FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 11, November, 2025, pp 435 – 440
DOI: https://doi.org/10.33003/fjs-2025-0911-3944

PHISHING DETECTION IN SOCIAL MEDIA PLATFORMS USING MACHINE LEARNING

Shuaibu Tatu Haruna, Sunday Eric Adewumi and Victoria Ifeoluwa Yemi-Peters

Department of Computer science, Federal University Lokoja, Nigeria.

*Corresponding authors' email: dagasta2004@gmail.com.

ABSTRACT

A Phishing attack is the process where online fraudsters gain unwanted access to internet users' private information through phishing links. the continuous advancement in technology has lead increase and more sophisticated phishing attacks. phishing attacks continue to pose serious threats to internet users, particularly social media users. different studies have focused on the detection on phishing attempt, but have focused on url-based, attributes, neglecting the possibility of contextual indicators in textual messages.to address these gaps, this study develops a message-based phishing detection model that focused on features such as message length, link presence, and urgency cues. this study used a combination of dataset collected from a nigerian online forum (nairaland), and kaggle phishing dataset, which provides a balanced and diverse sample of both legitimate and phishing messages. the data was preprocessed and used to train three different algorithms which includes random forest (rf), support vector machine (svm), and deep neural network (dnn), and was evaluated using standard evaluation matrices including accuracy, precision, recall, and f1-score. the results from the models evolution shows that the dnn model outperforms all the compared models achieving 99% across all evolution matrices used. this study was further compared with the baseline papers, and the proposed dnn model was able to outperform the baseline model which focused on url-based phishing detection. future work can focused on real-time deployment of the model and integration with mobile messaging systems for proactive defense.

Keywords: Phishing, Detection, Cyber security, Machine learning, Deep learning

INTRODUCTION

Phishing is the process by which cyber criminals gain access to private information of internet users' through some malicious links called phishing link (Van-Geest et al., 2024). Cyber criminals use social engineering techniques to carryout phishing attacks to trick social media users into disclosing their personal information. Over the year phishing attacks has targeted email users, but the recent emergence of social media platforms like Twitter Facebook, Telegram, Instagram, and Tic-tok has provided other avenue for cybercriminals to carryout phishing attacks (Sarpong et al., 2022). Social media large number of users and instant messages feature has made it susceptible to phishing attempts. Little awareness has been made on phishing attacks on social media, making the users less careful, and easy prey for phishing attacks, which takes advantage of social trust (Safi& Singh, 2023). Detection of phishing attack early is important to prevent attacks. Accurate and reliable phishing detection system designed for social media platform is needed to fight against phishing attacks on social media.

Machine learning is a branch of Artificial Intelligence (AI), which has the ability to learn from historical data, and make prediction, and can be used to identify trends and anomalies that suggest phishing attempts (Basit et al., 2021). Recent advancement of marching learning techniques has made more efficient in the detection of phishing attempt. Recent study on phishing detection has shown the ability of machine learning models such as random forest, decision trees, support vector machine and deep learning models like deep neural networks and artificial neural networks in the detection of phishing attempts.

While studies on phishing detection have focused on email, websites and Internet of Things (IoT) devices (Tabassum et al., 2021), fewer studies have focused on message-based phishing detection on social media platforms, which features rapidly changing content and distinct user behaviors. Also, no study have compared the performance of both deep learning and machine learning approach to know which approach performs best on the detection of phishing attacks. This research tries to address these gaps by proposing an efficient phishing detection method that compared machine learning and deep learning algorithms on social media dataset. This study will utilized the Support Vector Machines (SVM), Random Forest, and a Deep Neural Network (DNN) algorithms.

MATERIALS AND METHODS

Literature Review

Different studies have utilized the machine learning and deep learning approach for phishing detection on social medial. Some of these studies are reviewed in this section.

