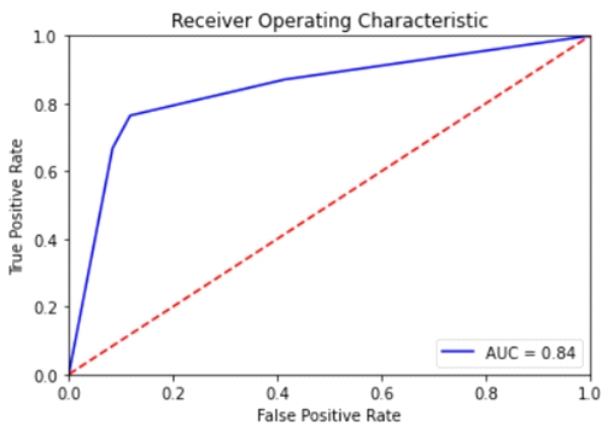


Figure 14. ROC Curve for-K nearest Neighbor

4. Discussion for Experiment # 1

This experiment shows that using machine learning methods to look for DoS attacks in network traffic works well. The outcomes demonstrate that these algorithms can spot patterns that point to bad behavior, showing that

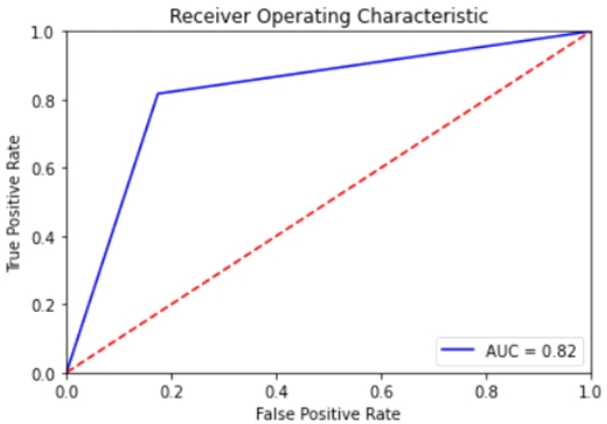


Figure 15. ROC Curve for –Random Forest Classifier

```
plt.plot(classifiers,scores)
plt.title("Accuracy Graph")
plt.figure(figsize=(6, 4))
plt.show()
```

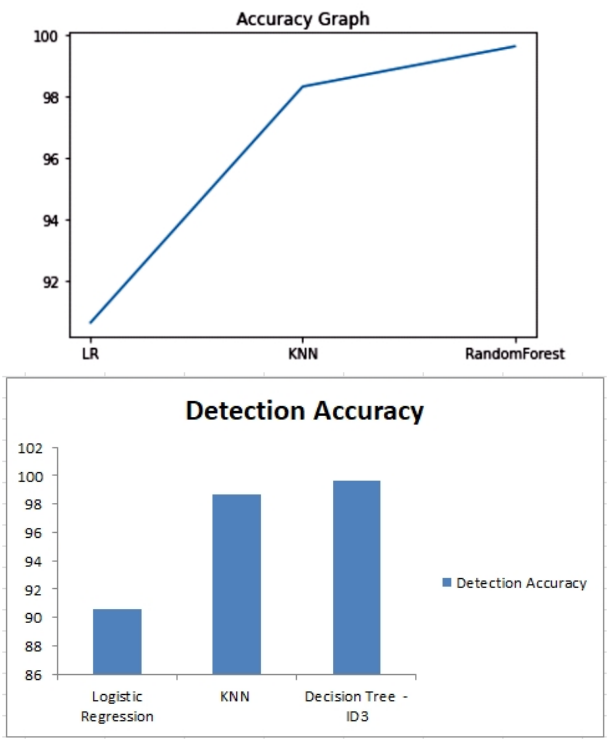


Figure 16. Detection Accuracy comparison Graph

they could be used to make network settings safer & secure. From the Figures 10 to 15 we can observe that Random forest classifier gives the highest performance. Its detection / prediction accuracy detected is 99.65%, while K-Nearest Neighbor is 98.63% and Logistic Regression 90.62%. Thus, the suggested method may detect or predict dos kind malicious network traffic or normal network traffic which helps to secure network servers. Figure 16 depicts comparison graph for the detection accuracy.

5. Experiment #2: (Another Approach) :

A practical approach of launching flood based dos attack with Kali liunix hping3 and detecting it with wireshark:

5.1 Proposed Methodology

In this method we can see how to use Kali Linux (hping3) to initiate a TCP SYN Flood based DoS attack and how to use the Wireshark network protocol analyzer to detect it. An attacker can easily use up all of a target's server resources by sending a lot of SYN packets and not replying (ACK). In this state, the target is having a hard time handling traffic, which will cause its CPU and RAM to be used up faster. Eventually, it will run out of memory space and bandwidth. After this point, the server will not be able to respond to any valid client requests, which will cause a Denial-of-Service attack.

DoS attacks are simple to execute, destructive in nature, and not always transparent. As part of a Denial of Service (DoS) attack, an adversary employs the three-way handshake of the TCP protocol to rapidly halt service and the network via a SYN flood attack. Fortunately, tools like Wireshark make it simple to record it and confirm any signs of a DoS attack. When someone does a SYN flood, they send a lot of SYN packets to the server using fake IP addresses. This makes the server send a reply (SYN-ACK) and leave its ports half-open, waiting for a reply from a host that doesn't exist.

Proposed Methodology Architecture For Experiment #2:

The proposed methodology implementation is explained in Figure 17.

5.1.1 Proposed Methodology Steps

Client Side [Attacker-PC] Working Steps:

Step 1: Login to Kali Linux

Step 2: Open Terminal Prompt in Admin mode in Kali Linux. Here Attacker PC (Kali Linux setup pc) iteratively sends huge no of packets to Target Server (Windows PC) = 10.0.2.15 using below command depicted in Figure 18

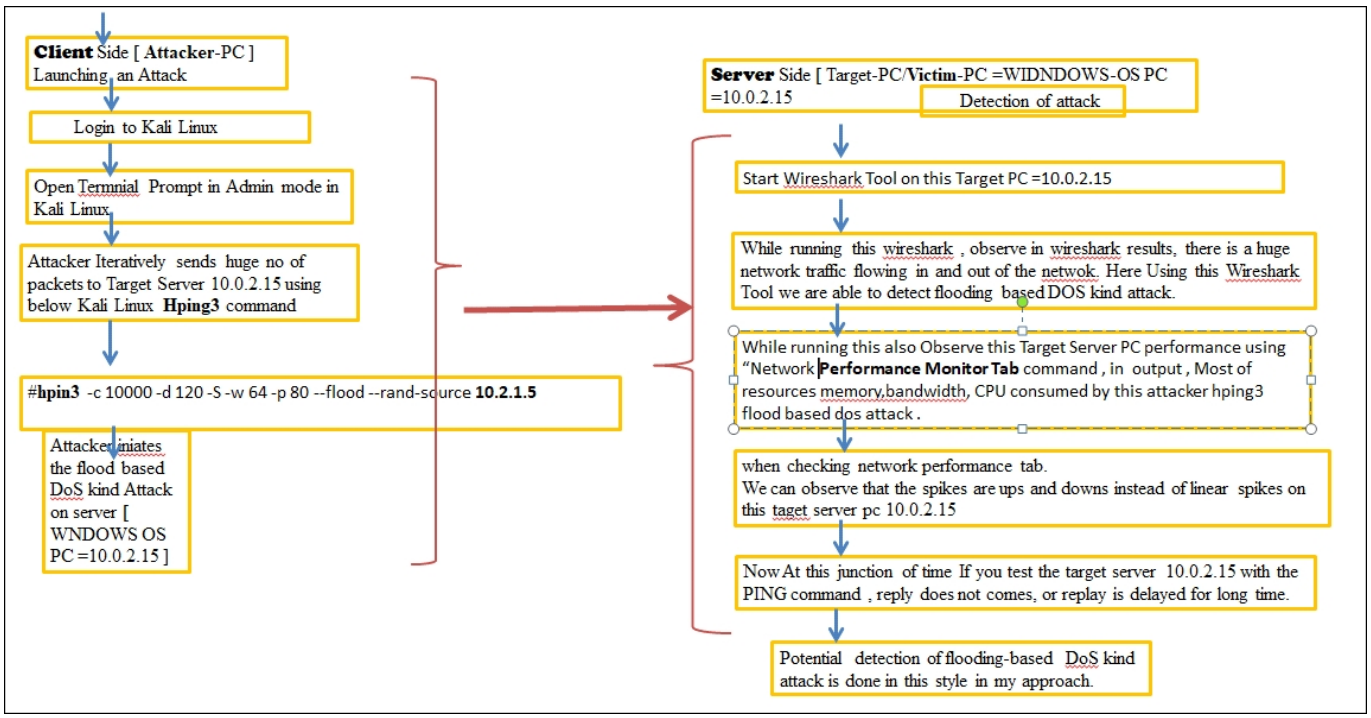


Figure 17. Architecture of Proposed Methodology

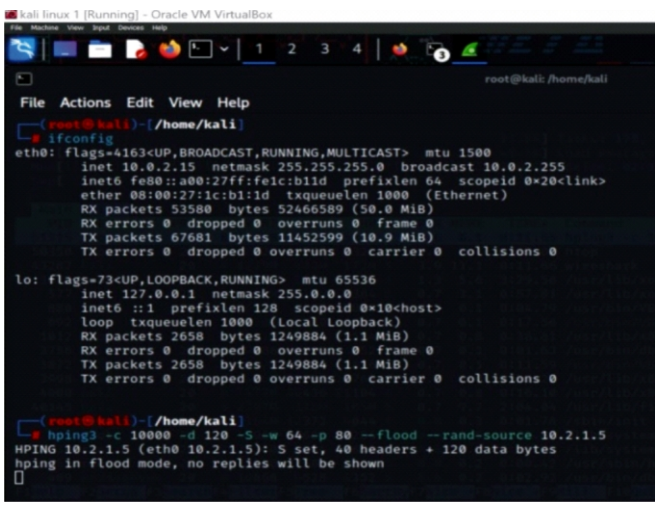


Figure 18. Flood-based DoS Attack Initiation using Kali Linux hping3

#hpin3 -c 10000 -d 120 -S -w 64 -p 80 --flood --rand-source 10.2.1.5

This above instruction command is defined as follows: Every packet has a length of 120 bytes (-d 120) and a weight of 15000 bytes (-c 15000) Enabling the SYN Flag (-S) and configuring the TCP window size to 64 (-w 64). The utilization of the flood parameter in conjunction with the port 80 directive (-p 80) expedites the assault against the HTTP web server belonging to our adversary. The --rand-source option produces illegitimate IP addresses, which is counterintuitive.

Step 3: This Attacker initiating the flood based DoS kind Attack with huge packets sent, is shown in Figure 19.

Server Side [Target-PC/Victim-PC] 10.0.2.15 Working Steps:

Step 1: Start Wireshark Tool on this Target PC [WINDOWS OS PC = 10.2.1.15]

Step 2: Observe the wireshark results as you execute this command; a tremendous amount of network traffic is entering and exiting the network. By employing the Wireshark tool, it is possible to identify massive traffic flooding-based DOS attacks as illustrated in Figure 20.

The significant quantity of SYN packets exhibits minimal temporal variability. SYN packets are characterized by

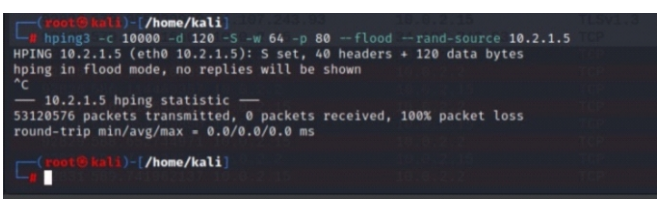


Figure 19. Dos Attack Launched

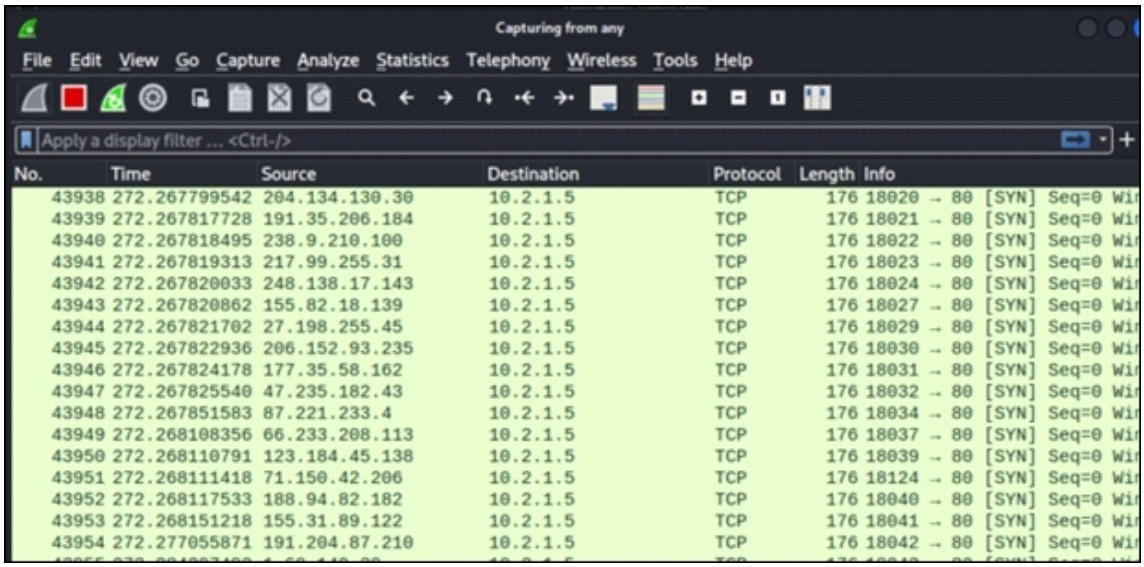


Figure 20. Wireshark Detected Huge Traffic on Server

their distinct originating IP address, port 80 (HTTP) destination, 120-byte length, and 64-byte window size. Cleaned with tcp flags, the quantity of SYN/ACKs is diminished. syn and ack both equal 1. Unquestionably a TCP SYN attack.

Step 3: Using the Network Performance Monitor Tab, we also observe the performance of this target server (a Windows PC). The output indicates that this attacker utilized the majority of the available resources; including memory, bandwidth, and CPU, in a hping3 flood-based DoS attack. The Windows server exhibits oscillating spikes as opposed to linear spikes, which are evident to the naked eye as illustrated in Figure 21.

Step 4: Now during this juncture of time, when the target server 10.0.2.15 is tested using the PING command, it either fails to respond to the command or provides a significantly delayed response (Figure 22).

6. Result & Discussion for Experiment #2

Potential detection of flooding-based DoS kind attack can be done in this way using above approach.

