

AI-Driven Phishing: Adaptive Defense Strategies

Joel Yim, Jason Xu
University of Washington

ABSTRACT

Phishing attacks remain one of the biggest Cybersecurity threats, using manipulative tactics to obtain confidential information. Modern anti-phishing tools aren't efficient enough in detecting hyper-personalized AI-powered phishing techniques, which can bypass many of the email-based spam filters and block lists. We will evaluate and examine the existing AI-based phishing detection techniques, including models such as ML, DL, and NLP models, exploring their advantages and limitations. Additionally, we explore Hybrid Learning models that incorporate AI detection tools with human oversight to mitigate false positives and improve accuracy. Our analysis will identify the key limiting factors in these different models and propose a solution to improve phishing detection.

Index Terms—Phishing detection, ML (Machine Learning), DL (Deep Learning), NLP (Natural Language Processing).

I. INTRODUCTION

Phishing attacks continue to remain and grow as one of the most prevalent cybersecurity threats today due to their nature of exploiting human vulnerabilities. They are low-cost, high-impact attacks that require little to no technical skills to fully execute, which makes them an attractive cyberattack option for threat actors. By creating a sense of urgency, phishing emails can easily bypass a lot of the conventional security measures used today, such as Secure Email Gateways (SEG) and blocklists. The FBI reported more than \$10.3 billion in financial losses related to phishing in 2022, highlighting how important it is to monitor this threat. [5]

SEG solutions and blocklists are quite ineffective against evolving AI phishing tactics, since SEG relies on predefined rules and signatures, making it unable to detect highly sophisticated generated emails that closely mimic legitimate messages. Also, blocklists fail on the same level, since they aren't effective when attackers use newly created or compromised accounts, leaving a lot of the phishing attempts undetected.

To counteract these challenges, AI-based phishing detection models such as ML & DL models have been used, which have been used to classify and detect advanced email-based anomalies and improve detection accuracy. However, challenges remain, specifically detecting new and emerging AI-generated content and minimizing false positives, which may compromise the effectiveness of current systems.

One of the ways the research was conducted was by developing an API and utilizing tools such as Python, Flask, Scikit-Learn, Pandas, and Joblib to test and analyze data. We employed ML techniques such as logistic regression for

classification and model evaluation to assess the effectiveness of our approach in real-world scenarios.

This paper aims to:

- Define the challenges in phishing detection.
- Evaluate existing AI-based solutions.
- Propose enhancements to mitigate false positives through hybrid learning approaches.

Our research highlights gaps in current models and suggests improvements to enhance phishing detection efficiency and scalability.

II. RELATED WORK

Prior research on AI-based phishing detection has primarily focused on ML, DL, & NLP models, which have shown significant potential in identifying phishing attacks, but they come with inherent limitations.

- **Machine Learning (ML):** Traditional ML algorithms, such as Decision Trees, K-nearest neighbors (KNN), and Random Forests, have been widely used for phishing detection. These methods analyze features such as URL patterns, sender reputation, and email content to classify messages as phishing or legitimate. While these techniques are effective at identifying known attack patterns, they often produce high false positive rates and can struggle with detecting novel or sophisticated phishing attempts [1].
- **Deep Learning (DL):** As noted, CNNs and RNNs have demonstrated great potential in learning intricate patterns from extensive datasets and identifying phishing attacks. DL techniques can automatically extract relevant features from data and adapt to evolving phishing strategies. However, they typically require large amounts of labeled data and significant computational power to achieve high accuracy [2].
- **Natural Language Processing (NLP):** NLP methods, such as sentiment analysis and semantic similarity models, are increasingly used for text-based phishing detection. These models focus on understanding the language and intent behind phishing emails to identify suspicious patterns. While NLP can effectively capture subtle linguistic cues, it often struggles with highly personalized attacks and can be computationally expensive [3].

Despite these advancements, existing methods face several challenges:

- **High False Positives:** Many AI-based systems produce high false positive rates, where legitimate emails are

mistakenly flagged as phishing attempts. This can lead to user frustration and reduced trust in the system.

- **Scalability Issues:** As phishing attacks become more sophisticated, the volume of data that needs to be processed increases. Traditional ML models struggle to scale effectively with large datasets, leading to delays in real-time detection and higher computational costs.
- **Inability to Detect New Attack Techniques:** Since many models rely on historical data to identify phishing patterns, they are less effective at detecting novel or evolving attack methods. For example, attacks that use new tactics, such as QR codes or social engineering, may bypass existing detection systems.

To address these limitations, hybrid approaches that combine the strengths of ML, DL, and NLP, along with human oversight, have been proposed. These hybrid systems aim to reduce false positives, improve scalability, and enhance the detection of new phishing techniques. Reinforcement learning, in particular, is being explored to continuously adapt models to new data and threats, improving their ability to detect both known and novel phishing attacks.