Sakhare et al. (2024) focused on the detection of phishing attacks by utilizing a Graph Neural Network (GNN) with different machine learning algorithms which includes extreme gradient boosting (XGBoost), LightGBM, NB, and CatBoost. The phishing detection model was train with the use of a URL based dataset, considering URL features like the length, dots, numbers, and special characters. The experimental result shows that the proposed GNN model outperformed all other compared algorithms achieving more than 90% accuracy.

Abed et al. (2023) conducted a research that on combating evolving phishing attacks. The study presenteda phishing detection model that was built based on URLs features. The study made use of multiple machine learning algorithms. The result from the study shows that the proposed models achieved good result, with accuracies up to 88.3%.

Narayana et al. (2023) carried out a research on the detection of phishing attacks. The study utilized the RF algorithm on the detection of phishing attacks.The study used the wider dataset which consist of HTML, JavaScript, URLs, and domain-based attributes. The experimental result shows that the proposed RF model was able to achieve over 98% accuracy and over 98% recalls.

Kaibassova et al. (2022) focused on the detection of phishing websites with the use of ensemble learning approach. The study utilized the AdaBoost, CatBoost, and Gradient Boosting Classifier algorithms. Data preprocessing and exploratory data analysis was carried out on a URL based dataset. The experimental result shows that the AdaBoost outperform other algorithm with an average ROC AUC score of 99%.

A research that focused on a the identification of phishing URL was carried out by Thakur et al. (2022). The study utilized three deep learning algorithms which are the Convolution Neural Networks (CNN), Long-Term-Short-Term memory (LSTM), and Deep Neural Network (DNN). The result from the study shows that the DNN was able to outperform all other compared algorithms, achieving 96.95% accuracy, 99% precision, 100% recall, and 99% F1-score.

Putra et al. (2022) carried out a research that focused on bot attacks. The study utilized a techniquesthat identifies patterns in bot communications. The study utilized machine learning models which includes Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF), and Naïve Bayes (NB). The result from the experiment shows that the KNN was able to achieve the best result, getting the highest accuracy OF 99.9%.

A research that focused of phishing attempt detection was carried out by Kim (2024). The study adopted the machine learning approach, using RF and LR to build the phishing detection model. The dataset used to train the model was collected from DataCo Global. The result from the study shows that the proposed model was able to achieve 94.71% accuracy, 95.74 precision, 95.23 % recall, 95.15% f1_score.

Al-Ruwili&Mostafa (2023) carries out a research on ransomware detection. The study utilized several machine learning algorithm which include NB, DT, and SVM. The algorithms were trained on the SEL datasets. The experimental result shows that the SVM outperform all other compared algorithm, achieving 99.31% accuracy on the SEL dataset.

Adane & Beyene (2023) conducted a research that applied machine learning approach on the detection of website-based phishing attacks. The study used the Cat-Boost Classifier and other machine learning algorithms. The study applied Uni-variate Feature Selection method, which achieved 97.46% accuracy rate with the Cat-Boost Classifier.

Almalki &Abdelmajeed (2023) carried out a research of website-based phishing detection. The study used the neural network algorithm to build the model. The dataset used was a URL based dataset. The Neural Network was compared to other state-of-the-art algorithms, and was able to outperform the compared algorithm by achieving the highest accuracy which is 98.27%.