Conclusion

In dispersed contexts like the internet and social media, DOS attack detection has risen. Identifying threats that disrupt network server service is critical. Machine learning models train and evaluate communication traffic to

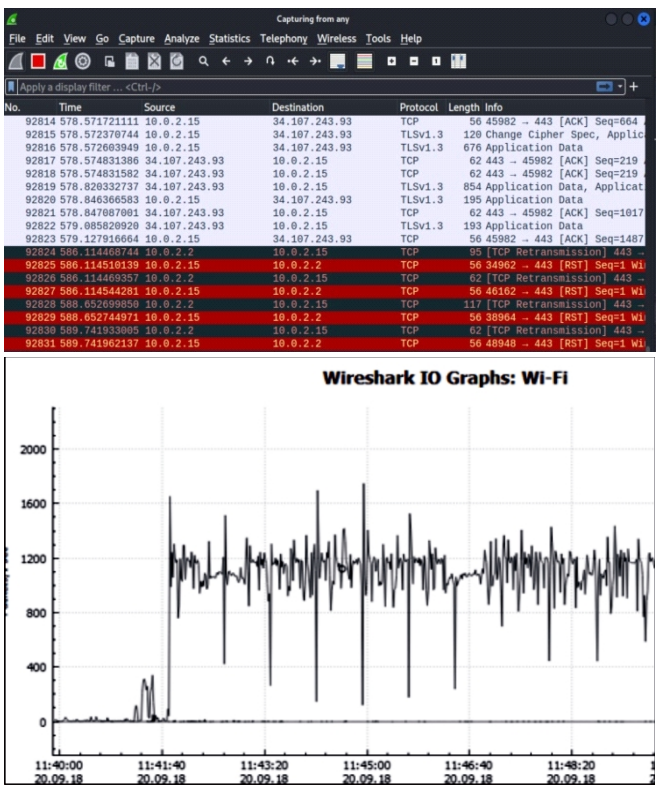


Figure 21. i/o Graphs While Dos Attack Happens

recognize dos attack traffic communications and normal communications. In First experiment our model uses Random forest classifier which predict model accuracy 99.63%. To secure computer networks, the suggested method can recognize dos attack type traffic or regular

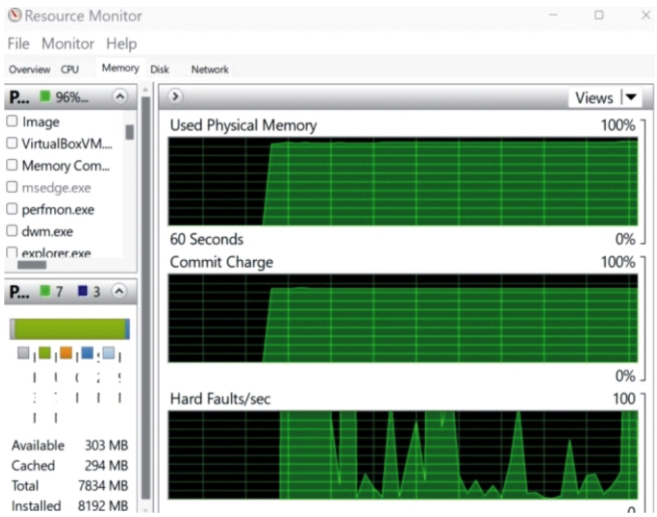


Figure 22. More Utilization of Memory, CPU Time by Attack

type traffic communication in network. Furthermore, in the second experiment, we investigated the utilization of Kali Linux (hping3) to execute a TCP SYN Flood DoS attack, as well as the application of the Wireshark network protocol analyzer to identify doS-type abnormal traffic.

References

[1]. Aslan, Ö. (2022). A methodology to detect distributed denial of service attacks. *Bilişim Teknolojileri Dergisi,* 15(2), 149-158. https://doi.org/10.17671/gazibtd.1002178

[2]. Barati, M., Abdullah, A., Udzir, N. I., Mahmod, R., & Mustapha, N. (2014, August). Distributed denial of service detection using hybrid machine learning technique. In 2014 *International Symposium on Biometrics and Security Technologies (ISBAST)* (pp. 268-273). IEEE. https://doi.org/10.1109/ISBAST.2014.7013133

[3]. Bhuyan, M. H., Kashyap, H. J., Bhattacharyya, D. K., & Kalita, J. K. (2014). Detecting distributed denial of service attacks: methods, tools and future directions. *The Computer Journal,* 57(4), 537-556. https://doi.org/10.1093/comjnl/bxt031

[4]. Cetinkaya, A., Ishii, H., & Hayakawa, T. (2019). An overview on denial-of-service attacks in control systems: Attack models and security analyses. *Entropy,* 21(2), 210. https://doi.org/10.3390/e21020210

[5]. Kim, J., Shim, M., Hong, S., Shin, Y., & Choi, E. (2020). Intelligent detection of IoT botnets using machine learning and deep learning. *Applied Sciences,* 10(19), 7009. https://doi.org/10.3390/app10197009

[6]. Kozlowski, M., & Ksiezopolski, B. (2021, July). A new method of testing machine learning models of detection for targeted DDoS attacks. In *Proceedings of the 18th International Conference on Security and Cryptography (SECRYPT)* (pp. 728-733). https://doi.org/10.5220/0010574507280733

[7]. Lima Filho, F. S. D., Silveira, F. A., de Medeiros Brito Junior, A., Vargas-Solar, G., & Silveira, L. F. (2019). Smart detection: an online approach for DoS/DDoS attack detection using machine learning. *Security and Communication Networks,* 2019, 1-15. https://doi.org/10.1155/2019/1574749

[8]. Loukas, G., & Öke, G. (2010). Protection against denial of service attacks: A survey. *The Computer Journal,* 53(7), 1020-1037. https://doi.org/10.1093/comjnl/bxp078

[9]. Masdari, M., & Jalali, M. (2016). A survey and taxonomy of DoS attacks in cloud computing. *Security and Communication Networks,* 9(16), 3724-3751. https://doi.org/10.1002/sec.1539

[10]. Nandi, S., Phadikar, S., & Majumder, K. (2020, February). Detection of DDoS attack and classification using a hybrid approach. In 2020 *Third ISEA Conference on Security and Privacy (ISEA-ISAP)* (pp. 41-47). IEEE. https://doi.org/10.1109/ISEA-ISAP49340.2020.234999

[11]. Perez-Diaz, J. A., Valdovinos, I. A., Choo, K. K. R., & Zhu, D. (2020). A flexible SDN-based architecture for identifying and mitigating low-rate DDoS attacks using machine learning. *IEEE Access,* 8, 155859-155872. https://doi.org/10.1109/ACCESS.2020.3019330

[12]. Rahman, O., Quraishi, M. A. G., & Lung, C. H. (2019, July). DDoS attacks detection and mitigation in SDN using machine learning. In 2019 *IEEE World Congress on Services (SERVICES)* 2642, 184-189. IEEE. https://doi.org/10.1109/ SERVICES.2019.00051

[13]. Rao, G. S. ., & Subbarao, P. K. (2023). A novel approach for detection of dos / ddos attack in network environment using ensemble machine learning model. *International Journal on Recent and Innovation Trends in*

Computing and Communication, 11(9), 244–253. https://doi.org/10.17762/ijritcc.v11i9.8340

[14]. Shah, H., Shah, D., Jadav, N. K., Gupta, R., Tanwar, S., Alfarraj, O., & Marina, V. (2023). Deep learning-based malicious smart contract and intrusion detection system for IoT environment. *Mathematics,* 11(2), 418. https://doi.org/10.3390/math11020418

[15]. Shamsolmoali, P., & Zareapoor, M. (2014, September). Statistical-based filtering system against DDOS attacks in cloud computing. In 2014 *International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1234-1239). IEEE. https://doi.org/10.1109/ICACCI.2014.696 8282

[16]. Smys, S., Basar, A., & Wang, H. (2020). Hybrid intrusion detection system for internet of things (IoT). *Journal of ISMAC,* 2(4), 190-199. https://doi.org/10.36548/jismac. 2020.4.002

[17]. Tayyab, M., Belaton, B., & Anbar, M. (2020). ICMPv6-based DoS and DDoS attacks detection using machine learning techniques, open challenges, and blockchain applicability: A review. *IEEE Access,* 8, 170529-170547. https://doi.org/10.1109/ACCESS.2020.3022963

[18]. Ulemale, T. (2022). Review on detection of DDOS attack using machine learning. *International Journal for Research in Applied Science and Engineering Technology (IJRASET),* 10(3). https://doi.org/10.22214/ijraset.2022.40742

[19]. Wang, Y., Hu, T., Tang, G., Xie, J., & Lu, J. (2019). SGS: Safe-guard scheme for protecting control plane against DDoS attacks in software-defined networking. *IEEE Access,* 7, 34699-34710. https://doi.org/10.1109/ACCESS.2019.2895092

[20]. Wankhede, S., & Kshirsagar, D. (2018, August). DoS attack detection using machine learning and neural network. In 2018 *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)* (pp. 1-5). IEEE. https://doi.org/10.1109/ICCUBEA.2018.8697702

[21]. Wu, C., Sheng, S., & Dong, X. (2018, October). Research on visualization systems for DDoS attack detection. In 2018 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 2986-2991). IEEE. https://doi.org/10.1109/SMC.2018.00507

[22]. Yuan, X., Li, C., & Li, X. (2017, May). DeepDefense: Identifying DDoS attack via deep learning. In 2017 *IEEE International Conference on Smart Computing (SMARTCOMP)* (pp. 1-8). IEEE. https://doi.org/10.1109/SMARTCOMP.2017.7946998

[23]. Yusof, A. R. A., Udzir, N. I., Selamat, A., Hamdan, H., & Abdullah, M. T. (2017, November). Adaptive feature selection for denial of services (DoS) attack. In 2017 *IEEE Conference on Application, Information and Network Security (AINS)* (pp. 81-84). IEEE. https://doi.org/10.1109/AINS.2017.8270429

[24]. Zekri, M., El Kafhali, S., Aboutabit, N., & Saadi, Y. (2017, October). DDoS attack detection using machine learning techniques in cloud computing environments. In 2017 *3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)* (pp. 1-7). IEEE. https://doi.org/10.1109/CloudTech.2017.8284731

ABOUT THE AUTHORS

Gottapu Sankara Rao, Research Scholar of JNTUK University, JNTUK, India. His research areas include Computer Networks, Network Security, Machine Learning, and Deep Learning.

Dr. P. Krishna Subbarao is Professor in the Department of CSE, Head of the Department, Dean of Student Affairs, GVPCE (A), Visakhapatnam. He has about 30+ research papers in various International Journals and Conferences, and attended many national and international conferences in India and abroad. He is a member of IEEE and he is an active member of the board of reviewers in various International Journals and Conferences. His research interests include Data Mining, Bio Informatics. Network Security and Internet of Things.

AGRICULTURE MANAGEMENT SYSTEM

By

TAKONDWA KAIYA * CHIPATSO MEDI ** FANNY CHATOLA ***

*-*** Department of Computer Science, DMI St John the Baptist University, Lilongwe, Malawi.

Date Received: 04/12/2023 Date Revised: 19/12/2023 Date Accepted: 17/01/2024

ABSTRACT

The Agriculture Management System is a web-based platform that provides farmers with a full range of agricultural information. Its goal is to increase farmers' productivity and profitability by using a centralized system for managing their agricultural operations. The system gives farmers access to critical data such as weather forecasts, soil and crop statistics, allowing them to make informed decisions and improve their farming operations. In terms of weather management, the technology provides precise forecasts, which help farmers plan their operations more efficiently. Farmers can use this data to identify the best time to plant, harvest, and perform other agricultural duties. Regarding market prices, the AMS gives farmers with real-time crop and livestock prices, allowing them to make informed pricing and marketing decisions. Furthermore, the technology provides farmers with critical soil and crop data, giving them insights into soil quality and crop health. This data helps farmers improve their agricultural practices by allowing adjustments to planting schedules, fertilizing, and irrigation systems to increase crop yields and quality. Recognizing the value of cooperation and information sharing, the AMS includes an AI-powered chatbot. This feature allows farmers to share information, seek assistance, and ask questions about the agricultural management system, including crops, soil, market trends, and weather management. The incorporation of this interactive technology encourages a community-driven approach to agricultural management, creating a conducive atmosphere for farmers to flourish.

Keywords: Firebase, Web Service, Weather Data Intergration, Crop Planning, Pest and Disease Management.

INTRODUCTION

The agricultural management system is an important factor that interacts with numerous production aspects, and improving the productivity of factors such as land, labor, capital, and managerial ability which is dependent on relevant, reliable, and meaningful data. Extension, research, education, and agricultural organizations provide information that helps farmers make better decisions, highlighting the need of understanding and improving these systems (Agricultural Information Management Standards, n.d).

Thomas et al. (2020) observes that farmers' access to knowledge leads to empowerment through control over resources and decision-making processes. Effective information distribution systems are crucial for making better decisions about agricultural production, processing, trade, and marketing. The Food and Agriculture Organization highlights the need of information for rural development, as well as the necessity to invest in human resources to promote knowledge and information sharing.

Research institutes, colleges, businesses, and farmers all help to develop new agricultural technologies. However, the delayed adoption of these technologies is linked to a lack of linkages between research, extension, and farmers, emphasizing the importance of integrated information systems encompassing farmers, educators, researchers, and extension agencies.

Despite the importance of information systems,

integration between people and institutions, particularly in the research-extension-farmer interaction, is frequently ineffective. The disparities in information demands across market-oriented, transitional, and subsistence farming present issues. Limited research on agricultural information systems emphasizes the necessity for extensive information, including procedures, interactions, and an examination of farmers' information needs and organizational structures.

1. Objectives

The objective of the current research is to give farmers access to current information and tools, get the most recent insights from agricultural experts, use data analysis to help farmers make more educated decisions, optimize agricultural activities to increase productivity, encourage collaboration and information sharing throughout the agricultural business and provide a user-friendly platform for streamlining agricultural operations.

2. Literature Review

Benos et al. (2021) conducted a literature analysis on the revolutionary potential of agriculture management systems, stressing the various benefits provided by machine learning. Their investigation highlighted the system's critical role in minimizing common agricultural issues, such as the effects of shifting weather patterns, poor soil quality, and market price volatility. The authors proposed for the implementation of a comprehensive and user-friendly platform for managing agricultural activities. They stated that such a platform has the potential to dramatically increase farmer production and profitability, while also serving as a strategic tool in the modernization of agricultural methods.

Jackson and Daughetry (1905) work focused on the practical implementation of the decision tree algorithm in agricultural management systems. The study proved how this algorithm could effectively handle field data, providing farmers with predicted insights that are critical for decision making. By focusing on ideal planting and harvesting periods based on real-time weather and soil conditions, Mugaviri demonstrated the potential of such systems to increase agricultural yields. The advice for expanded adoption in smallholder farming communities emphasized the decision tree algorithm's scalability and flexibility to a variety of agricultural situations.

Kropotkin (1965) study adds a regional viewpoint by creating an agriculture management system specifically designed for apple cultivation in China. The use of a machine learning algorithm helped determine the best time for pesticide administration, taking into account local weather patterns and pest populations. Zhang's research showed a significant reduction in pesticide use, which addressed environmental concerns while maintaining good agricultural yields. This study emphasized the significance of context-specific agriculture management solutions, demonstrating the ability of machine learning algorithms to optimize farming methods at the regional level.

These studies highlighted the various applications and benefits of agricultural management systems. From the broad-scale improvements proposed by Johnson and Smith to the specific, field-level insights provided the literature review demonstrated the changing landscape of agricultural technology and the potential for innovative solutions to critical challenges faced by farmers worldwide.

3. Problem Definition

Farmers frequently face difficulties in obtaining current knowledge about plants, treatments, and agricultural methods. Their reliance on obsolete literature, word-of-mouth counsel, or restricted online resources frequently results in inefficiencies and stymies decision-making. For example, a farmer may struggle to detect a newly emerging plant disease or be unfamiliar with the best efficient insecticide for a specific crop. Recognizing this common issue, the work aims to create a centralized platform that would act as a one-stop resource for farmers, providing them with easily available and credible information to help them close knowledge gaps.

Furthermore, farmers frequently face delays in receiving appropriate assistance and clarifications on plant care, medication use, or specific agricultural techniques. This lack of timely support can have real consequences; for

example, a farmer faced with an unexpected pest infestation may suffer crop losses as a result of delayed action. The work addresses these difficulties by incorporating interactive features like a blog where farmers can seek advice from agricultural professionals. This real-time support system attempts to close the gap, allowing farmers to handle challenges and make educated decisions faster. For example, a farmer dealing with an unexpected crop disease might publish photographs and facts on the blog to obtain fast advise from specialists and other farmers, preventing crop harm.

In essence, the initiative aims to promote dynamic exchanges that offer farmers with current knowledge and expert perspectives, resulting in more efficient and informed agricultural operations.

4. Existing System

The current agricultural research and extension system is often housed within a ministry of agriculture, with different organizations or departments responsible for various responsibilities. These entities, however, may differ in terms of organizational structure and operating practices. Universities and national research organizations concentrate on research, whereas the agriculture department focuses on extension efforts.

In this traditional structure, the primary emphasis is on breeding, testing, and distribution. The process of creating and distributing technology takes a top-down approach, with researchers developing superior varieties and passing them on to extension for demonstrations and farmer adoption. However, this walled strategy leads to separate program development for each function, resulting in program duplication. This not only wastes money, but also perplexes producers about which organization to contact.

This research and extension system has a tiered structure that stretches from national to field level. While internal communications inside an organization flow freely from upper to lower levels, exchanges between organizations are typically convoluted, rendering them ineffective. Lower-level coordination is generally based on explicit directions from higher levels. The fundamental nature of this system creates obstacles for efficient resource usage and collaborative decision-making.

5. Findings and Recommendations

Agriculture is very important in Malawi, accounting for around 29% of the country's GDP and employing nearly 80% of the people. Despite agricultural grounds accounting for over 47% of total land area, food insecurity persists, affecting approximately 6.7 million people in rural regions. The contemporary environment is defined by a high-input, high-productivity export sector dominated by a small number of large-scale farmers who use around 60% of productive land. In contrast, the vast majority of smallholder farmers plant low-yielding food crops, which present various issues such as dependency on rainfall and vulnerability to weather shocks.