III. METHODOLOGY

A. Exploring AI-Driven Phishing Detection

We explored various AI-driven phishing detection models, such as Machine Learning (Decision Tree, Random Forest, KNN, SVM, Logistic Regression), Deep Learning (CNN, RNN, Transformers,.) and Natural Language Processing (BERT) approaches.

Our research compared the performance of several models, including KNN, Decision Tree, and Random Forest, alongside other machine learning techniques used in phishing detection studies. Based on prior research findings (Table 1), these models demonstrated varying levels of accuracy and false positive rates.

However, for our implementation, we selected Logistic Regression because of its usability, feasibility, and suitability for phishing detection.

Our phishing detection model was developed using Python, Flask, Scikit-Learn, and Pandas. The system was trained on a labeled dataset of phishing and legitimate emails, with feature selection focusing on email content analysis (keywords, frequency of suspicious words, and metadata).

In addition, we created a web interface to allow users to input email text and receive phishing classification results in real-time. This interface integrates with our trained Logistic Regression model and provides a simple way to test email samples.

B. Data Preparation

The dataset was sourced from Kaggle, specifically the Phishing Email Static Dataset. We processed the emails by extracting relevant features and preparing them for classification, ensuring the dataset was cleaned and ready for model training.

C. Model Training

We ran the processed dataset through our phishing detection model, using machine learning techniques such as Logistic Regression to classify phishing vs. non-phishing emails. The model was trained to recognize patterns in the data, focusing on identifying suspicious elements like keywords and email metadata.

D. Deployment

The phishing detection model was locally hosted, and the program was run to test the model's effectiveness. We also created a web interface that allowed users to input email text and receive phishing classification results in real time. This interface is integrated with our trained Logistic Regression model, offering an accessible way for users to test email samples.

Figure 1 shows the interface displaying test email results classified as either "phishing" or "legitimate."

Phishing Detector

This program detects phishing emails using a machine learning model. Use the buttons below to run the test script or import a new dataset.



Fig. 1. Web interface displaying phishing classification results for test emails.

E. Automation

To enhance the efficiency and reliability of the system, we implemented error handling and automated processes. This included the automatic skipping of empty cells in the dataset to ensure smooth execution and avoid interruptions during the processing phase.

Phishing Detector

This program detects phishing emails using a machine learning model. Use the buttons below to run the test script or import a new dataset.

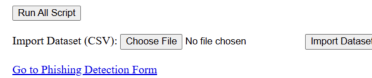


Fig. 2. Automation process in handling empty cells and ensuring smooth execution.

F. Future Improvements

Looking ahead, we plan to refine the user interface for a more seamless experience. Additionally, we aim to expand our model by incorporating more machine learning detection

methods and exploring deep learning (DL) approaches for enhanced performance and accuracy.

G. Addressing Existing Challenges

- **Detection of AI-Generated Phishing:** We will assess the effectiveness of deep learning-based anomaly detection in identifying AI-generated phishing attempts that bypass traditional filters.
- **Balancing False Positives and False Negatives:** Our research will explore reinforcement learning strategies for adaptive threshold tuning to optimize detection accuracy.
- **Scalability Considerations:** We will discuss how cloud-based AI models can support large-scale phishing detection, ensuring real-time performance and adaptability to evolving phishing techniques.

H. Addressing Existing Challenges

- **Inability to Detect AI-Generated Phishing:** We will explore pre-trained models and implement adversarial training techniques to improve robustness and detect AI-generated phishing attempts that may not match traditional patterns.
- **High False Positives:** Reinforcement learning techniques will be explored for adaptive threshold tuning to balance false positives and false negatives in real time. This approach can help the model learn over time and adjust thresholds based on real-world feedback.
- **Scalability Issues:** Cloud-based solutions will be employed to handle large datasets and perform real-time phishing detection. This will ensure that our model can scale to handle high volumes of email traffic without significant delays in processing.

IV. QUALITATIVE ANALYSIS

A. Evaluation Metrics

We assess phishing detection models based on:

- Accuracy and precision in classification.
- False positive and false negative rates.
- Model scalability and processing speed.

B. Comparison with Existing Solutions

Table 1 summarizes a qualitative comparison of existing phishing detection methods in terms of their accuracy rates, while Table 2 provides insights into their false positive rates and scalability.

Our analysis of the reported accuracy suggests that hybrid models (ML + NLP + Human) provide the highest performance, achieving an accuracy of approximately 95% across various studies. Other ML models such as Random Forest and SVM on the other hand, were able to reach accuracy rates of approximately 85%-90%, which are effective but less optimal for more complex phishing scenarios. Deep learning models like CNN and RNN showed impressive promise, achieving accuracies around 90% and above, but faced limitations in scalability.