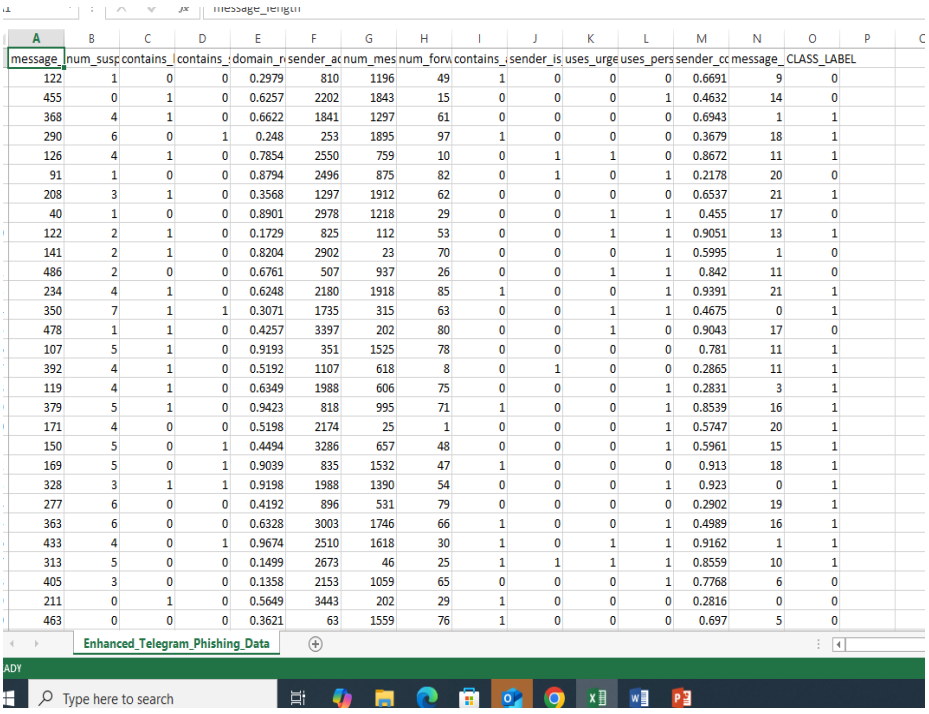
Based on the reviewed papers, we observed that previous studies on phishing detection have mainly focused on URL-based and website-based phishing detection models, and limited studies have focused on message-based methods, particularly social media messages. Also, most existing studies structured future from URL or web content, and fails to consider behavioral indicators and linguistic cues present in textual messages. Additionally, most existing studies often focus on traditional feature selection methods, ignoring the potential of feature correlation analysis for improving model robustness.These study aim to address these gaps by developing a message-based phishing detection model that compare deep learning and machine learning approaches with a focus of correlation-driven analysis to improve model performance and robustness.

Methodology

The method that was adopted in this research follows a unique pattern similar to the CRISP DM methodology. This method has different phases which include:

Data Collection

The dataset used in this study was collected from two different sources. The first dataset was collected from a social media platform (Nairaland), and the second dataset was collected from Kaggle. The two dataset were combined together and used to achieve the aim of this study. The combination of these dataset from different sources offers a diverse representative dataset containing both legitimate and phishing messages. Figure 1 shows a snapshot of the combined dataset.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
message_	num_susp	contains_	contains_	domain_r	sender_ac	num_mes	num_forw	contains_	sender_is	uses_urge	uses_pers	sender_cc	message_	CLASS_LABEL
122	1	0	0	0.2979	810	1196	49	1	0	0	0	0.6691	9	0
455	0	1	0	0.6257	2202	1843	15	0	0	0	1	0.4632	14	0
368	4	1	0	0.6622	1841	1297	61	0	0	0	0	0.6943	1	1
290	6	0	1	0.248	253	1895	97	1	0	0	0	0.3679	18	1
126	4	1	0	0.7854	2550	759	10	0	1	1	0	0.8672	11	1
91	1	0	0	0.8794	2496	875	82	0	1	0	1	0.2178	20	0
208	3	1	0	0.3568	1297	1912	62	0	0	0	0	0.6537	21	1
40	1	0	0	0.8901	2978	1218	29	0	0	1	1	0.455	17	0
122	2	1	0	0.1729	825	112	53	0	0	1	1	0.9051	13	1
141	2	1	0	0.8204	2902	23	70	0	0	0	1	0.5995	1	0
486	2	0	0	0.6761	507	937	26	0	0	1	1	0.842	11	0
234	4	1	0	0.6248	2180	1918	85	1	0	0	1	0.9391	21	1
350	7	1	1	0.3071	1735	315	63	0	0	1	1	0.4675	0	1
478	1	1	0	0.4257	3397	202	80	0	0	1	0	0.9043	17	0
107	5	1	0	0.9193	351	1525	78	0	0	0	0	0.781	11	1
392	4	1	0	0.5192	1107	618	8	0	1	0	0	0.2865	11	1
119	4	1	0	0.6349	1988	606	75	0	0	0	1	0.2831	3	1
379	5	1	0	0.9423	818	995	71	1	0	0	1	0.8539	16	1
171	4	0	0	0.5198	2174	25	1	0	0	0	1	0.5747	20	1
150	5	0	1	0.4494	3286	657	48	0	0	0	1	0.5961	15	1
169	5	0	1	0.9039	835	1532	47	1	0	0	0	0.913	18	1
328	3	1	1	0.9198	1988	1390	54	0	0	0	1	0.923	0	1
277	6	0	0	0.4192	896	531	79	0	0	0	0	0.2902	19	1
363	6	0	0	0.6328	3003	1746	66	1	0	0	1	0.4989	16	1
433	4	0	1	0.9674	2510	1618	30	1	0	1	1	0.9162	1	1
313	5	0	0	0.1499	2673	46	25	1	1	1	1	0.8559	10	1
405	3	0	0	0.1358	2153	1059	65	0	0	0	1	0.7768	6	0
211	0	1	0	0.5649	3443	202	29	1	0	0	0	0.2816	0	0
463	0	0	0	0.3621	63	1559	76	1	0	0	0	0.697	5	0