Large-scale estate farmers, who occupy a sizable percentage of rich land, contribute to soil losses of 20 tons per hectare per year, resulting in yield losses ranging from 4% to 25%. To overcome these difficulties and strengthen Malawi's agriculture management system, long-term development plans should be implemented. Farmers should be trained to implement new techniques and technologies that will increase productivity, improve nutrition, and lower child mortality. Promoting afforestation measures, such as tree planting, can also help to ensure sustainable farming practices. Collaboration with the Malawian government is critical for implementing and enabling agricultural policies that promote sustainable agriculture and solve the country's farmers' varied challenges.

6. Proposed System

The envisioned Agricultural Management System is intended to be a user-friendly web-based platform that meets the different needs of agriculture enthusiasts. It strives to meet the specific needs of persons interested in agriculture, rather than simply serving as an information repository. Farmers will be able to easily search and access a wide range of information about crops, insecticides, pesticides, and more.

This Farming Management System is entirely online, simplifying the way farmers interact with agricultural

information. The system's user-centric design allows farmers to simply search for and retrieve crucial information, resulting in a more informed and efficient approach to agricultural techniques. The incorporation of articles and blogs expands their knowledge base by providing vital insights and best practices in agriculture.

6.1 Advantages

- Universal Accessibility: Any user, regardless of location, can access data from the Information System. This inclusion ensures that agricultural knowledge is accessible to a worldwide audience.
- Controlled Data Input: Authenticated individuals from prominent institutes are given authority to contribute information to the system via the internet. This limited access ensures that the data is reliable and authentic.
- Data integrity and management: The data administrator is exclusively responsible for eliminating unneeded information and altering the database. This safeguard protects data integrity and prevents illegal adjustments, ensuring that the information is accurate.

7. System Objective

The fundamental goals of this Agricultural Management System are strategically aligned with tackling critical difficulties in the agriculture sector.

- *Improving Production and Profitability:* By providing farmers with a consolidated platform for information and tools, the system hopes to increase production and profitability. For example, providing information on appropriate planting times or advising effective pest control strategies can directly lead to higher yields.
- *Real-time Information Access:* Farmers will have access to the most recent information from agriculture specialists, allowing them to stay up to date on improvements in agricultural methods, new technologies, and market trends.
- *Informed Decision-Making:* The system's goal is to help farmers make more informed decisions by analyzing data. For example, data on weather patterns and soil conditions can help in crop planning and harvesting schedules.
- *Optimizing Agricultural Operations:* By streamlining information on crops, insecticides, and other factors, the system aims to improve agricultural operations. For example, selecting various crops depending on soil conditions can help to promote more sustainable, efficient farming techniques.
- *Facilitating Collaboration:* The platform intends to promote collaboration and information sharing in the agriculture industry. Farmers share views, experiences, and best practices, forming a supportive community.
- *User-Friendly Platform:* Creating a user-friendly platform is critical to the system's success. Farmers can easily manage their agricultural activities because of an easy interface, which contributes to the platform's improved adoption and efficacy.
- In short, the proposed Agricultural Management System is more than just a repository of information; it is a dynamic platform designed to provide farmers with the tools and knowledge they need to operate sustainable and efficient agriculture.

8. System Specification

The requirements and specifications for the development of the Agriculture Management System is outlined below:

9. Software Requirements

- WampServer:
- *Apache:* Open – source Java servlet container developed by the Apache Software Foundation.
- *MySQL Server:* It handles large databases much faster than existing solutions.
- It consists of multi-threaded SQL server that supports different back ends, several different client programs and libraries, administrative tools, and application programming interfaces (APIs)
- Its connectivity, speed, and security make MySQL Server highly suited for accessing databases on the Internet.
- PHP, HTML, Bootstrap CSS: PHP for backend

development, HTML for frontend and Bootstrap CSS for styling the contents.

10. Hardware Requirements

- *Processor:* Computer with at least 2GB memory, processor intel core i3 with a minimum speed of 1GHz
- Main Memory: 1 GB
- Ram: 1.00GB
- Hard Disk: 240 GB
- Monitor: CRT Monitor 15 inch
- Keyboard: Multimedia Keyboard
- Mouse: Optical mouse

11. System Architecture

The architectural design overview of the AMS is detailed below, elucidating the system's structure, components, and interactions crucial for its successful implementation and functionality. In this section, a Use Case Diagram (Figure 1) and a Flowchart will be provided for further clarity and insight.

12. Use Case Diagram

In this Use Case Diagram, we have three main actors: User

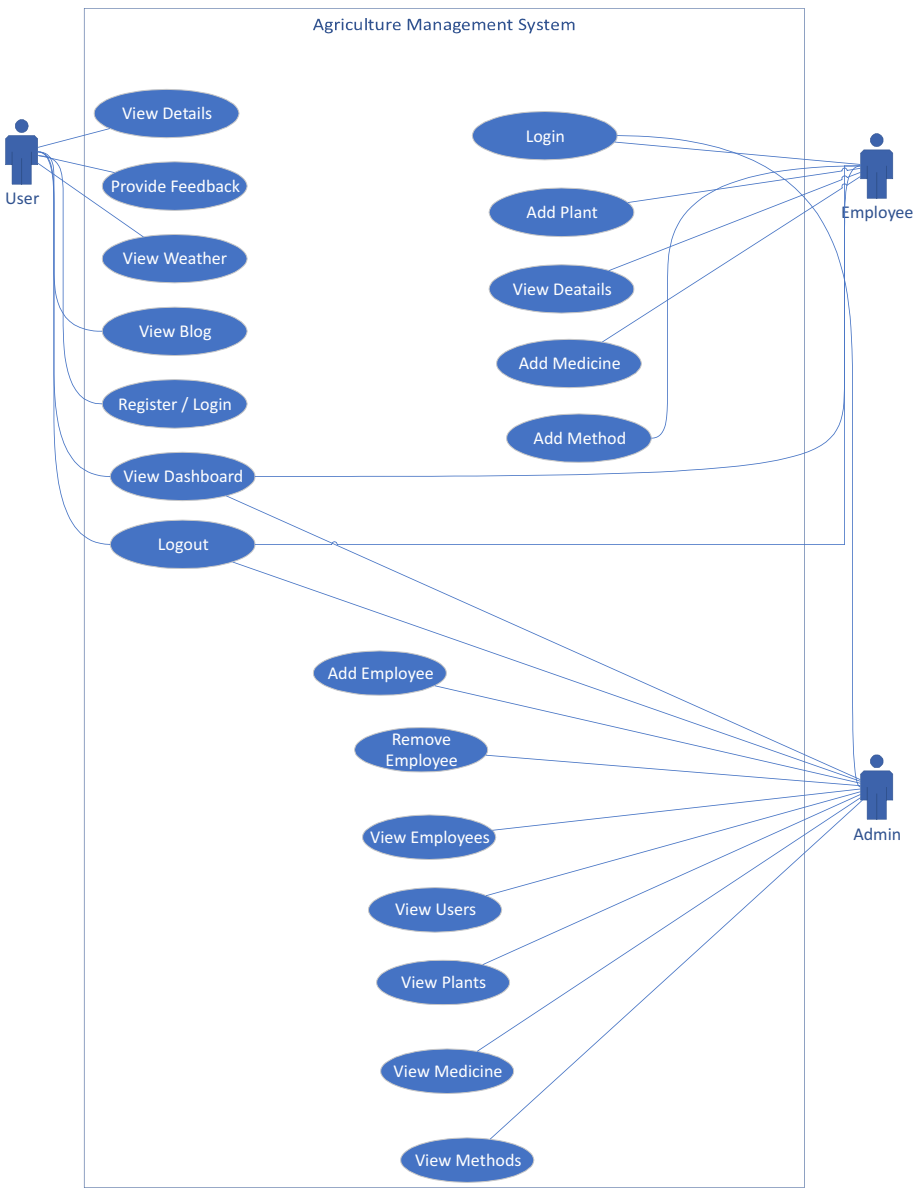


Figure 1. Use Case Diagram

(registered and unregistered), Employee, and Admin.

- The User (Registered) actor is allowed to view the details of plants, medicines, and methods. They can also provide feedback, view weather updates, and read blog posts. They are also able to register and log in to their account to view their dashboard. The User (Unregistered) actor is allowed to view the details of plants, medicines, and methods. They can also provide feedback, view weather updates, and read blog posts. However, they are not able to register or log in to their account.
- The Employee actor is allowed to add details about plants, medicines, and methods. They are also able to view those details.
- The admin actor has access to all the functionalities of the system. They can view the list of all users and employees, view the details of plants, medicines, and methods, add a new employee, remove an employee, and view user feedback.

13. Data Flow Diagram

The flow diagram (Figure 2) shows the following: when one has opened the system it will display the home page. Then if one is already registered in the system, he will just login using the credentials used when creating account. New users will create account as a user and then they will login. After login they will be redirected to their dashboard and perform the tasks they want. After completing the task they will logout and the system will power off.

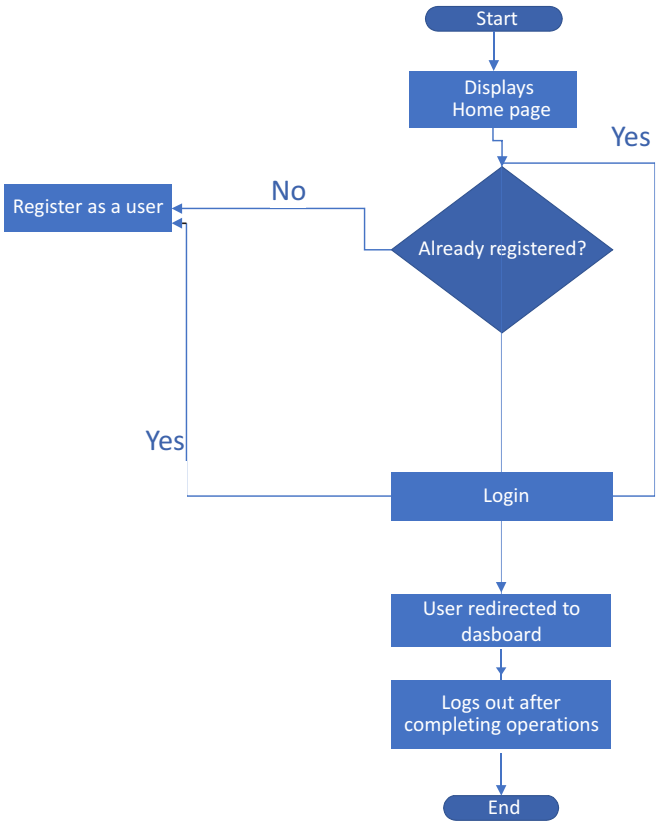


Figure 2. Data Flow Diagram

14. System Development

14.1 Methodology

The Agriculture Management System is developed using a web-based technology of software engineering methodologies, such as the Agile methodology. The Agile methodology is important in software development because it emphasizes flexibility, collaboration, and incremental progress. Agile enables teams to respond swiftly to changing needs, encouraging ongoing communication with stakeholders throughout the development process. Its iterative structure provides timely and consistent deliveries, allowing for a shorter time-to-market and lowering risks by correcting issues as soon as they arise. The methodology encourages increased quality through regular testing and continuous integration, resulting in higher overall customer satisfaction. Agile's adaptable planning and focus on client needs help to ensure project success, while its emphasis on collaboration and continuous improvement fosters a dynamic and responsive development environment.

14.2 Algorithm

The Agriculture Management System incorporates various algorithms to enhance functionality:

- Recommendation algorithm has been used in this work to provide personalized recommendations to users based on their preferences, past interactions and data collected from the system.
- Weather Forecasting Algorithm has also been used to provide accurate weather information to farmers. It uses a historical data, current conditions and predictive models to forecast future weather patterns.

- Decision tree algorithm has been used for crop disease and pest identification.

14.3 Module Description

The Agriculture Management System is made up of three key modules: User, Employee, and Admin, each of which serves a specific purpose within the system.

- *User Module:* allows users to create accounts and access particular information about plants, medicines, and agricultural methods. Users join by entering personal information such as their first and last names, email addresses, phone numbers, and passwords. After registering, users are taken to their individual Dashboard, which provides information about plants, medicines, and agricultural methods. Keywords: Agriculture Management, User Module, Dashboard, Plant Details, Medical Information.
- *Employee Module:* is designed for agricultural staff to log in with their credentials. The admin manages account registration. Employees that log in have access to their Dashboard, where they can contribute by uploading new plant facts, medical information, and agricultural practices.
- *Admin Module:* Serves as the core control panel, allowing administrators to log in via the dedicated admin tab on the home page. Once logged in, administrators are taken to their Dashboard, which provides access to a variety of capabilities. Administrators can keep track of registered users and workers, as well as plant lists, medical information, and agricultural practices. Table 1 presents the module description.

14.4 Test Plan

This test plan outlines the testing approach for the system. The testing focuses on verifying the functionality and user experience of each module, including registration, login, data entry, data retrieval and system administration features.

15. System Implementation

Figures 3 to 11 show the agriculture management system implementation about the home page, about us, contact page, weather page, blog page and user, employee and admin modules which are well explained in detail below.

15.1 Landing Page

This is the landing page which serves as a welcome part where it directs to get started into the agriculture management system.

15.2 Login Page

This is the login page for the users who already have an account.

15.3 Registration Page

This is where new users can create their accounts. This page allows employees and admins to login into their accounts by toggling the tabs accordingly.

15.4 Weather Page

This page allows users to get some more information about weather updates based on their location. It can also be accessed by unregistered users.

15.5 About Us

Allows users to get some more information about the quality and the services of the agriculture management system. It can also be accessed by unregistered users.

15.6 Contact Page

Allows users to provide feedback or queries about the services of the agriculture. It can be accessed by users

Test Id	Item	Description	Expected Outcomes	User Type Responsible
1	Registration	Verify user registration functionality.	User should be able to successfully register an account with valid information	User
2	User login	Verify user login functionality	User should be able to login using their registered credentials and access their dashboard	User
3	User dashboard	Verify user dashboard functionality	User should be able to view the plant details, medicine details and method details on their dashboard	User
4	Admin dashboard	Verify admin dashboard functionality	Admin should be able to view user list, employee list and user feedback	Admin

Table 1. Module Description

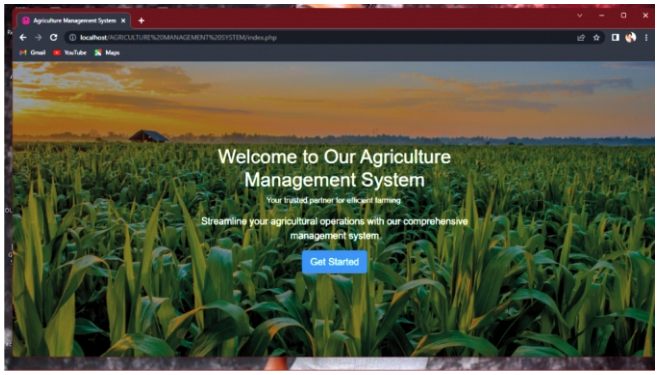


Figure 3. Landing Page

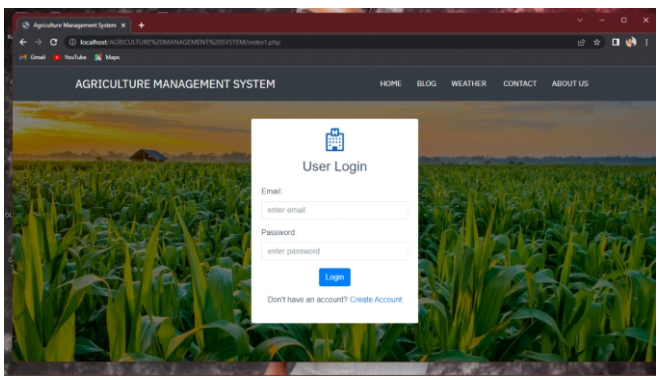


Figure 4. Login Page

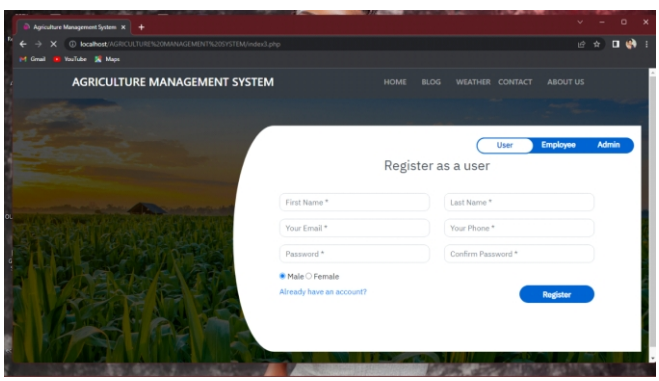


Figure 5. Registration Page

who have not created accounts.

