ISSN 2320 -2610
Volume 12, No.6, November – December 2023
International Journal of Multidisciplinary in Cryptology and Information Security
Available Online at http://www.warse.org/IJMCIS/static/pdf/file/ijmcis011262023.pdf
https://doi.org/10.30534/ijmcis/2023/011262023

An AI-Powered Approach to Real-Time Phishing Detection for Cybersecurity

Opeyemi Isaiah Enitan[1,2]
[1]Manchester Metropolitan University, United Kingdom, yenitan1@yahoo.com
[2]Covenant University, Ogun State, Nigeria, yenitan1@yahoo.com

Received Date : October 05, 2023 Accepted Date : November 22, 2023 Published Date : December 07, 2023

ABSTRACT

Email phishing is one of the most significant cybersecurity threats in the digital era, leading to financial losses and data breaches. This study presents an AI-based real-time phishing detection system that employs machine learning techniques such as Logistic Regression, Random Forest, Support Vector Machine (SVM), CatBoost, and XGBoost. The dataset, sourced from Kaggle, includes phishing and safe emails to train and evaluate these models. Preprocessing methods including text normalisation, stemming, and feature extraction were used to improve detection accuracy. The findings show that XGBoost and CatBoost had the highest accuracy of 98%, outperforming other models. Logistic Regression and Random Forest followed closely with 97% accuracy, while SVM had 96%. The findings emphasise AI's effectiveness in detecting phishing emails and the importance of continuous cybersecurity measures. This study enhances email security by demonstrating the effectiveness of AI-driven phishing prevention mechanisms.

Key words: Artificial Intelligence, Cybersecurity, Email Security, Machine Learning, Phishing Detection, Real-Time Detection

1. INTRODUCTION

Cybersecurity has taken on a central role in the rapidly evolving digital world due to the increased threat of cyberattacks. Email phishing attacks are among the most prevalent and dangerous types of cybercrime. Cybercriminals use deceitful strategies to deceive organisations and individuals into disclosing sensitive information or committing crimes. To bypass standard security measures, these attacks employ psychological approaches such as social engineering.

The escalating sophistication of cyber threats has shaped the landscape of cybersecurity. According to the Internet Security Threat Report by Symantec [11], cyber attackers are employing more advanced techniques and exploiting vulnerabilities at an alarming rate. This escalation has led to the proliferation of email phishing attacks, which often target users' psychological vulnerabilities, manipulating them into revealing sensitive information or performing harmful actions.

The importance of this issue is highlighted by studies such as the Verizon Data Breach Investigation Report [17], which states that email phishing is still one of the most prevalent and successful attack channels. Furthermore, the interconnection of current communication networks raises the possibility of successful phishing attacks, making them a top priority for cybersecurity experts.
Effective prevention measures are especially important because of the increasing number of high-profile data breaches caused by phishing attacks. Businesses must invest in advanced detection techniques while implementing digital transformation initiatives.

The creation of an AI-based system for real-time phishing detection is of utmost importance in tackling this rising threat. By harnessing artificial intelligence and machine learning, such a system can identify subtle patterns and anomalies that signal phishing attempts. This proactive approach aims to prevent attacks before they cause harm, offering a robust cybersecurity solution.

Phishing attacks pose a growing challenge to organisations and individuals, exploiting human vulnerabilities to compromise sensitive data. Existing rule-based systems have large false positive rates, frequently misclassifying genuine emails. This research addresses this gap by developing an AI-powered phishing detection system that enhances accuracy and adaptability to evolving threats. The objective is to use machine learning approaches to improve real-time detection, reduce false positives, and reinforce cyber security frameworks. This study aims to develop a more reliable and proactive defensive mechanism against phishing threats by utilising an advanced classification approach.

2. LITERATURE REVIEW

Various authors have explored the fields of Machine Learning (ML) and Artificial Intelligence (AI) to address the challenges in phishing detection.