Enhanced_Telegram_Phishing_Data

Figure 1: Dataset Structure

Data Preprocessing

After collection, we subjected the dataset to several preprocessing steps to prepare it for model training. This included cleaning, feature selection, and encoding categorical features. Feature correlation was checked to see feature that are highly correlated in the dataset. Figure 2 shows the correlation heatmap that was generated.

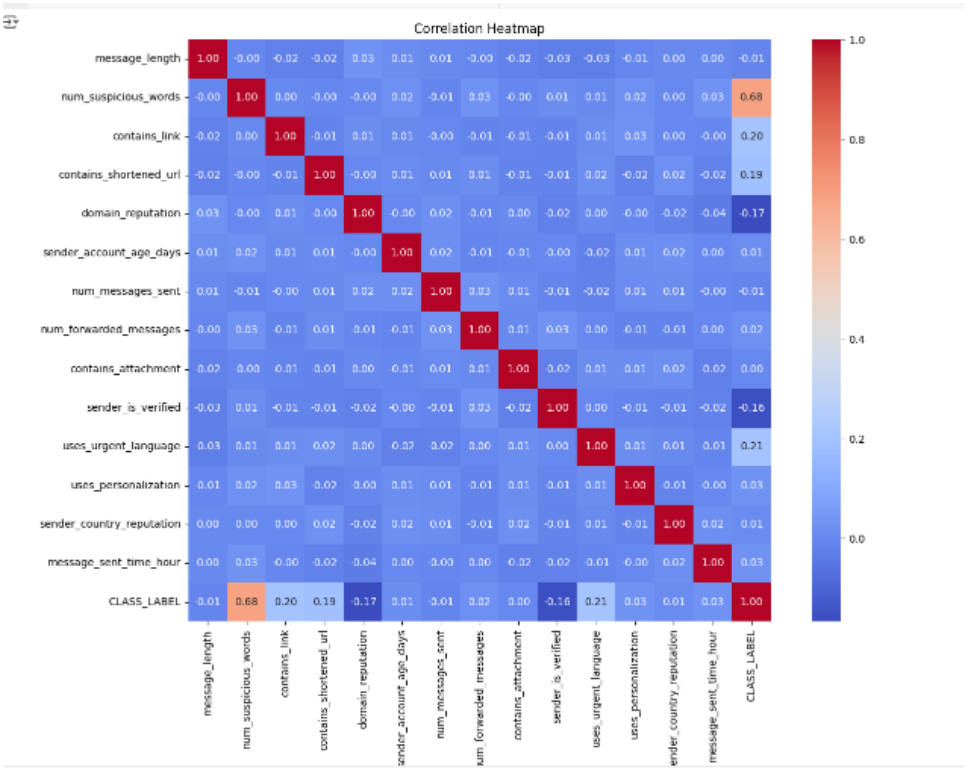


Figure 2: Correlation Heatmap

Model Building

After the data preprocessing has been done, the preprocessed data was further used to train the models which are RF, SVM, and a DNN. The DNN algorithm uses three dense layers with ReLU activation function, followed by a sigmoid output layer suitable for binary classification, and was trained over 20 epochs. The raining history is shown in Figure 3. Which demonstrate strong convergence and consistent learning.

```
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
125/125 ━━━━━━━━━━━━━━━━━━━━ 4s 15ms/step - accuracy: 0.6705 - loss: 0.6056 - val_accuracy: 0.9460 - val_loss: 0.2130
Epoch 2/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 2s 8ms/step - accuracy: 0.9645 - loss: 0.1696 - val_accuracy: 0.9800 - val_loss: 0.0932
Epoch 3/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 7ms/step - accuracy: 0.9821 - loss: 0.0835 - val_accuracy: 0.9830 - val_loss: 0.0598
Epoch 4/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 5ms/step - accuracy: 0.9890 - loss: 0.0503 - val_accuracy: 0.9880 - val_loss: 0.0441
Epoch 5/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 5ms/step - accuracy: 0.9934 - loss: 0.0370 - val_accuracy: 0.9890 - val_loss: 0.0356
Epoch 6/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9971 - loss: 0.0240 - val_accuracy: 0.9880 - val_loss: 0.0319
Epoch 7/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9971 - loss: 0.0202 - val_accuracy: 0.9910 - val_loss: 0.0272