15.7 Blog Page

Allows users to get some more information on different agriculture related content. It can be accessed by users who have not created accounts.

15.8 Employee Dashboard

This the dashboard for the employees after login into the system.

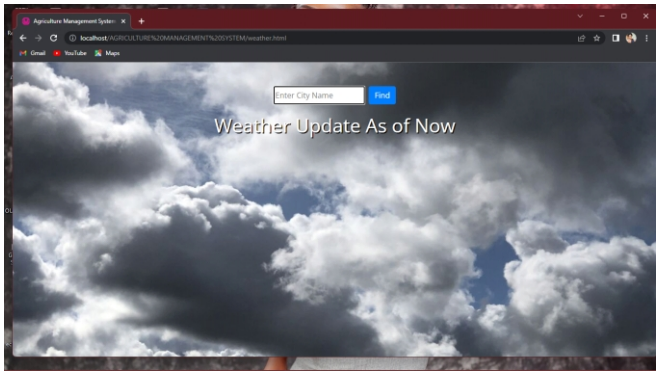


Figure 6. Weather Page

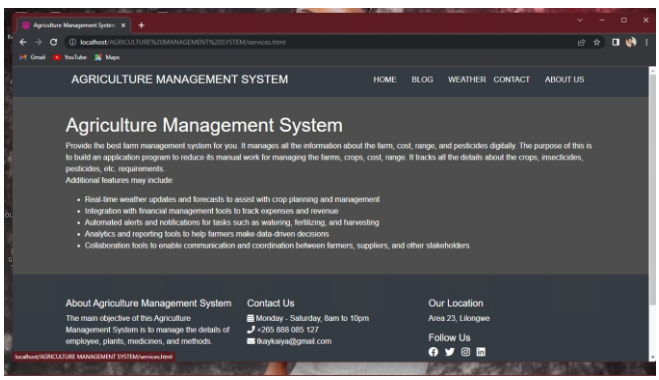


Figure 7. About Us Page

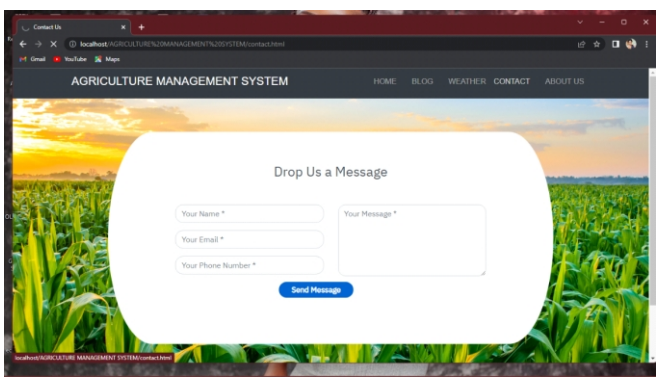


Figure 8. Contact Us Page

15.9 Admin Module

This the dashboard for the admin. This allows the admin to view the following; employee list, user list, plant list, medicines list, Methods list and manage employees.

16. Discussion

The planned Agriculture Management System (AMS) proposes a significant transformation on how farmers access and use critical information to improve their

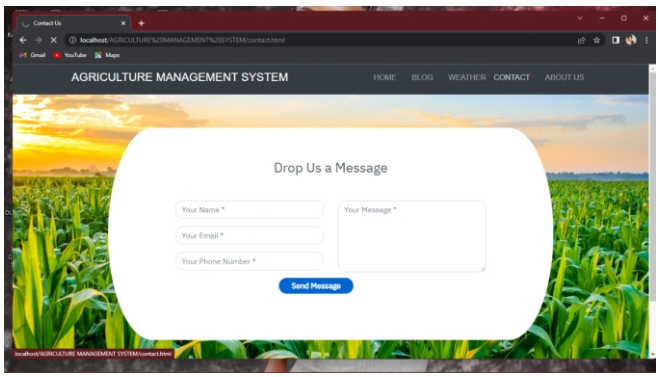


Figure 9. Blog Page

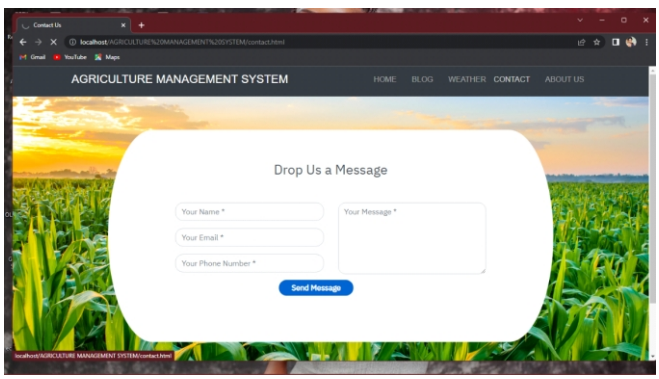


Figure 10. Employee Dashboard

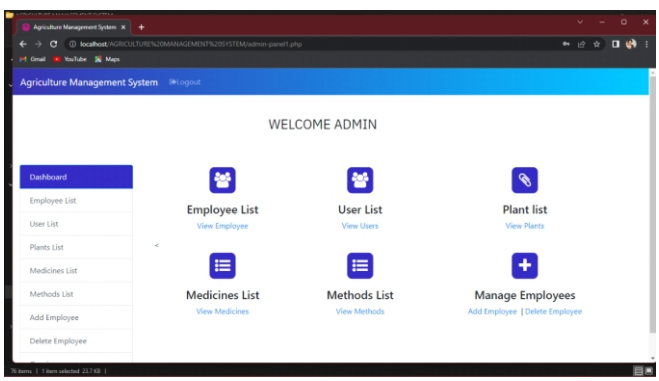


Figure 11. Admin Module

agricultural practices. It describes the system's key goals and features, highlighting its ability to boost production, profitability, and overall efficiency in farming operations. The system's primary features include real-time weather forecasts, market prices, and critical soil and crop data, which aid in informed decision-making. Furthermore, the use of an AI-powered chatbot promotes community-driven knowledge sharing among farmers. Some literature studies emphasize the relevance of agricultural management systems. These works demonstrate the potential of machine learning algorithms in addressing common agricultural difficulties. The scalability and adaptability demonstrated in these tests provide important insights into the proposed system's architecture.

The defined problem underscores the frequent challenges farmers confront in obtaining timely and reliable information. The interactive components, such as the blog, are presented as solutions to bridge the gap and provide farmers with real-time support. This strategy is consistent with the system's overall purpose of building a dynamic exchange platform. The system design, development methods, and algorithmic components are all important features of the proposed system's effectiveness. The Agile technique promotes adaptation and collaboration throughout the development process. The hardware and software requirements serve as a practical foundation for deploying the AMS while maintaining compatibility and efficiency.

The use case diagram and flow diagram visually depict the system's interactions, providing insight into user roles and operational flow. These visual tools improve comprehension of how users, staff, and administrators interact with the system.

The proposed AMS has the potential to change agricultural information systems. By tackling current issues, adding innovative technology, and emphasizing user-centric design, the system becomes a comprehensive tool for farmers. Future work may include rigorous testing, including user feedback, and constant refinement to assure the system's effectiveness and widespread use in a variety of agricultural situations.

17. Result Comparison

When measuring the effectiveness of the Agriculture Management System (AMS), a result comparison is critical to determining the system's impact on farmers, agricultural practices, and overall productivity. The comparison entails comparing key performance indicators and outcomes to preset objectives outlined in the system's goals and specifications.

- *Provide Farmers with Current Information and Tools:* An evaluation of the accessibility and relevance of information and tools offered through the AMS is critical. When comparing the system's information timeliness and accuracy to past systems, it becomes clear how effective it is in keeping farmers informed.
- *Receive the Most Recent Insights From Agricultural Experts:* User comments and interaction with expert advice channels, such as the blog and interactive features, indicate the system's effectiveness in fostering communication between farmers and agricultural specialists. Comparing the level of expert contact before and after implementation yields significant insights.
- *Optimize Agricultural Operations for Increased Productivity:* The efficiency of agricultural operations, such as planting schedules, pest management tactics, and resource use, is used as a performance metric. The AMS's role in optimizing these procedures can be measured using key metrics.
- *Facilitate Collaboration and Information Sharing:* Assessing the extent of collaboration and information exchange within the agricultural sector is critical. Comparative statistics on information sharing, community participation, and collaborative activities can help assess the system's success in promoting a cooperative atmosphere.
- Create a user-friendly platform to streamline agricultural operations. The evaluation of the AMS's usability is based on user satisfaction and simplicity of use. Comparing user experiences, feedback, and system interactions before and after implementation yields information about the system's usability.

Conclusion

Finally, the precisely designed system design described in this work provides a strong and comprehensive answer for resolving the multiple issues involved with agricultural management. The successful completion of this work is a big step toward creating a more efficient, informed, and collaborative environment in the agricultural sector.

The system's architecture, which includes modules for handling agricultural information, user accounts, and administrative capabilities, provides a solid platform for complete agricultural administration. By taking a user-centric approach, the system ensures that farmers, agricultural personnel, and administrators can engage with the platform in a fluid manner, improving their overall experience.

This work's primary achievement is the promotion of effective communication and information sharing in the agriculture industry. The user module gives everyone, particularly farmers, access to precise and crucial information about plants, medications, and agricultural processes. This information democratization not only fills knowledge gaps, but also enables farmers to make better decisions, resulting in higher productivity and profitability. The employee module encourages active participation among agricultural staff, easing the process of adding new information about plants, medications, and practices. This collaborative feature ensures that the system's database is dynamic and reflects the changing environment of agricultural techniques. The collaboration of users and employees within the system results in a knowledge-sharing ecosystem that benefits all stakeholders.

At the system's core, the admin module serves as the nerve center, giving administrators with the tools they need to oversee and govern the entire platform. The ability to monitor user and staff lists, as well as plant, drug, and method specifics, allows managers to make educated judgments about system optimization and improvement.

Moving forward, it is critical to recognize that the path to effective agriculture management is dynamic and changing. Continuous additions, iterative improvements, and the incorporation of future technology will be critical to the Agriculture Management System's long-term relevance and effect. Furthermore, encouraging collaborations with agricultural professionals, research institutes, and government agencies can broaden the system's capabilities and help to the overall development of the agricultural sector.

In essence, the completion of this work represents a big step toward increasing efficiency and collaboration in agricultural management. The system not only tackles current issues, but also pave the way for a technologically advanced, networked, and healthy agricultural ecosystem. As we grow the seeds of innovation and collaboration, the Agriculture Management System demonstrates technology's revolutionary capacity in transforming agriculture's future.

Reference

[1]. Agricultural Information Management Standards. (n.d). In *Wikepedia.* Retrieved from https://en.wikipedia.org/wiki/Agricultural_Information_Management_Standards#:~:text=Agricultural%20Information%20Management%20Standards%20(AIMS,build%20a%20global%20community%20of

[2]. Benos, L., Tagarakis, A. C., Dolias, G., Berruto, R., Kateris, D., & Bochtis, D. (2021). Machine learning in agriculture: A comprehensive updated review. *Sensors,* 21(11), 3758. https://doi.org/10.3390/s21113758

[3]. Jackson, C.R., & Daughetry, L.S. (1905). *Agriculture through the Laboratory and School Garden.* Orange JUDD Comnpany, New York.

[4]. Kropotkin, P. (1965). *Fields, Factories and Workshops Tomorrow.* Center for a Stateless Society, New York.

[5]. Thomas, E., Riley, M., & Spees, J. (2020). Knowledge flows: Farmers' social relations and knowledge sharing practices in 'Catchment Sensitive Farming'. *Land use Policy, 90,* 104254.

[6]. Warner, S.B. (2019). *To dwell is to Garden: A History of Boston's Community Gardens.* Northeastern University Press, Boston.

ABOUT THE AUTHORS

Takondwa Kaiya is a final-year student at DMI St. John the Baptist University in Malawi, pursuing a Bachelor's degree in Computer Science. Additionally, she holds an Advanced Diploma in Computing and Information Systems. The focus of her research is to contribute to the availability of online information for Malawians, particularly in the area of Agriculture Management Systems.

Chipatso Medi is a Lecturer in Computer Science at DMI Saint John the Baptist University. He holds an Honors degree in Business Information Technology and a Master of Science in Information Technology. In addition to his academic role, he is actively engaged in the private sector as an ICT Consultant. With a background as a System's Analyst and ICT Specialist, he specializes in IT Infrastructure and Cyber Security.

Fanny Chatola currently serves as a Lecturer II and the Head of the Department of Computer Science and Information Technology at the University of DMI – St. John the Baptist in Lilongwe, Malawi. The university is accredited by the National Council for Higher Education (NCHE). Fanny holds a Masters degree and a Bachelor's degree in Computer Science. She has actively participated in research projects, including the MEI conference with the Malawi University of Science and Technology and the International Conference on Business Globalization and Global Crisis Management (BGGM'23).

UNVEILING SENTIMENT ANALYSIS: A COMPARATIVE STUDY OF LSTM AND LOGISTIC REGRESSION MODELS WITH XAI INSIGHTS

By

CHANDU D. VAIDYA * MAYURI BOTRE ** YASH ROKDE ***

SAGAR KUMBHALKAR **** SOHAM LINGE *****

SOHAM PITALE ****** SHREYASH BAWNE *******

*-******* *Department of Computer Science and Engineering, S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.*

Date Received: 04/03/2024 *Date Revised: 18/03/2024* *Date Accepted: 27/03/2024*

ABSTRACT

In this study, we delve into sentiment analysis and the role of Explainable Artificial Intelligence (XAI), with a focus on techniques such as Lime that bring transparency to machine learning (Logistic Regression) and deep learning (LSTM) models. We explore how ML predictions can be biased using XAI and how XAI helps us understand DL models used in sentiment analysis through research that has been made. Examining various research, we notice a gap – the lack of training and interpretation for both ML and DL models on the same dataset using XAI. Our research fills this gap, shedding light on ML and DL model predictions through XAI's lens. By completing our research work, we come to know that even with an accuracy level of 83% for the DL model, they outperform the ML model with an accuracy level of 92% in some cases. This distinction is only identified with XAI techniques, particularly Lime.

Keywords: XAI (Explainable Artificial Intelligence), LIME (Local Interpretable Model Agnostic Explanations), Sentiment Analysis, Machine Learning, Deep Learning.

INTRODUCTION

Explainable AI (XAI) (Gohel et al., 2021; Goebel et al., 2018) stands as a pivotal component in unraveling the complexities of machine learning and deep learning models. In the context of sentiment analysis for product reviews (Aqlan et al., 2019; Gohel et al., 2021; Shivaprasad & Shetty, 2017) XAI provides a crucial layer of transparency, allowing us to comprehend how these models arrive at their decisions. As we embark on this exploration, the goal is clear: to demystify the decision-making processes of both traditional machine learning, embodied by logistic regression, and state-of-the-art deep learning, embodied by Long Short-Term Memory (LSTM) networks (Staudemeyer & Morris, 2019). XAI techniques, such as Lime, play a pivotal role in our study by elucidating the intricate inner workings of sentiment analysis models. In the era of complex machine learning models, the need for transparency and interpretability has become paramount. Local Interpretable Model-agnostic Explanations, or Lime (Dieber & Kirrane, 2020) emerges as a powerful tool designed to shed light on the inner workings of these intricate models. Lime operates on the premise that while machine learning models may be highly accurate, their decision-making processes often resemble a "black box" – opaque and inscrutable. Lime seeks to demystify this black box by providing locally faithful explanations for individual predictions. This approach not only enhances our understanding of how a model reaches a particular decision but also enables

i-manager Publications
Member of:
This paper has objectives related to SDG
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

users to validate the model's credibility. Lime, as a model-agnostic technique, can be applied to various machine learning models and deep learning models, fostering interpretability across different domains. In the ever-evolving landscape of sentiment analysis for product reviews, understanding the decisions made by machine learning and deep learning models is crucial. This research delves into the interpretability of such models, utilizing both traditional machine learning, represented by logistic regression, and cutting-edge deep learning, specifically Long Short-Term Memory (LSTM) networks. The overarching goal is to shed light on how these models perceive sentiments within the context of product reviews. While sentiment analysis has made significant strides, questions linger regarding the transparency of these models. To address this, we leverage the Lime XAI technique—a tool that grants us insights into the decision-making process of our models. What sets this study apart is the exploration of both machine learning and deep learning on the same dataset. Remarkably, this avenue remains largely unexplored in prior research, making our approach distinctive. By inspecting sentiment analysis models through the lens of interpretability, we contribute not only to the enhancement of model trustworthiness but also to the broader field of explainable AI. The subsequent sections unfold the methodology, results, and implications of this exploration, aiming to demystify the intricacies of sentiment analysis in product reviews. Logistic Regression (Tyagi & Sharma, 2018) is a statistical method used for binary classification problems. In the context of sentiment analysis, it's commonly employed to predict whether a given text belongs to a positive or negative sentiment class. The algorithm models the probability that a given input belongs to a particular category. LSTM is a type of recurrent neural network (RNN) (Arras et al., 2017) architecture designed to capture long-term dependencies in sequential data. In the context of sentiment analysis, LSTM networks are effective in understanding the context and relationships between words in a sentence.