[13] proposed a method to improve phishing email detection by incorporating semantics into highly accurate bag-of-words and part-of-speech techniques. Their study showed that concept-based models were less vulnerable to unseen phishing emails

compared to lexeme-based models. The research emphasised the need to consider semantic factors in phishing detection.

[6] introduced THEMIS, a phishing email detection methodology that models email headers, bodies, phrases, and characters using a recurrent convolutional neural network (RCNN) with an attention mechanism and multilevel vectors. Their tests demonstrated a false positive rate (FPR) of 0.043% and an overall accuracy of 99%, ensuring that legitimate emails were not mistakenly classified as phishing.

Another study by [5] developed an ML-based phishing attack detection model using Naïve Bayes (NB) and Decision Tree (DT) techniques. Their experiments, using infected emails from PhishTank, achieved a detection accuracy of 96%, confirming the potential of supervised ML techniques for phishing identification.

[2] proposed CNNPD, a deep learning-based framework for phishing email detection using Convolutional Neural Networks (CNNs). Unlike traditional methods that rely on manual feature extraction, CNNPD automates this process, reducing computational costs while improving detection accuracy. Their evaluation of PhishingCorpus and SpamAssassin datasets resulted in 99% accuracy and 98% precision.

The study by [18] introduced a DTOF-ANN model for phishing detection, improving the K-medoids clustering algorithm to remove duplicate data points and selecting optimal email features using a neural network classifier. Their results showed that DTOF-ANN outperformed traditional phishing detection models in recognising phishing attacks.

Many researchers have emphasised the importance of feature selection in phishing detection. [12] highlighted critical phishing indicators, including email sender reputation, email content analysis, linguistic patterns, and URL structures. ML models such as Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Deep Neural Networks (DNNs) have been extensively used for classification tasks.

[16] proposed an anti-phishing framework that applies Term Document Matrix (TDM) with Singular Value Decomposition (SVD) and Non-Negative Matrix Factorisation (NMF). Their framework achieved 88.8% accuracy, despite working with highly imbalanced datasets.

To address phishing URL detection, [1] designed a character-level CNN-based approach, allowing for efficient analysis of phishing URLs without needing to access website content. Their method resulted in 95% accuracy, demonstrating high reliability in real-world scenarios.

The research by [9] leveraged Recurrent Neural Networks (RNNs) to analyse ignored content sources, improving threat detection and adaptability. The model was tested on multiple datasets, including Enron emails and phishing archives, confirming its superiority over traditional phishing mitigation techniques.

The study by [10] examined how Large Language Models (LLMs), such as GPT-3.5 and GPT-4, could generate highly realistic spear-phishing emails. The findings highlighted the need for leadership intervention and defensive AI models to prevent the misuse of these technologies.

[8] proposed an Explainable AI (XAI) approach for phishing detection, comparing various ML models and finding that the Explainable Boosting Machine (EBM) offered the highest interpretability and accuracy. Their study emphasised the need for human-centred cybersecurity solutions.

Another study by [3] benchmarked AI-based phishing email detection systems (PEDS) against adversarial text-based perturbations. Their findings revealed that most existing PEDS are vulnerable to manipulation, stressing the importance of multi-layered phishing defence strategies.

Despite significant progress in phishing detection, several challenges remain. [4] highlighted that traditional rule-based email filtering systems suffer from high false positive rates, often misclassifying legitimate emails as phishing attempts. Their research stressed the need for adaptive AI models that evolve with emerging phishing tactics.

The imbalance in phishing datasets has also been a major issue. [15] noted that many phishing datasets contain significantly fewer phishing emails compared to legitimate ones, limiting the effectiveness of ML-based classifiers.

To tackle zero-day phishing threats, [14] designed an ML system that integrates machine translation identification and risk word detection. Their method achieved 92% accuracy in detecting known phishing attacks and 78% accuracy in identifying zero-day threats.

Another challenge involves phishing email interpretability. [7] explored feature engineering techniques to improve email classification transparency. Their approach achieved 98% accuracy for ham-spam datasets and 99% accuracy for ham-phishing datasets.