Epoch 8/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.9978 - loss: 0.0172 - val_accuracy: 0.9900 - val_loss: 0.0271
Epoch 9/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9983 - loss: 0.0126 - val_accuracy: 0.9890 - val_loss: 0.0254
Epoch 10/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9991 - loss: 0.0118 - val_accuracy: 0.9900 - val_loss: 0.0239
Epoch 11/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 0.0079 - val_accuracy: 0.9860 - val_loss: 0.0253
Epoch 12/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 0.9996 - loss: 0.0067 - val_accuracy: 0.9880 - val_loss: 0.0234
Epoch 13/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 0.0062 - val_accuracy: 0.9870 - val_loss: 0.0243
Epoch 14/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 1.0000 - loss: 0.0057 - val_accuracy: 0.9890 - val_loss: 0.0231
Epoch 15/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 0.0042 - val_accuracy: 0.9930 - val_loss: 0.0216
Epoch 16/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - accuracy: 1.0000 - loss: 0.0034 - val_accuracy: 0.9910 - val_loss: 0.0206
Epoch 17/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 1.0000 - loss: 0.0034 - val_accuracy: 0.9920 - val_loss: 0.0237
Epoch 18/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 1.0000 - loss: 0.0028 - val_accuracy: 0.9910 - val_loss: 0.0240
Epoch 19/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 1.0000 - loss: 0.0023 - val_accuracy: 0.9900 - val_loss: 0.0212
Epoch 20/20
125/125 ━━━━━━━━━━━━━━━━━━━━ 1s 4ms/step - accuracy: 1.0000 - loss: 0.0019 - val_accuracy: 0.9900 - val_loss: 0.0221
32/32 ━━━━━━━━━━━━━━━━━━━━ 0s 4ms/step
DNN Classification Report:
```

Figure 3: DNN Training History for 20 Epochs

Model Evaluation

The built models were further evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score. Each of the models was evaluated on the text dataset which is 20% of the dataset.