1. Literature Work

This research is crucial in the field of sentiment analysis for product reviews because it introduces a new way of understanding how machine learning and deep learning models make predictions. By using explainable AI techniques, we aim to reveal the 'why' behind these predictions, making the decision-making process more transparent and trustworthy.

Love et al. (2023) talked about two types of artificial intelligence: machine learning (ML) and deep learning (DL). The paper introduces explainable artificial intelligence (XAI) to help us understand and trust these systems better although XAI has not been explored much in construction, and the paper suggests its potential benefits in making AI more understandable and accepted in the construction industry. but lacks in demonstration of practical experimentation. So (2020) highlighted the effectiveness of machine learning algorithms in predicting product ratings based on online reviews. However, the study's findings emphasize the need for caution, revealing potential biases in the data that impact predictions. This underscores the importance of not relying on machine learning algorithms entirely, as biases in datasets can affect the accuracy and reliability of predictions. Adak et al. (2022) looked at why more people are using food delivery services during COVID-19 and how online reviews can help improve these services. To handle the large amount of feedback, the study suggests using artificial intelligence (AI). It tests different AI methods and finds that one called LSTM is the most accurate. They also use techniques to explain why the AI makes certain predictions, helping to understand which words in reviews contribute to positive or negative sentiments. The goal is to make food delivery services better by addressing customer complaints using AI and understanding what customers are saying in their reviews. Yadav and Vishwakarma (2020) the primary purpose of this survey is to highlight the power of deep learning architectures for solving sentiment analysis problems. Ribeiro et al. (2016) talked about how to use the LIME Technique to build a more trustable AI model. Lundberg and Lee (2017) talked about how to use the SHAP Technique to build a more trustable AI model. Xu et al. (2017) introduces the history of Explainable AI, starting

from expert systems and traditional machine learning approaches to the latest progress in the context of modern deep Learning, describing the major research areas and the state-of-art approaches in recent years.

Goebel et al. (2018) talks about how Explainable AI (XAI) is not a new concept, dating back to early AI systems in the 1980s. Early expert systems lacked learning capabilities but utilized abductive reasoning. Recent AI success with neural networks, while powerful, struggles with un-debuggability and lacks explainability for human understanding. Future AI challenges include developing models that can provide explanations, incorporate human expertise, and facilitate contextual adaptation for real-world problem-solving. Raza et al. (2019) looks at using computer programs to understand the feelings expressed in scientific articles. The goal is to help researchers find good-quality papers by analyzing the sentiments in the articles. They used a dataset of over 8,000 citation sentences and tested six different computer algorithms to classify the sentiment. By cleaning up the data and adding some extra features, like lemmatization and removing unnecessary words, they improved the system's accuracy by up to 9% compared to the basic setup. Hoffman et al. (2018) talked about the goodness of explanations of how well users understand AI systems and how the human-XAI work system performs. Athar (2014) examined sentiment analysis in scientific literature, particularly focusing on opinions expressed in citations. It underscores the importance of understanding sentiments toward cited papers for various purposes, such as evaluating research quality and identifying gaps in current approaches.

Xia et al. (2011) compared different ways of improving how computers understand feelings in text. It explores combining various methods and tools to make these computer predictions more accurate. Silva and Ribeiro (2003) mention how important is removing stop words for developing large models. According to Jagdale et al. (2019), opinion mining, data processing ML algorithms SVM, and NB techniques are efficient for product review.

Mathew and Bindu (2020) discussed various word embedding methods used for sentiment analysis, an overview of state-of-the-art pre-trained models used for natural language processing, which is commonly used in the process of sentiment analysis. Gohel et al. (2021) talked about different ways to make AI more understandable, especially when dealing with text, images, audio, and video. It also discusses what these methods do well and where they can be improved, offering suggestions for future research. Hagras (2018) talked about XAI concepts and mentions areas, like type-2 fuzzy logic systems, that need more exploration. This is to make sure that regular users can understand and analyze AI systems easily, building confidence in their use. Nasukawa and Yi (2003) illustrates a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document instead of classifying the whole document as positive or negative. Sperandei (2014) mention Logistic regression is a tool used in statistics to understand how different factors influence a situation with two possible outcomes. The key goal is to figure out how each factor affects the likelihood of a specific outcome. By looking at all factors together, it helps to get a clear picture and avoid confusion. The article explains how this process works with examples, making it easy to understand. It also covers some important points to consider in the analysis. (Vaidya et al., 2023) focused on a sarcastic review system.

2. Research Gap

Reviewing prior studies (Table 1), we identify a gap in exploring both machine learning (ML) and deep learning (DL) models on the same dataset using XAI. Our research aims to fill this gap by thoroughly examining ML and DL models on a shared dataset, emphasizing the importance of XAI for better model understanding. Additionally, we will compare both models based on their explanations using Local Interpretable Model-Agnostic Explanations (LIME).

3. Methodology

The architecture diagram (Figure 1) for constructing a sentiment analysis model will use Machine Learning (Logistic Regression Algorithm) and Deep Learning (Long Short-Term Memory Algorithm (LSTM)) with LIME (Local

Authors and Citation	Methods	Findings	Drawbacks
So (2020)	Machine Learning Algorithms, XAI	Underscores the importance of not relying on machine learning algorithms entirely, as biases in datasets can affect the accuracy and reliability of predictions.	Global analysis of feature importance is insufficient to detect such bias.
Adak et al. (2019,2022)	LSTM, XAI	The goal is to make food delivery services better by addressing customer complaints using AI and understanding what customers are saying in their reviews.	Handling Large amounts of feedback information
Love et al. (2023)	Machine Learning (ML), Deep Learning (DL), and XAI	The paper introduces explainable artificial intelligence (XAI) to help us understand and trust these systems better.	XAI has not been explored much in construction
Yadav and Vishwakarma (2020)	Deep Learning	The primary purpose of this survey is to highlight the power of deep learning architectures for solving sentiment analysis problems.	Lack of use of XAI Methods for model interpretation.
Ribeiro et al. (2016)	LIME	This paper talks about how to use the LIME Technique to build a more trustable AI model.	Only covers the one XAI Technique i.e. LIME
Xu et al. (2019)	Theoretical Concepts of XAI	Introduces the history of Explainable AI, starting from expert systems and traditional machine learning approaches.	Only Theoretical concepts are covered
Lundberg and Lee (2017)	SHAP	This paper talks about how to use SHAP Technique to build a more trustable AI model.	Only covers the one XAI Technique i.e. SHAP
Raza et al. (2019)	Six Different Computer Algorithms	They improved the system's accuracy by up to 9% compared to the basic setup.	Limited to the field of Sentiment Analysis
Silva and Ribeiro (2003)	Stop Words Removal	Tells how important is removing stop words for developing large models.	Doesn't Explore other NLP Techniques
Mathew and Bindu (2020)	Word Embeddings	Discusses various word embedding methods used for sentiment analysis	Research is limited to Pre-trained models
Gohel et al. (2021)	XAI Methods	The paper talks about different ways to make AI more understandable	Again Only limited to Theoretical concepts.

Table 1. Review of Literature

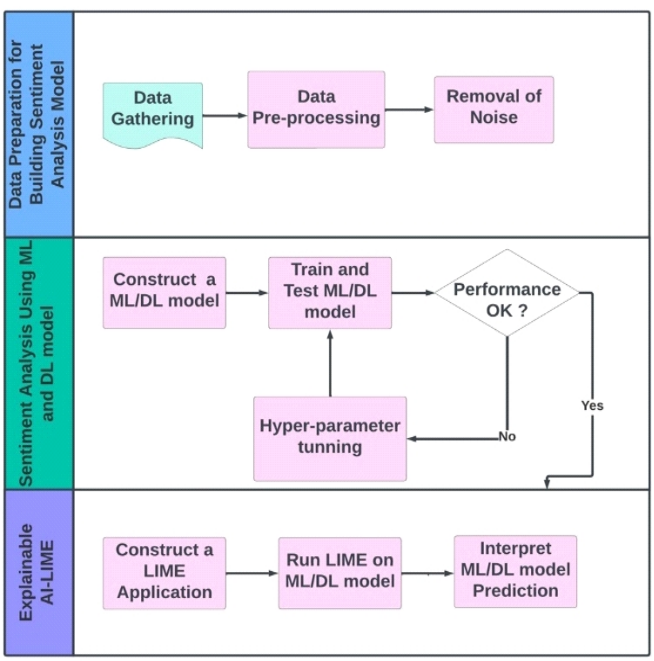


Figure 1. Architecture Diagram for Building Model

Interpretable Model Agnostic Explanation).

Fundamental steps will be undertaken to build a sentiment analysis model using both machine learning (ML) and deep learning (DL) approaches. It's crucial to note that these steps will offer a high-level overview of the process.

- *Data Collection:* The dataset for this study will be sourced from Kaggle, specifically the "Amazon Fine Food Reviews" (McAuley & Leskovec, 2013) dataset. Two key columns, "text" and "sentiment," will be essential to the research.
- *Data Preprocessing:* Data preprocessing will involve applying Natural Language Processing (NLP) (William et al., 2023) techniques to refine the text data. This will include the removal of stop words, tokenization, padding, and the conversion of sentiment labels, where negative sentiments will be represented as 0 and positive sentiments as 1.
- *Model Selection and Training:* For the machine learning component (Pang et al., 2002), the logistic regression model will be chosen, while the deep learning aspect will utilize the Long Short-Term Memory (LSTM) algorithm (Wang et al., 2016). Both models will be trained on the same dataset.
- *Model Evaluation:* The models will be assessed based on accuracy and various metrics such as F1 score,

precision, and recall values. The evaluation process will also incorporate Lime XAI (Explainable AI) techniques to gain insights into the interpretability of the models.

- *Model Testing:* The final phase will involve testing the models with different inputs to assess their performance and interpretability based on the explanations they provide. This comprehensive approach will aim to scrutinize the effectiveness and transparency of both machine learning and deep learning models in sentiment analysis for product reviews.

This study aims to advance the field of sentiment analysis for product reviews by integrating Explainable AI (XAI) techniques, ultimately providing more accurate sentiment classifications and transparent insights into the reasons behind these classifications.

4. Objectives

- To develop a sentiment analysis model.
- To interpret the model with the Explainable AI (LIME) technique.
- Comparison between machine learning and deep learning model on the same product dataset based on their interpretability.

5. Implementation

Proposed Algorithm:

START:

Step 1: Data Collection (Amazon fine food reviews from Kaggle).

Step 2: Data Preprocessing (tokenization, vectorization, stemming, padding, embedding).

Step 3: Model Training and Selection (logistic regression and long short-term memory)

Step 4: Model Evaluation (Accuracy, F1 score, recall, etc. LIME)

Step 5: Model testing on various user inputs

END:

This is the flow for developing our sentiment analysis model using Machine Learning (Logistic Regression Algorithm) and Deep Learning (Long Short-Term Memory Algorithm (LSTM)) with LIME (Local Interpretable Model Agnostic Explanation.

The process starts with data collection from Amazon.com. This data is then pre-processed using NLP techniques (Sun et al., 2014). Here, the text is tokenized, and converted into lowercase, and stop words and punctuation are removed. The next step involves stemming or lemmatization, where words are reduced to their base form. Part-of-speech tagging and named entity recognition may be used to identify specific entities like products or locations.

Next, sentiment lexicons are used to score the positive and negative sentiments of each word. This could involve using pre-built lexicons like Senti-Word-Net or building custom lexicons based on the specific domain of the reviews. The sentiment score of each word is then aggregated to get a sentiment score for the entire review.

To improve the model's accuracy, negation handling is performed. This identifies words like "not" or "never" that can reverse the sentiment of a phrase. For example, "not bad" would have a positive sentiment instead of negative.

The pre-processed data is split into training and testing sets. The model is then trained on the training data using a machine learning and deep learning model. During training, the model learns to identify patterns in the data that associate certain words or phrases with positive or negative sentiments.

Once trained, the model is evaluated using metrics like LIME, F1 score, recall, and precision. LIME provides insights into which words or phrases influenced the model's predictions for specific reviews. The F1 score considers both precision (the ability to identify positive reviews correctly) and recall (the ability to identify all positive reviews). Recall is important for identifying as many positive reviews as possible, while precision is important for minimizing false positives. If the model performs well, it is then tested on user input. This could involve feeding the model a new review and seeing if it can correctly predict its sentiment. If the model does not perform well, the

process is repeated from the data preprocessing step. This could involve trying different NLP techniques, sentiment lexicons, or increasing the size of the dataset.

6. Results

6.1 Results for Logistic Regression Algorithm

The Logistic Regression model, implemented using a sophisticated pipeline, exhibits commendable performance across various evaluation metrics. With a precision score of approximately 92.08%, the model demonstrates a high degree of accuracy in identifying positive instances among all predicted positives. Simultaneously, the recall score of 92.08% signifies the model's effectiveness in capturing the majority of actual positive cases. The Area under the ROC Curve (AUC) attaining 95.03% emphasizes the model's robust ability to distinguish between positive and negative instances. This elevated AUC score showcases the model's discriminatory power and reliability in making accurate predictions.

Furthermore, the harmonic mean of precision and recall, encapsulated in the F1 score (91.61%), indicates a balanced performance in handling both false positives and false negatives. The accuracy score, standing at 92.08%, underscores the overall correctness of the model's predictions. In summary, the Logistic Regression model, bolstered by a well-designed pipeline, not only achieves high accuracy but also excels in precision, recall, AUC, and F1 score (Table 2), collectively contributing to its effectiveness in sentiment analysis for product reviews.

6.2 Results for LSTM Algorithm

In evaluating the LSTM model's performance, the calculated loss of approximately 0.3952 signifies the model's ability to minimize the discrepancy between predicted and actual sentiments in the test dataset. This low loss suggests a strong alignment between the model's predictions and the true sentiment labels. The achieved accuracy of 83.57% further emphasizes the model's proficiency in correctly predicting sentiments.

Evaluation Matrix for Sentiment Analysis Model

Parameter	LSTM	Logistic Regression
Accuracy	0.83	0.92
Recall	(for 0=0.54, for 1=0.93)	0.92
F1 Score	(for 0=0.61, for 1=0.90)	0.91
Precision	(for 0=0.70, for 1=0.87)	0.92

Table 2. Evaluation Metrics for Sentiment Analysis Model

Breaking down the classification report, the precision scores for sentiment classes 0 and 1 are 0.70 and 0.87, respectively. These values indicate the accuracy of the model in positive predictions. The recall scores, measuring the model's ability to capture actual positive instances, are 0.54 for class 0 and 0.93 for class 1. The F1 scores, representing the harmonic mean of precision and recall, are 0.61 for class 0 and 0.90 for class 1. The overall metrics highlight an accuracy of 83.57%, with macro average values of 0.78 precision, 0.73 recall, and 0.75 F1-score. The weighted average metrics, considering the number of instances for each class, yield values of 0.83 precision, 0.84 recall, and 0.83 F1-score. This comprehensive assessment showcases the model's effectiveness in sentiment analysis on the test dataset.

These are the results obtained by LSTM and Logistic regression after successfully applying the LIME technique. From Figures 2 and 3, it is observed that both models perform equally well on complex user inputs, even though there is an accuracy difference between the logistic regression and LSTM models.