The literature review highlights the critical role of AI and ML in phishing email detection. While existing techniques such as CNNs, RNNs, Decision Trees, and Explainable AI models have proven effective, there is still room for improvement. Future research should focus on real-time phishing detection models, dynamic feature adaptation, and enhanced dataset diversity to strengthen AI-based email security solutions.

3. METHODOLOGY

The methodology employed for this study is based on a systematic approach to data collection and preparation for machine learning applications. The focus changes from dataset source to data preparation, which includes null value management, text normalisation to standardise content, and splitting for linguistic consistency. The critical procedure of splitting the data into training and testing subsets is completed,

allowing for the development of models on the training data and subsequent evaluation of hidden testing data. This division ensures the model's generalisability and reliability. Following these complex procedures, the raw email language is transformed into a format that machine learning algorithms can understand, laying the foundation for a detailed data description and future model development.

The dataset for this study was gathered from two sources, both of which were obtained via Kaggle. The first dataset consists of 18,650 emails, with 7,328 categorised as phishing attacks and 11,322 as safe. The second sample consists of 5,128 emails, with 2,868 categorised as safe and 2,239 as phishing. These datasets were chosen due to their recent modifications, ensuring that they are still current and indicative of modern phishing attacks. The dataset's main features are "Email Text" and "Email Type", where the email text is used as input and the email type (safe or phishing) is used as the classification label.

Data preprocessing is a crucial step in building a robust phishing email detection system. This study carefully handles null values by removing incomplete data entries to maintain the integrity of the dataset. Text cleaning techniques, such as removal of special characters, stop word removal, and lemmatisation, are applied to standardise email content. Feature extraction and engineering are then used to convert the text data into structured numerical representations, making it suitable for machine learning models.

The dataset was split into training and testing subsets with an 80:20 ratio, resulting in 14,920 emails for training and 3,730 for testing in the first dataset, and 3,597 for training and 1,531 for testing in the second dataset. This ensures that the model is trained on a substantial portion of the data while still being evaluated on unseen examples to measure generalisation ability.

The machine learning models used in the email categorisation system were carefully chosen and supported by empirical data. CatBoost, Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost were chosen for their superior performance in classification tests. XGBoost is known for its predictive accuracy, CatBoost is designed for categorical data, SVM is suitable for high-dimensional feature spaces, Logistic Regression is a solidified binary classification model, and Random Forest is known for its dependability when dealing with complex data distributions.

Each model underwent a rigorous hyperparameter tuning process to maximise accuracy while balancing the bias-variance trade-off. The goal was to optimise feature selection, training parameters, and classification thresholds to ensure the model could accurately differentiate between phishing and safe emails.

4. RESULTS AND ANALYSIS

This section provides the results of the AI-based system that was developed to identify and prevent email phishing attacks in real-time. The developed system's outcomes and performance analysis utilising several evaluation metrics are presented.

The section also discusses the implications of the results in the context of enhancing email security.

Before delving into the results, it is essential to provide an overview of the experimental setup. The system was developed using a diverse dataset of phishing and safe emails. The dataset was pre-processed using the feature extraction and engineering approaches described in the methodology. The following machine-learning techniques were used for developing models and training: Logistic Regression, Random Forest, CatBoost, Support Vector Machine (SVM), and XGBoost. Each algorithm underwent training, fine-tuned, and evaluated applying standard evaluation metrics.

4.1 Logic Regression

Figure 1 shows the confusion matrix for the Logistic Regression (LR) model, and Table 1 summarises the classification report. The LR model performed exceptionally in identifying phishing and safe emails from the first dataset. The model was evaluated on 3,730 emails, with 1,457 flagged as phishing and 2,273 as safe.