Experimental Results

This phishing detection models were trained and evaluated using standard evaluation matrices. The confusion matric for random forest, SVM, and DNN is presented inf Figure 4. Figure 5, and Figure 6 respectively while the summary of the result from the models evaluation is presented in Table 1 for clarity and comparison.

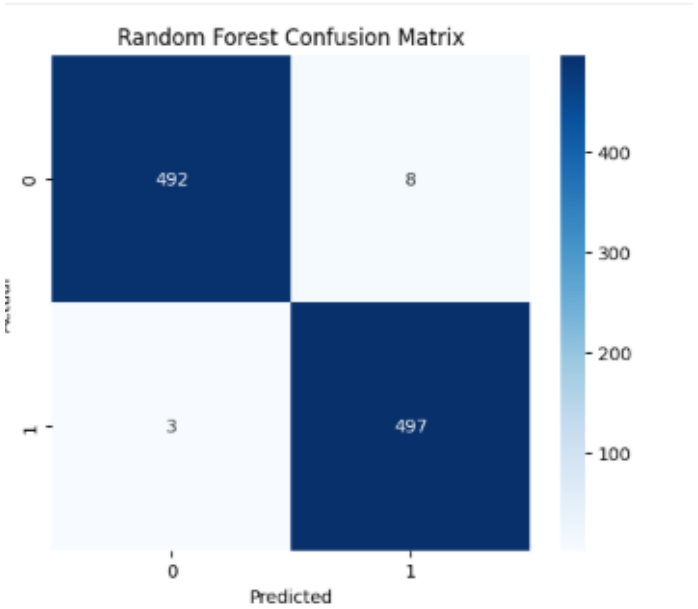


Figure 4: Random Forest Confusion Matrix

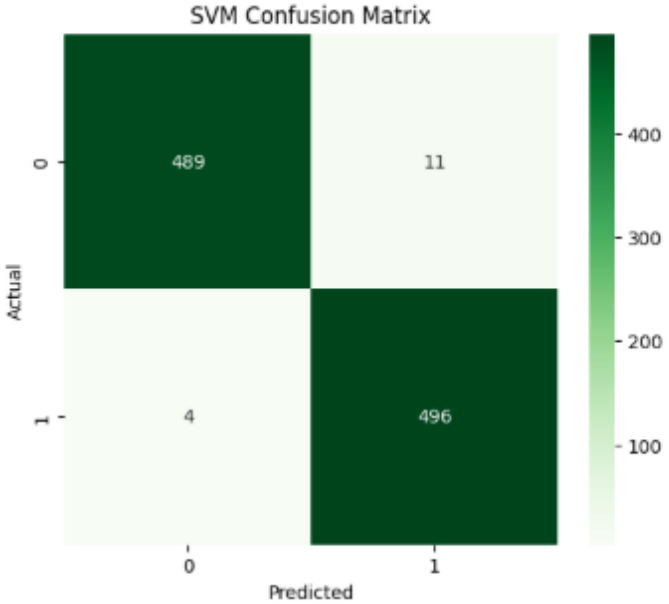


Figure 5: SVM Confusion Matrix

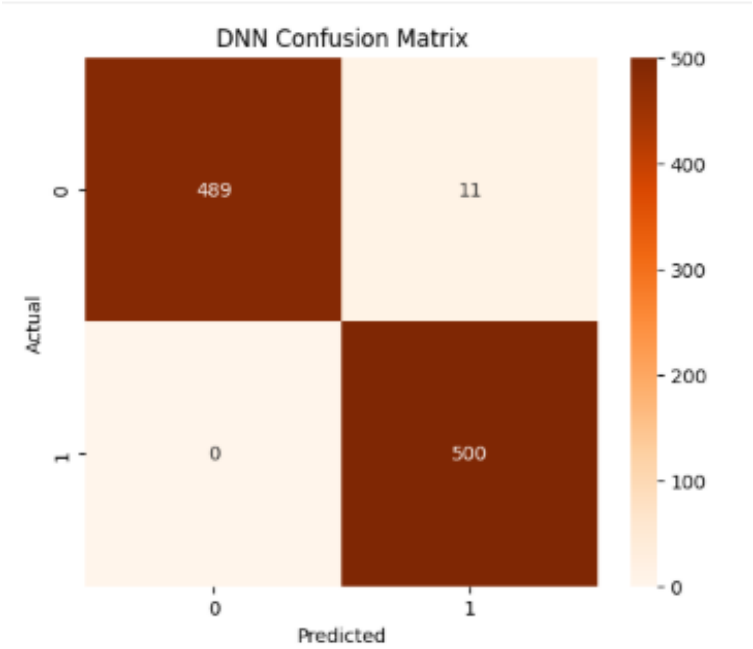


Figure 6: DNN Confusion Matrix

Table 1: Classification Performance Metrics for Each Model

Model	Accuracy	Precision	Recall	F1-Score
RF	0.99	0.98	0.99	0.99
SVM	0.98	0.98	0.98	0.99
DNN	0.99	0.99	0.99	0.99

From the table, we observed that the RF, SVM, and DNN model perform excellently well, showing machine learning and deep learning ability to detect phishing attacks. The deep learning based model slightly outperforms the machine learning-based models across all evaluation matrices. While the SVM achieves lower precision, recall, and F1-score of 98% each, the RF has accuracy, recall, and F1-score of 99% each but a slightly lower precision of 98%, showing its superiority over the SVM in phishing message detection. DNN achieves 99% across all evaluation metrics, indicating deep learning effectiveness and strong capability in the identification of phishing messages.

Comparative Analysis

This result of this study was compared with that of the baseline paper study by Kwon et al. (2023), which focused on the detection of malicious URLs with the uses of Neural Networks (NN). Known et al. (2021) utilized dataset collected from Whois and Alexa's top 1 million domains. The experimental result from Known et al (2021) study shows that the proposed NN model build on built on TensorFlow achieved 92.94%F1-score and 97.8% accuracy. On the other hands, this study's result shows the improved performance of the proposed models, particularly the DNN which outperforms all other compared algorithms and the baseline papers, achieving 99% across all evaluation matrices used. The improvement maybe due to the enhance feature engineering used in this study (integration of behavioral attributes like message length, link presence), the domain specific dataset used or the correlation-based feature analysis. Also, while the baseline paper focused URL-based phishing attacks detection, this