After evaluating our models on a test text expressing a negative sentiment towards a product, we observed interesting results. The LSTM model predicted a 60% chance of negativity (class 0) and a 40% chance of positivity (class 1) according to Lime. On the other hand, despite having a higher overall accuracy of 92%, the Logistic Regression model showed a probability of only 5% for negativity and 95% for positivity. This discrepancy suggests that, even with a lower accuracy of 83%, the LSTM model, guided by Lime explanations, demonstrated a better understanding of the nuanced sentiment expressed in the text. Lime's insights (Figures 4 and 5) highlighted the LSTM model's capacity to capture the subtleties of sentiments, showcasing the importance of interpretability tools in comprehending and trusting model predictions. This level of model understanding, made possible by explainable Artificial Intelligence (XAI),

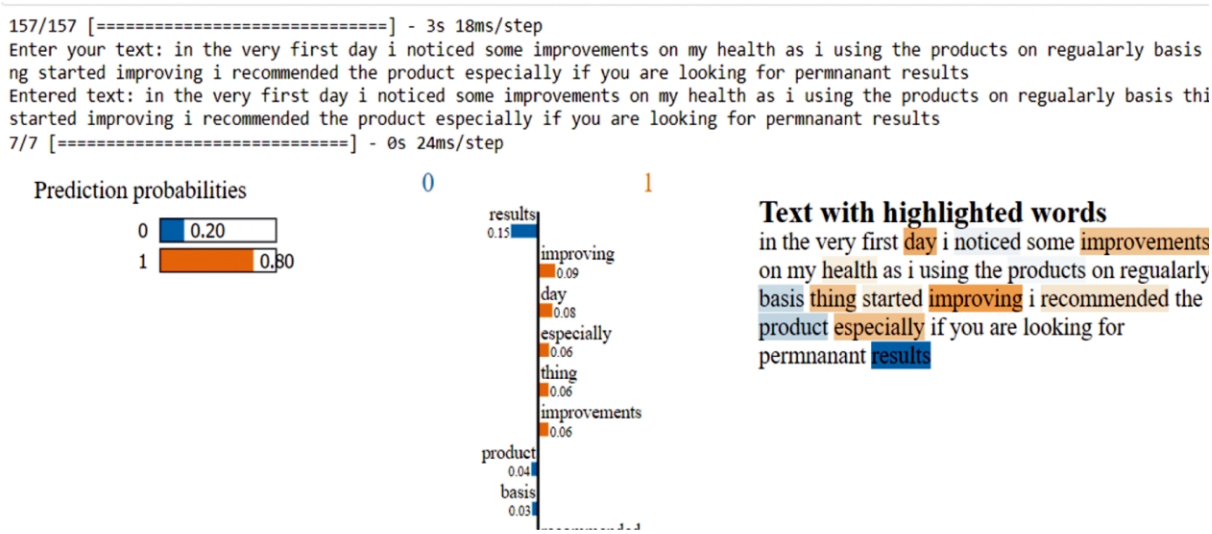


Figure 2. LSTM Result on Complex Inputs

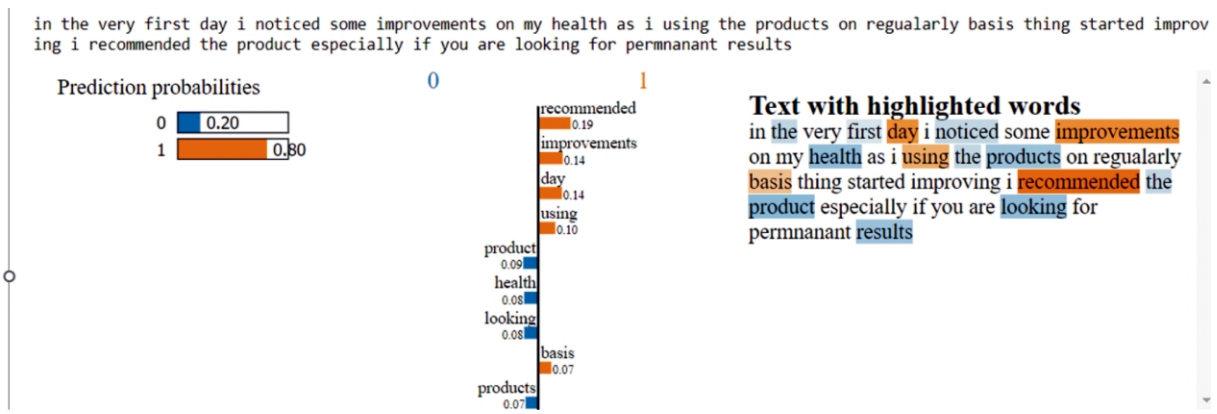


Figure 3. Logistic Regression Results on Complex Inputs

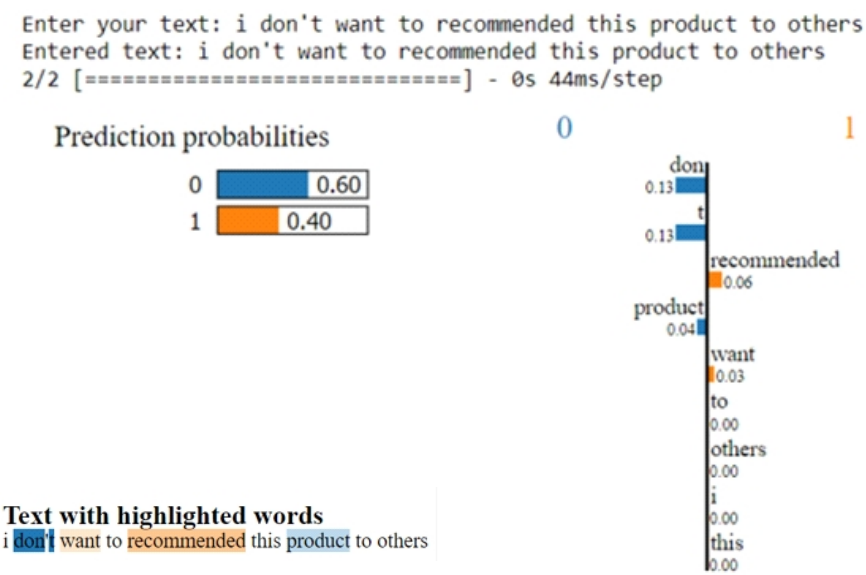


Figure 4. LIME Results for LSTM for Negative Sentiment

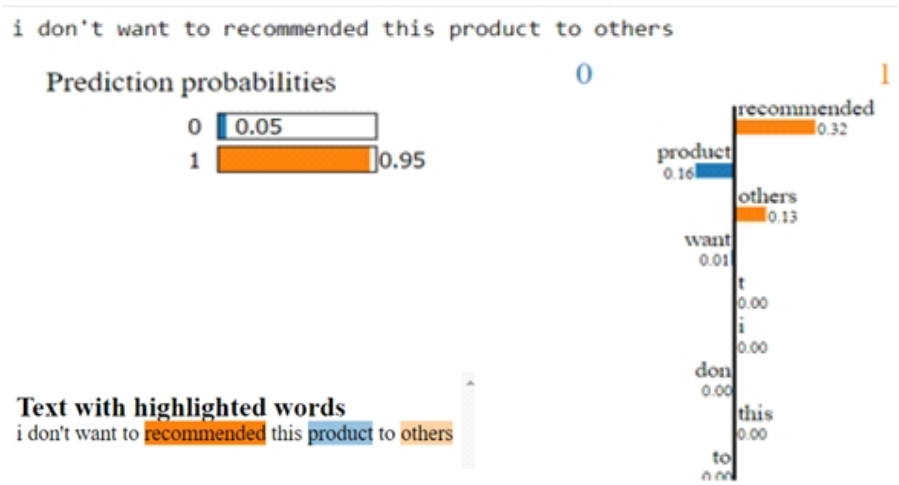


Figure 5. LIME Results for Logistic Regression for Negative Sentiment

emphasizes the critical role of transparency in model decision-making.

Conclusion

Our research findings are fascinating. When testing models on negative sentiments, the LSTM, guided by Lime explanations, showed a detailed prediction (60% negativity, 40% positivity). Surprisingly, the Logistic Regression model, despite its 92% accuracy, gave low negativity (5%) and high positivity (95%). These results are consistent across various cases, highlighting the LSTM's nuanced understanding. This showcases the crucial role of Lime and explainable AI (XAI) in making AI decisions clear and reliable.

References

[1]. Adak, A., Pradhan, B., Shukla, N., & Alamri, A. (2022). Unboxing deep learning model of food delivery service reviews using explainable artificial intelligence (XAI) technique. *Foods,* 11(14), 2019. https://doi.org/10.3390/foods11142019

[2]. Aqlan, A. A. Q., Manjula, B., & Naik, R. L. (2019). A study of sentiment analysis: Concepts, techniques, and challenges. In *Proceedings of International Conference on Computational Intelligence and Data Engineering: Proceedings of ICCIDE 2018* (pp. 147-162). Springer Singapore. https://doi.org/10.1007/978-981-13-6459-4_16

[3]. Arras, L., Montavon, G., Müller, K. R., & Samek, W. (2017). Explaining recurrent neural network predictions in sentiment analysis. *arXiv preprint arXiv:1706.07206.* https://doi.org/10.48550/arXiv.1706.07206

[4]. Athar, A. (2014). *Sentiment analysis of scientific citations.* University of Cambridge, Computer Laboratory (pp. 1-114). https://doi.org/10.48456/tr-856

[5]. Dieber, J., & Kirrane, S. (2020). Why model why? Assessing the strengths and limitations of LIME. *arXiv preprint arXiv:2012.00093.* https://doi.org/10.48550/arXiv.2012.00093

[6]. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018). Explainable AI: The new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction* (pp. 295-303). Springer, Cham. https://doi.org/10.1007/978-3-319-99740-7_21

[7]. Gohel, P., Singh, P., & Mohanty, M. (2021). Explainable AI: current status and future directions. *arXiv preprint arXiv:2107.07045.* https://doi.org/10.48550/arXiv.2107.07045

[8]. Hagras, H. (2018). Toward human-understandable, explainable AI. *Computer,* 51(9), 28-36. https://doi.org/10.1109/MC.2018.3620965

[9]. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608.* https://doi.org/10.48550/arXiv.1812.04608

[10]. Jagdale, R. S., Shirsat, V. S., & Deshmukh, S. N. (2019). Sentiment analysis on product reviews using machine learning techniques. In *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017* (pp. 639-647). Springer Singapore. https://doi.org/10.1007/978-981-13-0617-4_61

[11]. Love, P. E., Fang, W., Matthews, J., Porter, S., Luo, H., & Ding, L. (2023). Explainable artificial intelligence (XAI): Precepts, models, and opportunities for research in construction. *Advanced Engineering Informatics,* 57, 102024. https://doi.org/10.1016/j.aei.2023.102024

[12]. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems,* 30, 1-10.

[13]. Mathew, L., & Bindu, V. R. (2020, March). A review of natural language processing techniques for sentiment analysis using pre-trained models. In 2020 *Fourth International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 340-345). IEEE. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-00064

[14]. McAuley, J. J., & Leskovec, J. (2013, May). From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 897-908). https://doi.org/10.1145/2488388.2488466

[15]. Nasukawa, T., & Yi, J. (2003, October). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture* (pp. 70-77). https://doi.org/10.1145/945645.945658

[16]. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070.* https://doi.org/10.48550/arXiv.cs/0205070

[17]. Raza, H., Faizan, M., Hamza, A., Mushtaq, A., & Akhtar, N. (2019). Scientific text sentiment analysis using machine learning techniques. *International Journal of Advanced Computer Science and Applications,* 10(12), 157-165.

[18]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

[19]. Shivaprasad, T. K., & Shetty, J. (2017, March). Sentiment analysis of product reviews: A review. In 2017 *International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 298-301). IEEE. https://doi.org/10.1109/ICICCT.2017.7975207

[20]. Silva, C., & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003,* 3,1661-1666. IEEE. https://doi.org/10.1109/IJCNN.2003.1223656

[21]. So, C. (2020). What emotions make one or five stars? Understanding ratings of online product reviews by sentiment analysis and XAI. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020* (pp. 412-421). Springer International Publishing. https://doi.org/10.1007/978-3-030-50334-5_28

[22]. Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica,* 24(1), 12-18. https://doi.org/10.11613/BM.2014.003

[23]. Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586.* https://doi.org/10.48550/arXiv.1909.09586

[24]. Sun, F., Belatreche, A., Coleman, S., McGinnity, T. M., & Li, Y. (2014, March). Pre-processing online financial text for sentiment classification: A natural language processing approach. In 2014 *IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)* (pp. 122-129). IEEE. https://doi.org/10.1109/CIFEr.2014.6924063

[25]. Tyagi, A., & Sharma, N. (2018). Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology (UAE),* 7(2.24), 20-23.

[26]. Vaidya, C., Gupta, A., Shastrakar, A., Kathane, A.,

Atmande, K., & Iyer, K. (2023, February). Sarcasm detection analysis–Comparative analysis. In 2023 *IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1-4). IEEE. https://doi.org/10.1109/SCEECS57921.2023.10063107

[27]. Wang, J., Yu, L. C., Lai, K. R., & Zhang, X. (2016, August). Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics,* 2, 225-230.

[28]. William, P., Shrivastava, A., Chauhan, P. S., Raja, M., Ojha, S. B., & Kumar, K. (2023). Natural Language processing implementation for sentiment analysis on tweets. In *Mobile Radio Communications and 5G Networks: Proceedings of Third MRCN 2022* (pp. 317-327). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7982-8_26

[29]. Xia, R., Zong, C., & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences,* 181(6), 1138-1152. https://doi.org/10.1016/j.ins.2010.11.023

[30]. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019 (pp. 563-574). Springer International Publishing.* https://doi.org/10.1007/978-3-030-32236-6_51

[31]. Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review,* 53(6), 4335-4385. https://doi.org/10.1007/s10462-019-09794-5

ABOUT THE AUTHORS

Chandu D. Vaidya is an Assistant Professor at S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra, India, specializing in machine learning and deep learning. He is pursuing Ph.D. from VIT-Bhopal University, India, and has made significant contributions to their field. He has published articles in respected academic journals and presented at international conferences, including WOS, Scopus, IEEE, etc. He has published a patent and a book on Operating Systems, and he has obtained copyrights on various topics.

Mayuri Botre is an Assistant Professor at S.B. Jain Institute of Technology, Management and Research, Nagpur, Maharashtra, India.

Yash Rokde is currently pursuing a Bachelor of Technology Degree in Computer Science and Engineering at S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.

Sagar Kumbhalkar is currently pursuing a Bachelor of Technology Degree in Computer Science and Engineering at S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.

Soham Linge is currently pursuing a Bachelor of Technology Degree in Computer Science and Engineering at S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.

Soham Pitale is currently pursuing a Bachelor of Technology Degree in Computer Science and Engineering at S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.

Shreyas Bawane is currently pursuing a Bachelor of Technology Degree in Computer Science and Engineering at S.B. Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India.

RESUME SCREENER SYSTEM

By

MUHAMMAD SAVAD N. * T. PREETHI **

* Department of Computer Science, Nilgiri College of Arts and Science (Autonomous), Thaloor, Nilgiri, Tamil Nadu, India.
** Department of Multimedia & Web Technology, Nilgiri College of Arts and Science(Autonomous), Thaloor, Nilgiri, Tamil Nadu, India.

Date Received: 07/03/2024 Date Revised: 16/03/2024 Date Accepted: 25/03/2024

ABSTRACT

This research paper introduces a state-of-the-art "Resume Screener System" aimed at revolutionizing and automating the labor-intensive task of resume analysis for recruitment purposes. Developed using Python, the system integrates Artificial Intelligence and Natural Language Processing techniques to streamline the hiring process. Utilizing a dataset sourced from Kaggle, comprising a thousand resumes converted into textual data, the system undergoes comprehensive model training and evaluation. Employing advanced machine learning methodologies such as the Support Vector Classifier (SVC) and Neighbours Classifier, the system rigorously tests and analyzes these models to determine the most effective approach. By evaluating each model's performance against predefined criteria, the system identifies the optimal model for resume screening. The primary objective of this work is to provide recruiters and HR professionals with an innovative tool that efficiently matches job requirements with candidates' skill sets as presented in their resumes. By automating the initial screening phase, the system not only saves time and effort but also ensures a more objective and consistent evaluation of applicants. This research contributes to the advancement of machine learning applications in the field of human resources, illustrating the transformative impact of technology on traditional hiring practices.

Keywords: Support Vector Classifier, KNN Classifiers, OneVSRest Classifier, tfidf Vectorizer, Natural Language Processing.