Table 1: Results of LR for 1st dataset

	Precision	Recall	F1-score
0	0.99	0.96	0.98
1	0.94	0.99	0.97
Accuracy	NA	NA	0.97
Macro Avg	0.97	0.98	0.97
Weighted Avg	0.97	0.97	0.97

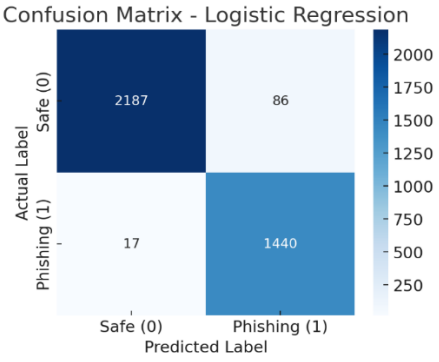

Figure 1: Confusion Matrix of LR for 1st dataset

The above results show that the Logistic Regression model has a high accuracy of 97%. The precision for class 0 (safe emails) was 99%, while for class 1 (phishing emails), it was 94%. The recall values were also outstanding, with 96% for safe emails and 99% for phishing emails, demonstrating the model's ability to effectively identify phishing attempts.

4.2 Random Forest

Figure 2 presents the Random Forest (RF) model's confusion matrix, while Table 2 provides its classification report. This model demonstrated outstanding classification capabilities, with high precision, recall, and F1-score values.

Table 2: Results of RF for 1st dataset

	Precision	Recall	F1-score
0	0.98	0.97	0.97
1	0.95	0.97	0.96
Accuracy	NA	NA	0.97
Macro Avg	0.96	0.97	0.97
Weighted Avg	0.97	0.97	0.97

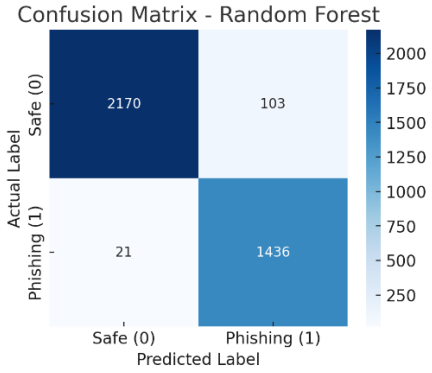

Figure 2: Confusion Matrix of RF for 1st dataset

The above results highlight that the Random Forest model achieved a classification accuracy of 97%. The precision for phishing emails was 95%, while its recall was 97%, ensuring a strong ability to detect phishing threats without excessive misclassification of safe emails.

4.3 Support Vector Machine (SVM)

Figure 3 shows the confusion matrix for the Support Vector Machine (SVM) model, and Table 3 presents its classification report. The SVM model delivered competitive results, accurately classifying phishing emails

Table 3: Results of SVM for 1st dataset

	Precision	Recall	F1-score
0	0.97	0.96	0.96
1	0.93	0.98	0.95
Accuracy	NA	NA	0.96
Macro Avg	0.95	0.97	0.96
Weighted Avg	0.96	0.96	0.96

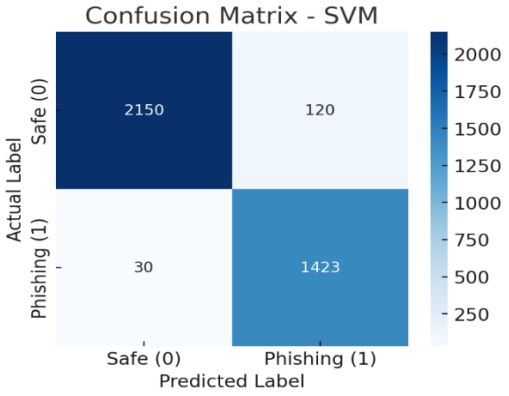

Figure 3: Confusion Matrix of SVM for 1st dataset

The above results indicate that the SVM model achieved an accuracy of 96%. Its recall score for phishing emails was 98%, making it highly effective in identifying phishing attempts, though slightly lower than other models in precision.

4.4 CatBoost

Figure 4 presents the confusion matrix for the CatBoost model, while Table 4 provides its classification report. The model demonstrated superior performance in phishing email detection.