study extend it focus to message-based phishing attack detection, which provides a better and more comprehensive approach that reflects real-world phishing tactics, involving urgent language cues in messaging apps. These finding shows that targeted dataset curation and context-aware feature engineering can improve the performance of phishing detection models across different platforms

CONCLUSION

This study focused on the development of a message-based phishing attempt detection on social media platforms. This study utilized a deep learning (DNN) and two machine learning algorithms (RF, and SVM) to train dataset collected from two source which are Nairaland (a Nigerian social platform), and Kaggle, this combinations of two dataset is to ensure diversity and real-world relevance of the phishing and legitimate messages. This study adopt a methodology inspired by the CRISP-DM methodology, this method has four different phases which include the data collection, data preprocessing, feature engineering, model training, and model evaluation. This study lay emphasis on correlation-based feature analysis to identify highly correlated feature and remove redundancy to enhance training efficiency. The three model were evaluated using standard evaluation metrics such as accuracy, precision, recall, and F1-score. The model achieve excellent result, with high performance with not less than 98% across all evaluation matrices used, with the DNN achieving the best result of 99% across all the evaluation matrices used. The result from this study was compared with that of the baseline paper carried out by Kwon et al. (2023). While the baseline paper achieves 97.8% accuracy and 92.94% F1-score on a URL-based phishing detection, this study's DNN achieves 99% across all evaluation metrics, outperforming the baseline paper. This study's result shows that combining domain-specific datasets, message-based analysis, and correlation-based behavioral feature selection

can significantly improve phishing detection performance. This study contributes to the war against phishing attacks and improvement of cyber security by shifting phishing detection focus from traditional URL-based detection methodsto a modern message-based detection method where phishing is increasingly prevalent. Future work can focused on real-time deployment of the model and integration with mobile messaging systems for proactive defense.