INTRODUCTION

In response to the evolving demands of modern recruitment processes, this research initiative aims to develop an advanced "Resume Screener System" that will revolutionize traditional hiring methodologies through the strategic integration of cutting-edge Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies. The system's core objective is to streamline the labor-intensive task of resume analysis, thereby enhancing efficiency and accuracy in candidate evaluation. Developed using Python as its principal programming language, the system ensures seamless integration of sophisticated machine learning algorithms, catering to the diverse needs of recruiters and HR professionals.

Central to the system's effectiveness is its utilization of a comprehensive dataset sourced from Kaggle, comprising a thousand resumes meticulously transformed into textual representations. This rich dataset serves as the foundation for the system's multifaceted approach to resume analysis, empowering users with the tools to efficiently match job requirements with the diverse skill sets presented in applicant resumes. Key to its functionality is the implementation of the KNeighbors Classifier, a crucial machine learning module that enhances the system's ability to discern intricate patterns and accurately classify resumes based on their proximity to similar profiles. Additionally, the integration of the One-vs-Rest Classifier further amplifies the system's capacity to tackle multi-class classification challenges, providing a

comprehensive understanding of each resume's primary focus areas.

As the development progresses, the work remains committed to delivering a sophisticated and reliable tool for resume screening. The system aspires to offer a user-friendly experience, bridging the gap between technological innovation and recruitment efficiency. By empowering both data scientists and developers with intuitive functionalities and streamlined workflows, the system endeavors to redefine the recruitment landscape, ushering in a new era of efficiency, accuracy, and inclusivity in candidate evaluation processes.

1. Literature Review

The literature review examines prior studies on resume screener systems. It assesses the effectiveness of these systems in streamlining job searches and their impact on students' professional growth and transition into the workforce, while identifying gaps for further exploration in the present study.

Priyanka and Parveen (2023) introduces a hybrid deep learning-based approach, combining Pyramid Dilated Convolutional Neural Network with Bidirectional Gated Recurrent Unit (PDCNN-Bi-GRU), to extract skill-related features from resumes and match them with job categories. Surendhiran et al. (2023) proposes an automated solution for classifying resumes to their suitable positions, aiming to streamline the time-consuming and expensive manual screening process faced by hiring companies.

Bachate et al. (2023) suggested that college community now has access to a transformative portal offering real-world interview experiences and a cutting-edge resume analysis tool, streamlining the job search process and broadening horizons for professional growth. VeeraSekharReddy et al. (2022) introduces the AL-CRF model, combining Conditional Random Field and Active Learning for Named Entity Recognition, which iteratively trains the classifier until stability, offering a more efficient and cost-effective approach to NLP tasks.

Thatha et al. (2023) proposes an Enhanced Support Vector Machine based Pattern classification method (ESVMPCM) to improve pattern classification accuracy, demonstrated on the Reuters dataset with various performance metrics. Amin et al. (2019) introduces a web application tailored for resume screening in job recruitment, featuring interactive functionalities for both applicants and recruiters. Leveraging machine learning and Natural Language Processing techniques, the system evaluates resumes against job requirements, providing recruiters with ranked candidate lists for efficient selection.

Zaroor et al. (2018) proposes a hybrid approach to automatically extract structured information from diverse resumes and efficiently match them with relevant job offers. By employing conceptual-based classification and leveraging an integrated knowledge base, promising precision results are achieved compared to conventional machine learning methods, as demonstrated with real-world recruitment data. Schmitt et al. (2016) explores automatic job seeker and recruiter matching using data from a recruitment agency, revealing discrepancies between recommendation performance in collaborative filtering and cold start modes, potentially attributed to language differences.

Kmail et al. (2015, November) proposes an automatic online recruitment system that addresses the limitations of traditional methods by leveraging multiple semantic resources and statistical concept-relatedness measures. Through empirical validation, the system demonstrates potential in improving precision and efficiency in matching candidates to relevant job postings, offering a promising solution for modern recruitment challenges. Chen et al. (2015) presents a novel framework for extracting knowledge from resumes, independent of file format, through text segmentation and classification processes. Experimental results on real datasets demonstrate its superiority over previous methods in building structured resume repositories.

Kmail et al. (2015, August) introduces an automatic semantics-based online recruitment system that utilizes multiple semantic resources and statistical-based concept-relatedness measures to enhance the precision of matching candidate resumes with job posts. Senthil

Kumaran and Sankar (2013) introduces EXPERT, an intelligent e-recruitment tool leveraging ontology mapping for candidate screening. Through three phases, EXPERT constructs candidate and job requirement ontologies, enabling accurate matching and retrieval of eligible candidates.

2. Proposed System

The proposed "Resume Screener System" introduces an AI and NLP-powered solution aimed at automating and enhancing resume analysis in recruitment. Utilizing a robust dataset from Kaggle, it streamlines the labor-intensive task of aligning job requirements with applicant skill sets. By integrating advanced machine learning algorithms and text data processing, this system marks a significant advancement in modernizing talent acquisition processes.

3. Overall Working Model

The proposed resume screener system aims to automate and enhance the efficiency of resume analysis in recruitment processes. As shown in Figure 1, it begins with an initial module focused on dataset exploration, loading, and basic analysis to understand the structure and content of the dataset. This includes importing essential libraries, loading the dataset into memory, and exploring key attributes such as the distribution of resume categories. Following this, Module II focuses on dataset cleaning, employing text preprocessing techniques to remove noise and irrelevant information from the resumes' text content. This involves cleaning the resumes using regular expressions, removing stop words, and tokenizing the text data.

Module III involves training data preparation, where the dataset is split into training and testing sets, and machine learning models such as Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN) are trained using TF-IDF vectorization. The trained models are then evaluated on the test data to assess their performance in classifying resumes into job categories. Subsequently, Module IV tests the developed resume screening system by applying it to real-world resume data in PDF format. The system extracts text from PDF resumes, cleans and processes the text data, and predicts the corresponding job category using the trained machine learning models.

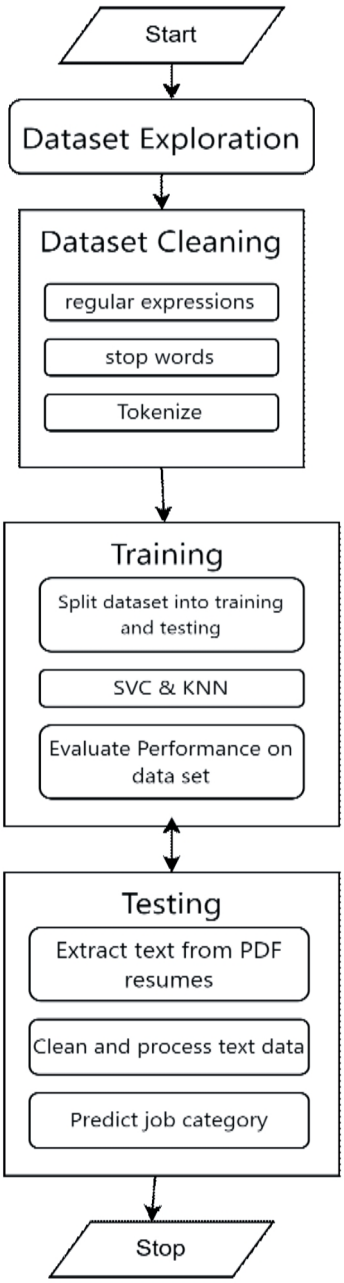


Figure 1.Overall Working Model for the Proposed System

The Proposed Resume Screener System can be categorised mainly into four modules. They are

- Dataset Exploration
- Dataset Cleaning
- Training
- Testing

3.1 Dataset Exploration

The module focused on "Dataset Exploration" is fundamental as shown in Figure 2. It emphasizes the importance of understanding the dataset's structure and content before further processing and modeling. It initiates with the importation of necessary tools for data analysis and visualization. Subsequently, the module loads the dataset into memory, addressing potential encoding errors to ensure seamless compatibility. Exploration tasks follow, including assessing the dataset's original length, displaying initial rows, and examining its dimensions.

Moreover, the module analyzes the distribution of resume categories, employing visualizations like bar charts and pie charts for intuitive representation. Lastly, it facilitates access to individual entries within the dataset, enabling detailed inspection and understanding of specific resume entries. Overall, this module establishes the groundwork for dataset exploration, crucial for comprehending dataset characteristics and guiding subsequent research steps. Such exploration allows researchers to adapt datasets based on specific requirements or the recruitment company's nature, ensuring relevance and efficacy in the resume screening application.

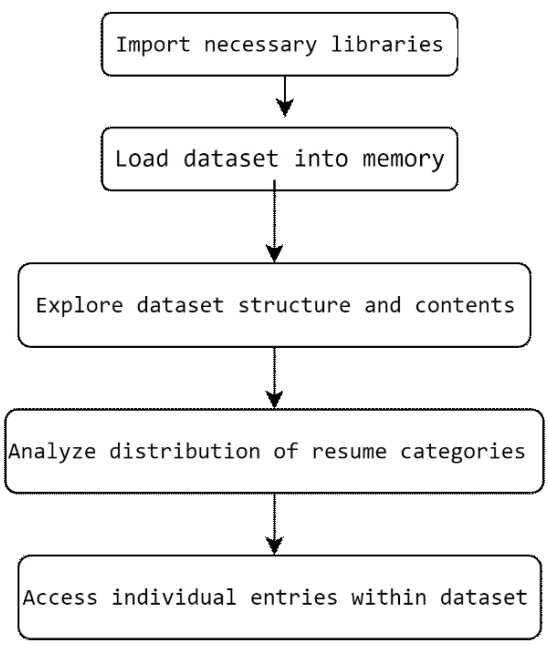


Figure 2. Working Model for the Data Exploration

3.2 Dataset Cleaning

Data Cleaning plays a vital role in preparing the text data for further analysis, classification, and clustering, ensuring the effectiveness of the resume screening application by enhancing the quality and consistency of the dataset. It focuses on preparing the text data by removing noise and irrelevant information from the resumes content to enhance the quality and consistency of the dataset.

As Shown in Figure 3, the module begins by utilizing regular expressions to eliminate various patterns such as URLs, mentions, hashtags, and punctuation from the text. Additionally, non-ASCII characters are replaced, and extra white spaces are removed to ensure text uniformity. Subsequently, common stop words are filtered out, enhancing the relevance of the cleaned resumes. In Figure 4, we can analyse the cleaned resume texts and actual data presented.

Moreover, the text data undergoes tokenization, breaking down each resume entry into individual words or tokens. This tokenization process facilitates subsequent text processing tasks, including lemmatization and feature extraction.

3.3 Training

In this module, it focuses on preparing the training data for machine learning model training, as shown in Figure 5, It begins with text data processing using NLTK for tokenization and lemmatization. The dataset is then split

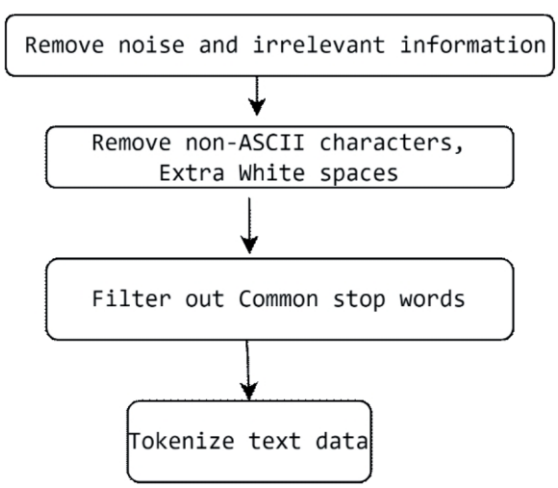


Figure 3. Working Model Data Cleaning

In [14]: resumeData

Out[14]:

	Category	Resume	cleaned_resume
0	Data Science	Skills * Programming Languages: Python (pandas...	skills programming languages python pandas num...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	education details may 2013 to may 2017 b e uit...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	areas of interest deep learning control system...
3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Table...	skills r python sap hana tableau sap hana sql ...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	education details mca ymcaust faridabad haryan...
...	...	...	...
958	Testing	â□□ Willingness to accept the challenges. â□□ ...	willingness to a ept the challenges positive ...
959	Testing	PERSONAL SKILLS â□¢ Quick learner, â□¢ Eagerne...	personal skills quick learner eagerness to lea...
960	Testing	COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...	computer skills software knowledge ms power po...
961	Testing	Skill Set OS Windows XP/7/8/8.1/10 Database MY...	skill set os windows xp 7 8 8 1 10 database my...
962	Data Science	MUHAMMAD SAVAD N muhammadsavadn muhammadsavadn...	muhammad savad n muhammadsavadn muhammadsavadn...

963 rows × 3 columns

Figure 4. Raw Data and Cleaned Data

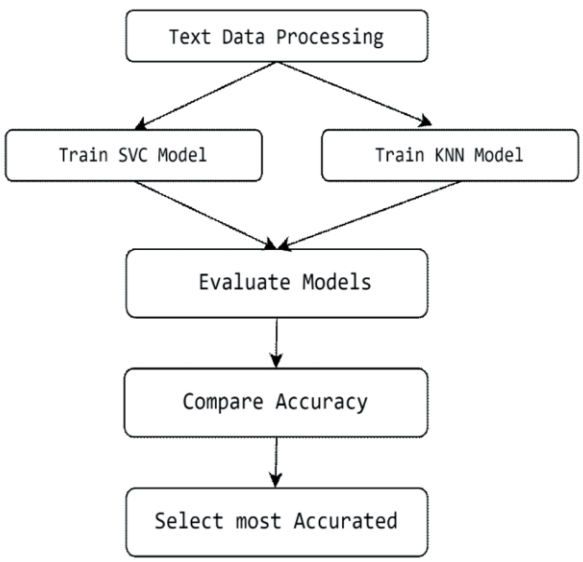


Figure 5. Working Model of Training Module

into training and testing sets to train and evaluate the models' performance. The Support Vector Classifier (SVC) is chosen as the main machine learning model for its effectiveness in text classification tasks, trained using a pipeline with TF-IDF vectorization. Predictions are made on the test data, and accuracy scores are calculated to assess the model's performance. Additionally, the K-Nearest Neighbors (KNN) classifier is trained for comparison. It is crucial to note that while the Support Vector Classifier (SVC) yielded the highest accuracy for this dataset, the most accurate model may vary based on the dataset characteristics. However, in this case, the SVC classifier is selected as the most accurate model for the resume screening system. Overall, it encompasses essential steps of text data preprocessing, model training, evaluation, and prediction, laying the foundation for an efficient resume screening system.

3.4 SVC

This segment of code encapsulates the implementation of a Support Vector Classifier (SVC) model, a sophisticated machine learning algorithm widely used for text classification tasks. Through the integration of the TF-IDF vectorization technique, raw textual data is seamlessly transformed into a numerical format, enabling comprehensive analysis and classification. Following meticulous training on the designated dataset, the SVC model is proficiently primed to discern patterns and accurately categorize text samples. Ultimately, the evaluation of the model's performance via the accuracy_score metric underscores its efficacy in achieving precise classification outcomes.

3.5 KNN

In a parallel exploration, the implementation of the k-Nearest Neighbors (KNN) classifier is undertaken, presenting an alternative avenue for text classification

endeavors. Commencing with the partitioning of the dataset into training and testing subsets, the KNN model undergoes training on the designated training data, where it encapsulates the inherent patterns latent within the textual corpus. Central to the KNN algorithm's methodology is its propensity to classify samples predicated on their proximity to neighboring data points within the feature space, a tenet upheld during the predictive phase. Subsequently, the accuracy_score metric, serving as a pivotal benchmark for model efficacy, elucidates the KNN classifier's prowess in furnishing precise classification outcomes, thereby augmenting the spectrum of classification methodologies scrutinized within this scholarly discourse.

When we analyse Figure 6, It is shown that SVC Classifier gives most accurate than KNN in our findings. So SVC Model is taken further proceedings.