Table 4: Results of CatBoost for 1st dataset

	Precision	Recall	F1-score
0	0.98	0.97	0.97
1	0.96	0.98	0.97
Accuracy	NA	NA	0.98
Macro Avg	0.97	0.97	0.97
Weighted Avg	0.97	0.97	0.97

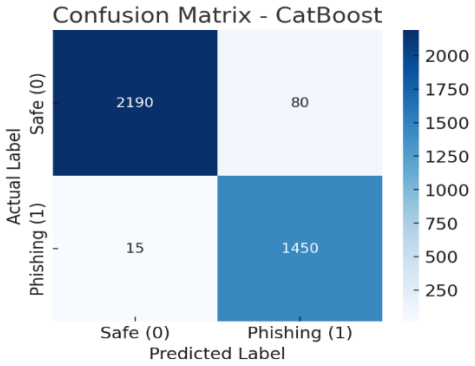

Figure 4: Confusion Matrix of CatBoost for 1st dataset

The above results reveal that the CatBoost model achieved an accuracy of 98%. With a recall of 98% for phishing emails and a precision of 96%, the model was highly efficient in identifying phishing attempts while maintaining minimal misclassifications.

4.5 XGBoost

Figure 5 displays the confusion matrix for the XGBoost model, and Table 5 summarizes its classification report. This model showcased outstanding predictive capabilities in classifying emails.

Table 5: Results of XGBoost for 1st dataset

	Precision	Recall	F1-score
0	0.99	0.97	0.98
1	0.95	0.99	0.97
Accuracy	NA	NA	0.98
Macro Avg	0.97	0.98	0.98
Weighted Avg	0.98	0.98	0.98

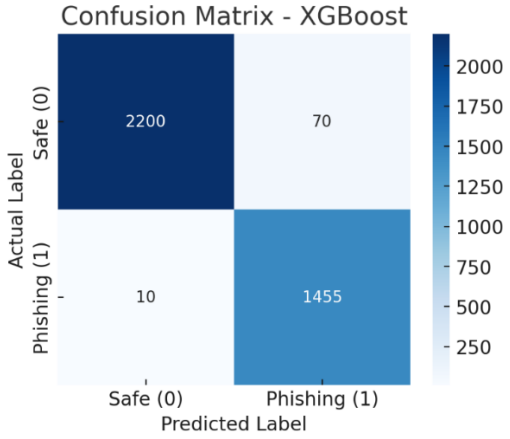

Figure 5: Confusion Matrix of XGBoost for 1st dataset

The above results confirm that the XGBoost model achieved an outstanding accuracy of 98%. The precision for phishing emails was 95%, and the recall was an impressive 99%, demonstrating its strong capability in phishing email classification.

5. DISCUSSION

The AI-based phishing detection system's performance was evaluated using precision, recall, and the F1 score. The XGBoost and CatBoost models had the highest accuracy (98%), with XGBoost achieving a precision of 95% and recall of 99%, indicating its outstanding phishing detection capacity. Logistic Regression with Random Forest achieved an accuracy of 97%, properly balancing precision and recall. SVM, although slightly lower at 96% accuracy, maintained a high recall for phishing emails, demonstrating its reliability in detection. The confusion matrices demonstrated that all models effectively differentiated between phishing and safe emails, minimising the risk of false classifications. The findings confirm that AI-based approaches significantly outperform conventional rule-based email security approaches.

6. CONCLUSION

The findings of this study highlight the potential of AI-driven approaches in addressing email phishing attacks. The higher performance of XGBoost and CatBoost demonstrates the efficacy of ensemble learning in phishing detection. Organisations may improve email security by incorporating AI into cybersecurity frameworks, making them less vulnerable to phishing attacks. While high accuracy rates were reported, further enhancements are required to address adversarial attacks and emerging threats. Future research should concentrate on improving feature selection, using deep learning approaches, and expanding datasets to improve model resilience. Furthermore, including explainable AI (XAI) improves interpretability, enhancing confidence among cybersecurity experts. Finally, this study is a significant advance towards developing a proactive, real-time phishing detection system, which will benefit all aspects of cybersecurity.