REFERENCES

Abed, L. H., Mohammed, H. J., & Yaseen, Y. S. (2023). Phishing identification through up-to-date features generation and exploration. *International Journal on Technical and Physical Problems of Engineering*, *15*(3).

Adane, K., & Beyene, B. (2023). Phishing Website Detection with and Without Proper Feature Selection Techniques: Machine Learning Approach. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 158). https://doi.org/10.1007/978-3-031-24475-9_61

Almalki, S., &Abdelmajeed, N. (2023). A New Intelligent Model for Phishing Web Sites Detection. *International Journal of Information & Digital Security*, *1*(1). https://doi.org/10.54878/66f0v378

Al-Ruwili, A. S. M., & Mostafa, A. M. (2023). Analysis of Ransomware Impact on Android Systems using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications*, *14*(11). https://doi.org/10.14569/IJACSA.2023.0141178

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., &Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. In *Telecommunication Systems* (Vol. 76, Issue 1). https://doi.org/10.1007/s11235-020-00733-2

Kaibassova, D., Nurtay, M., Tau, A., &Kissina, M. (2022). Solving the problem of detecting phishing websites using ensemble learning models. *Scientific Journal of Astana IT University*. https://doi.org/10.37943/12oyrs4391

Kim, B. H. (2024). Design of efficient phishing detection model using machine learning. *TehnickiGlasnik*, *18*(1). https://doi.org/10.31803/tg-20230219213151

Kwon, H., Park, S., & Kim, Y. (2021). Design of detection method for malicious URL based on Deep Neural Network. *Journal of Convergence for Information Technology*, *11*(5).

Narayana, G., Manchala, U. D., Naresh, U., Kiran, S., Kiran, M. A., & Ch, R. K. (2023). Improving Phishing Website Detection with Machine Learning: Revealing Hidden Patterns for Better Accuracy. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(9). https://doi.org/10.17762/ijritcc.v11i9.8353

Putra, M. A. R., Ahmad, T., &Hostiadi, D. P. (2022). Analysis of Botnet Attack Communication Pattern Behavior on Computer Networks. *International Journal of Intelligent Engineering and Systems*, *15*(4). https://doi.org/10.22266/ijies2022.0831.48

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, *35*(2). https://doi.org/10.1016/j.jksuci.2023.01.004

Sakhare, N. N., Bangare, J. L., Purandare, R. G., Wankhede, D. S., &Dehankar, P. (2024). Phishing Website Detection Using Advanced Machine Learning Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(12s).

SarpongA. K., KwasiAhiable, R., Kwame Appati, J., &Essel Mensah, E. (2022). Phishing Attacks in Social Engineering: A Review. *Journal of Cyber Security*, *4*(4). https://doi.org/10.32604/jcs.2023.041095

Tabassum, N., Neha, F. F., Hossain, M. S., & Narman, H. S. (2021). A Hybrid Machine Learning based Phishing Website Detection Technique through Dimensionality Reduction. *2021 IEEE International Black Sea Conference on Communications and Networking, BlackSeaCom 2021*. https://doi.org/10.1109/BlackSeaCom52164.2021.9527806

Thakur, I., Panda, K., & Kumar, S. (2022). Deep learning methods for malicious URL detection using embedding techniques as Logistic Regression with Lasso penalty and Random Forest. *PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing*. https://doi.org/10.1109/PDGC56933.2022.10053199

Van-Geest, R. J., Cascavilla, G., Hulstijn, J., & Zannone, N. (2024). The applicability of a hybrid framework for automated phishing detection. *Computers and Security*, *139*. https://doi.org/10.1016/j.cose.2024.103736