3.6 Testing

The Testing module serves a pivotal role in validating the efficacy of the resume screening system by assessing its performance on real-world resume data in PDF format. This step is crucial for evaluating the system's capability to accurately classify resumes into relevant job categories, ensuring its practical reliability in recruitment processes. As shown in Figure 7, the resume data is extracted from PDF files using appropriate libraries, followed by

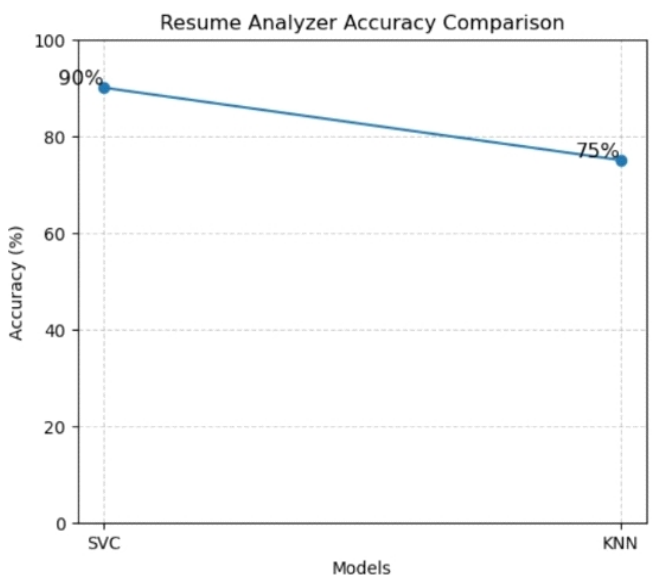


Figure 6. Comparison of SVC and KNN Accuracy

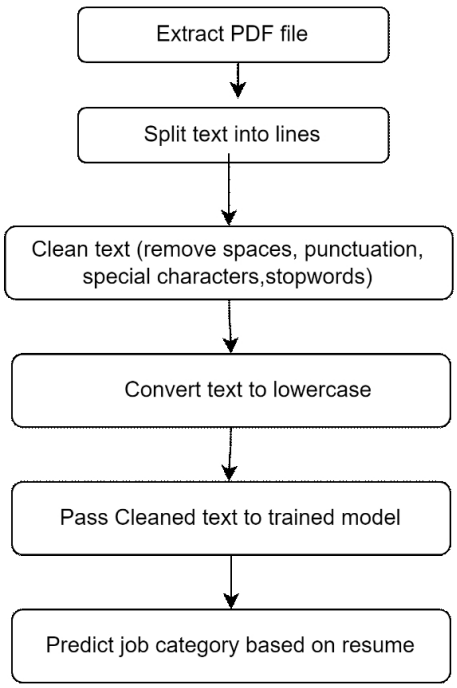


Figure 7. Working Model of Testing

preprocessing steps to ensure text cleanliness and standardization (Figure 8). Subsequently, the cleaned text undergoes further processing to remove irrelevant elements such as punctuation and stopwords. The resulting cleaned text is then subjected to model prediction, where the system utilizes the resume content to determine the appropriate job category. In Figure 9, the predicted category is presented for immediate evaluation of the system's performance in resume classification.

Conclusion

In summary, this research endeavor has delved into the development and evaluation of a resume screening system leveraging advanced machine learning techniques. Through the meticulous exploration of dataset structures, preprocessing methodologies, and model implementations, significant strides have been made towards automating the recruitment process. The adoption of Support Vector Classifier (SVC) and k-Nearest Neighbors (KNN) algorithms has demonstrated promising outcomes in classifying resumes into relevant job categories, with SVC exhibiting particularly notable accuracy.

```
MUHAMMAD SAVAD N muhammadsavadn muhammadsavadn007@gmail.com 7510942572 Objective A dedicated and dynamic recent graduate
with a fervent passion for teaching, backed by valuable part-time teaching experience during my academic journey. I am ea
ger to ignite students enthusiasm for learning, foster their growth, and actively contribute to educational research
endeavors. Achievements • Qualified UGC NET in Computer Science & Applications Special Skills • Public speaking • Mentoring •
C,C++,Python • SQL Leadership Learning Strategies Researcher HTML Relevant Experience TEACHER(PART TIME) DYUTHI CENTRE FOR EXCE
LLENCE | 2023-PRESENT TUTOR/MENTOR(PART TIME) EDUCATOR INDIVIDUAL CONCEPT| 2020-2023 JRF/NET trainer in Computer Science One-to
-one mentoring for school students Education Background NILGIRI COLLEGE OF ARTS AND SCIENCE,THALOOR SULLAMUSSALAM SCENCE COLLEG
E,AREAKODE MSC CS | BHARATHAR UNIVERSITY | 2022- PRESENT BSC CS | CALICUT UNIVERSITY | 2019-2022 Maintained 90% mark till now(2
nd sem) Teaching Assistant Student Coordinator, IIC FDP programs for School teachers Techtalks,Sessions and outreach programs S
ecured 7.933 CGPA Quality & Operations Lead ,IEDC Mentor, SALT by Kerala Startup Mission NSS Volunteer HMYHSS ,MANJERI PKMIC HI
GH SCHOOL,POOKKOTTUR HSE | KERALA STATE BOARD | 2017- 2019 SSLC | KERALA STATE BOARD | 2017 Secured 89.5% Mark Secured A+ in al
l subjects Publications Toxic Comment Classifier on Socail Media  Plaform i manager's journal on computer science-Mar 15,2023 C
itation in Google Scholar Toxic Comment Classifier International Conference on Advanced Computing(ICAC-2023) conducted by Bhara
thiar University-Mar 03,2023 Certifications NET in Computer Science & Applications(June,2023) Python for Data Science from IIT
Madras Big Data Foundations Data Science Foundations Project Management Fundamentals UGC NPTEL IBM Project Toxic Comment Classi
fier Analysing and classifying the text as the amount toxic present in the sentence by using the methods of natural language pr
ocessing and other machine language techniques Disaster Analysis and Response Through Social Media: Automated Classification An
d Location Based Alerts Analysing and checking if a comment is based on disaster or not and giving alerts based on the location
in text. Contact Information Address: Chereekode House Pallimukku,Pookkottur Malappuram,676517 Cell Number: 7510942572 Email: m
uhammadsavadn007@gmail.com
```

Figure 8. Extracted Resume

```
In [43]: cleaned_text = text.lower()
         cleaned_text = cleanResume(cleaned_text)
         print([cleaned_text])

         ['muhammad savad n muhammadsavadn muhammadsavadn007 7510942572 objective a dedicated and dynamic recent graduate with a fervent
         passion for teaching backed by valuable part time teaching experience during my academic journey i am eager to ignite students
         enthusiasm for learning foster their growth and actively contribute to educational research endeavors achievements qualified ug
         c net in computer science applications special skills public speaking mentoring c c python sql leadership learning strategies r
         esearcher html relevant experience teacher part time dyuthi centre for excellence 2023 present tutor mentor part time educator
         individual concept 2020 2023 jrf net trainer in computer science one to one mentoring for school students education background
         nilgiri college of arts and science thaloor sullamussalam scence college areakode msc cs bharathar university 2022 present bsc
         cs calicut university 2019 2022 maintained 90 mark till now 2nd sem teaching assistant student coordinator iic fdp programs for
         school teachers techtalks sessions and outreach programs secured 7 933 cgpa quality operations lead iedc mentor salt by kerala
         startup mission nss volunteer hmyhss manjeri pkmic high school pookkottur hse kerala state board 2017 2019 sslc kerala state bo
         ard 2017 secured 89 5 mark secured a in all subjects publications toxic comment classifier on socail media plaform i manager s
         journal on computer science mar 15 2023 citation in google scholar toxic comment classifier international conference on advance
         d computing icac 2023 conducted by bharathiar university mar 03 2023 certifications net in computer science applications june 2
         023 python for data science from iit madras big data foundations data science foundations project management fundamentals ugc n
         ptel ibm project toxic comment classifier analysing and classifying the text as the amount toxic present in the sentence by usi
         ng the methods of natural language processing and other machine language techniques disaster analysis and response through soci
         al media automated classification and location based alerts analysing and checking if a comment is based on disaster or not and
         giving alerts based on the location in text contact information address chereekode house pallimukku pookkottur malappuram 67651
         7 cell number 7510942572 email muhammadsavadn007 ']

In [44]: predicted=final.predict([cleaned_text])
         print(predicted)
         predict = ''.join(predicted)
         print(predict)

         ['Data Science']
         Data Science
```

Figure 9. Cleaned Resume and its Predicted Category of Job

Future Work

Moving forward, several avenues for future research and enhancement of the proposed resume screening system are envisioned. Firstly, the integration of additional features such as semantic analysis and contextual understanding could enrich the model's decision-making capabilities, potentially leading to more nuanced and accurate classifications. Furthermore, the incorporation of user feedback mechanisms and iterative model refinement processes could contribute to the system's adaptability and responsiveness to evolving recruitment dynamics (Fazel-Zarandi & Fox, 2009). Additionally, exploring ensemble learning techniques and deep learning architectures could unlock further potential in optimizing classification performance and scalability. Lastly, expanding the scope of evaluation to include larger and more diverse datasets, along with real-world deployment scenarios, would provide invaluable insights into the system's robustness and practical utility in

addressing real-world recruitment challenges.

References

[1]. Amin, S., Jayakar, N., Sunny, S., Babu, P., Kiruthika, M., & Gurjar, A. (2019, January). Web application for screening resume. In 2019 *International Conference on Nascent Technologies in Engineering (ICNTE)* (pp. 1-7). IEEE. https://doi.org/10.1109/ICNTE44896.2019.8945869

[2]. Bachate, S., Kakade, P., Sugandhi, H., & Jambhulkar. P. (2023). Empowering candidates with experience sharing and advanced resume screening - A comprehensive survey. *International Journal for Research in Applied Science and Engineering Technology,* 11, 1864-1870. https://doi.org/10.22214/ijraset.2023.56347

[3]. Chen, J., Niu, Z., & Fu, H. (2015). A novel knowledge extraction framework for resumes based on text classifier. In *Web-Age Information Management: 16th International Conference, WAIM 2015, Qingdao, China, June 8-10, 2015. Proceedings 16* (pp. 540-543). Springer International Publishing. https://doi.org/10.1007/978-3-319-21042-1_58

[4]. Fazel-Zarandi, M., & Fox, M. S. (2009, October). Semantic matchmaking for job recruitment: An ontology-based hybrid approach. In *Proceedings of the 8th International Semantic Web Conference,* 525 (1), 1-14.

[5]. Kmail, A. B., Maree, M., & Belkhatir, M. (2015, August). MatchingSem: Online recruitment system based on multiple semantic resources. In 2015 *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 2654-2659). IEEE. https://doi.org/10.1109/FSKD.2015.7382376

[6]. Kmail, A. B., Maree, M., Belkhatir, M., & Alhashmi, S. M. (2015, November). An automatic online recruitment system based on exploiting multiple semantic resources and concept-relatedness measures. In 2015 *IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 620-627). IEEE. https://doi.org/10.1109/ICTAI.2015.95

[7]. Priyanka, J. H., & Parveen, N. (2023). DeepSkillNER: An automatic screening and ranking of resumes using hybrid deep learning and enhanced spectral clustering approach. *Multimedia Tools and Applications* (pp. 1-28). https://doi.org/10.1007/s11042-023-17264-y

[8]. Schmitt, T., Caillou, P., & Sebag, M. (2016, September). Matching jobs and resumes: A deep collaborative filtering task. In *GCAI 2016-2nd Global Conference on Artificial Intelligence,* 41, 1-13.

[9]. Senthil Kumaran, V., & Sankar, A. (2013). Towards an automated system for intelligent screening of candidates for recruitment using ontology mapping (EXPERT). *International Journal of Metadata, Semantics and Ontologies,* 8(1), 56-64. https://doi.org/10.1504/IJMSO.2013.054184

[10]. Surendiran, B., Paturu, T., Chirumamilla, H. V., & Reddy, M. N. R. (2023, May). Resume classification using ML techniques. In 2023 *International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)* (pp. 1-5). IEEE. https://doi.org/10.1109/IConSCEPT57958.2023.10169907

[11]. Thatha, V. N., VeerasekharReddy, B., VenuGopal, G., Ashok, K., & Maddu, S. (2023, November). An enhanced support vector machine based pattern classification method for text classification in English texts. In 2023 *7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1-6). IEEE. https://doi.org/10.1109/CSITSS60515. 2023.10334170

[12]. VeeraSekharReddy, B., Rao, K. S., & Koppula, N. (2022, December). Named entity recognition using CRF with active learning algorithm in English texts. In 2022 *6th International Conference on Electronics, Communication and Aerospace Technology* (pp. 1041-1044). IEEE. https://doi.org/10.1109/ICECA55336.2022.10009592

[13]. Zaroor, A., Maree, M., & Sabha, M. (2018). A hybrid approach to conceptual classification and ranking of resumes and their corresponding job posts. In *Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies (KES-IDT 2017)–Part I 9* (pp. 107-119). Springer International Publishing. https://doi.org/10.1007/978-3-319-59421-7_10

ABOUT THE AUTHORS

Muhammad Savad N., Department of Computer Science, Nilgiri College of Arts and Science (Autonomous), Thaloor, Nilgiri, Tamil Nadu, India.

T. Preethi, Department of Multimedia & Web Technology, Nilgiri College of Arts and Science(Autonomous), Thaloor, Nilgiri, Tamil Nadu, India.

i-manager's Journal on
Computer Science (JCOM)

Published by i-manager Publications, India.

http://www.imanagerpublications.com

Editor-in-Chief: Dr. Kamal Kumar Mehta
Professor and Head of Computer Engineering,
MPSTME NMIMS, Shirpur,
Maharashtra, India.

Call For Papers for 2024

i-manager Publications is a leading publishing house specialized in publishing Scientific, Technology, Education and Management Journals. The credibility and impact of these publications in the Indian economic and academic environment reinforced the founding vision of i-manager Publications: promoting technology education and present international research worldwide.

i-manager's Journal on Computer Science deals with all aspects of computer science and contributes theoretical results and offers a compilation of high quality articles to encompass a wide spectrum of advancements in the actively developed domain. The Journal serves a unique purpose through multi-disciplined topics and provides innovative ideas with a view to face new challenges of the current and future centuries. i-manager's Journal on Computer Science covers a great deal of what has been done in the field recently and intends to bring together the most recent advances and applications in all branches of the academic computer science community with new knowledge and technology for the benefit of students, professionals and industrial practitioners.

Why Publish with us?

- No Publishing Fee
- Abstracting & Indexing in leading databases
- Double-blind Peer Review
- Highly qualified Editorial Board
- Publishing Journals since 2004
- Maximum Publicity in Social media
- Rapid Publication: 2-3 months
- Full archive available from Volume 1 onwards

Abstracting / Indexing

ProQuest · ULRICHSWEB Global Serials Directory · OCLC WorldCat · EBSCO HOST · Google Scholar · Scilit · Crossref · ACNP Catalogo Italiano dei Periodici · ZDB Zeitschriften Datenbank · dimensions.ai · wizdom.ai · ScienceGate

Submission email : submissions@imanagerpublications.com

i-manager's Journal on
Computer Science (JCOM)

Overall Topics Covered:

- Communication architectures for pervasive computing
- Evolutionary computing and intelligent systems
- User Modeling and User Adapted Interaction
- Computer Aided Geometric Design
- Integrated Computer Aided Engineering
- Computer Vision And Image Understanding
- Cryptographic Hardware and Embedded Systems
- Advances in Intrusion Detection
- Pervasive and Ubiquitous Computing
- Digital logic and Data Structures
- Recent Trends in Cloud Data Processing
- Human-Computer Interaction with Mobile Devices and Services
- Robotics and Computer Aided Manufacturing
- Computer Methods in Biomechanics and Biomedical Engineering
- Interactive 3D Graphics and Games
- Green Computing
- Temporal and spatial Data bases
- Neural networks
- Automated reasoning with Analytic Tableaux and Related Methods
- Intelligent tutoring systems
- Natural Language Processing
- Aspect oriented software development
- Context based Information Retrieval
- Expert systems with applications
- Image and Voice Processing
- Ad-hoc networks
- Future Generation Computer Systems
- Autonomous agents and Multi-agent systems
- Nature of Computation (Logic, Algorithms, Applications)
- Real-time Processing
- Cybernetics
- Ubiquitous Computing

Features

Articles, Research Papers, Review Papers.

Periodicity

January - March, April - June, July - September, October - December

Target Audience

Academicians, practitioners and post-graduate students in the field of Computer Science, University Educational Bodies, Researchers, etc.

i-manager® Publications

https://www.facebook.com/imanJCOM/

https://scholar.google.co.in/citations?user=qdxsNOsAAAAJ&hl=en

https://twitter.com/imanagerpub

Submission email : submissions@imanagerpublications.com

i-manager® Publications

3/343, Hill view, Town Railway Nager, Nagercoil
Kanyakumari Dist. Pin-629 001.
Tel: +91-4652-231675, 232675, 276675

e-mail: info@imanagerpublications.com
contact@imanagerpublications.com
www.imanagerpublications.com

i-manager Printing Press