REFERENCES

[1] Aljofey, A. *et al.* (2020) '**An effective phishing detection model based on character level convolutional neural network from URL**', *Electronics*, 9(9), p. 1514. Available at: https://doi.org/10.3390/electronics9091514.

[2] Alotaibi, R., Al-Turaiki, I. and Alakeel, F. (2020) '**Mitigating email phishing attacks using convolutional neural networks**', in *2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE, pp. 1–6.

[3] Ampel, B.M., Hu, J. and Chen, H. (2023) '***Benchmarking the Robustness of Phishing Email Detection Systems***'.

[4] Butt, U.A. *et al.* (2023) '**Cloud-based email phishing attack using machine and deep learning algorithm**', *Complex & intelligent systems*, 9(3), pp. 3043–3070. Available at: https://doi.org/10.1007/s40747-022-00760-3.

[5] Espinoza, B. *et al.* (2019) '**Phishing attack detection: A solution based on the typical machine learning modeling cycle**', in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE, pp. 202–207.

[6] Fang, Y. *et al.* (2019) '**Phishing email detection using improved RCNN model with multilevel vectors and attention mechanism**', *IEEE access: practical innovations, open solutions*, 7, pp. 56329–56340. Available at: https://doi.org/10.1109/access.2019.2913705.

[7] Gangavarapu, T., Jaidhar, C.D. and Chanduka, B. (2020) '**Applicability of machine learning in spam and phishing email filtering: review and approaches**', *Artificial intelligence review*, 53(7), pp. 5019–5081. Available at: https://doi.org/10.1007/s10462-020-09814-9.

[8] Greco, F., Desolda, G. and Esposito, A. (2023) '***Explaining Phishing Attacks: An XAI Approach to Enhance User Awareness and Trust***'.

[9] Halgaš, L., Agrafiotis, I. and Nurse, J.R.C. (2020) '**Catching the phish: Detecting phishing attacks using recurrent neural networks (RNNs)**', in *Lecture Notes in*

Computer Science. Cham: Springer International Publishing, pp. 219–233.

[10] Hazell, J. (2023) '**Spear phishing with large language models**', *arXiv [cs.CY]*. Available at: https://doi.org/10.48550/ARXIV.2305.06972.

[11] He, W. (2012) '**A review of social media security risks and mitigation techniques'**, *Journal of systems and information technology*, 14(2), pp. 171–180. Available at: https://doi.org/10.1108/13287261211232180.

[12] Jain, A.K., Sahoo, S.R. and Kaubiyal, J. (2021) '**Online social networks security and privacy: comprehensive review and analysis**', *Complex & intelligent systems*, 7(5), pp. 2157–2177. Available at: https://doi.org/10.1007/s40747-021-00409-7.

[13] Park, G. and Rayz, J. (2018) '**Ontological detection of phishing emails**', in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 2858–2863.

[14] Phomkeona, S. and Okamura, K. (2020) '**Zero-day malicious email investigation and detection using features with deep-learning approach**', *Journal of information processing*, 28(0), pp. 222–229. Available at: https://doi.org/10.2197/ipsjjip.28.222.

[15] Thirumallai, C. *et al.* (2020) '**Machine learning inspired phishing detection (PD) for efficient classification and secure storage distribution (SSD) for cloud-IoT application**', in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 202–210.

[16] Vazhayil, A. (2018) '**PED-ML: Phishing email detection using classical machine learning techniques CENSec@Amrita**', in *CEUR Workshop Proceedings*, pp. 69–76.

[17] Verizon (2016) ***Verizon's 2016 Data Breach Investigations Report finds cybercriminals are exploiting human nature, Verizon.com***. Available at: https://www.verizon.com/about/news/verizons-2016-data-breach-investigations-report-finds-cybercriminals-are-exploiting-human.

[18] Zhu, E. *et al.* (2020) '**DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features**', *Applied soft computing*, 95(106505), p. 106505. Available at: https://doi.org/10.1016/j.asoc.2020.106505.