When it comes to scalability, methods like Hybrid (ML + NLP + Human) and ML (K-Nearest Neighbors, SVM) performed well, demonstrating the potential for deployment in larger systems with high scalability. However, some methods, especially deep learning-based approaches (CNN, RNN), showed low scalability, which could limit their practical applications in large-scale phishing detection systems.

These findings underline the importance of considering both accuracy and scalability when selecting or developing phishing detection models, as high accuracy does not always guarantee the feasibility of deployment in real-world systems with large datasets.

TABLE I
ACCURACY RATES REPORTED IN DIFFERENT STUDIES

Study	Method	Accuracy
Yahya et al. (2021)	ML (Decision Tree, Random Forest)	85%–94%
Yahya et al. (2021)	ML (K-Nearest Neighbors, SVM)	Up to 97.6%
Basit et al. (2021)	ML (Random Forest, SVM, Decision Tree)	85%
Basit et al. (2021)	DL-Based (CNN, RNN)	90%
Basit et al. (2021)	Hybrid (ML + NLP + Human)	95%
Ahmad et al. (2024)	ML (RF, SVM, Logistic Regression)	85%- 90%
Ahmad et al. (2024)	DL/NLP (Transformers, BERT)	95%

Note: CNN - Convolutional Neural Network, RNN - Recurrent Neural Network, RF - Random Forest, KNN - K-Nearest Neighbors, SVM - Support Vector Machine, BERT - Bidirectional Encoder Representations from Transformers.

TABLE II
OBSERVATIONS ON FALSE POSITIVE RATES AND SCALABILITY

Study	Method	Accuracy	Scalability
Yahya et al. (2021)	ML (DT, RF)	Moderate	Moderate
Yahya et al. (2021)	ML (KNN, SVM)	High	Moderate
Basit et al. (2021)	ML (RF, SVM, DT)	High	Moderate
Basit et al. (2021)	DL-Based (CNN, RNN)	Moderate	Low
Basit et al. (2021)	Hybrid (ML + NLP + Human)	High	High
Ahmad et al. (2024)	ML (RF, SVM, LR)	Not Detailed	Moderate
Ahmad et al. (2024)	DL/NLP (Transformers, BERT)	Improved	Low

Note: CNN - Convolutional Neural Network, RNN - Recurrent Neural Network, RF - Random Forest, KNN - K-Nearest Neighbors, SVM - Support Vector Machine, BERT - Bidirectional Encoder Representations from Transformers.

V. CONCLUSIONS AND FUTURE WORK

Our study identifies key weaknesses in current phishing detection systems and emphasizes the importance of hybrid AI-human approaches. Our research-validated that machine learning (ML) models form a strong foundation for phishing detection. However, refinements are necessary, as these models can struggle with false positives, dataset limitations, and adapting to new phishing tactics.

Deep learning (DL) and natural language processing (NLP) models show considerable promise for analyzing phishing patterns more effectively and adapting to increasingly sophisticated attack strategies. However, these models require more computing power and larger datasets, limiting their practical deployment in real-world environments.

Future research should focus on:

- Enhancing adversarial training to counter AI-generated phishing.
- Developing real-time phishing detection with low-latency AI models.
- Expanding dataset diversity for improved model generalization.
- Exploring hybrid AI models that combine ML, DL, and NLP techniques to create more robust and adaptable phishing detection systems.
- Integrating real-world phishing datasets, including meta-data such as sender authentication, to improve detection accuracy.

In conclusion, while machine learning provides a strong starting point for phishing detection, continuous advancements in AI techniques and better data sources are necessary to stay ahead of evolving cyber threats.

ACKNOWLEDGMENT

The authors would like to thank Dr. Geetha for her invaluable assistance and support throughout the research process.

REFERENCES

- [1] F. Yahya *et al.*, "Detection of Phishing Websites using Machine Learning Approaches," *2021 International Conference on Data Science and Its Applications (ICoDSA)*, Bandung, Indonesia, 2021, pp. 40-47, doi: 10.1109/ICoDSA53588.2021.9617482.
- [2] S. Ahmad *et al.*, "Across the Spectrum In-Depth Review AI-Based Models for Phishing Detection," *IEEE Open Journal of the Communications Society*, doi: 10.1109/OJCOMS.2024.3462503.
- [3] A. Basit *et al.*, "A Comprehensive Survey of AI-Enabled Phishing Attacks Detection Techniques," *Telecommunications Systems*, vol. 76, pp. 139-154, 2021, doi: 10.1007/s11235-020-00733-2.
- [4] I. Naseer, "The Role of Artificial Intelligence in Detecting and Preventing Cyber and Phishing Attacks," *European Journal of Advances in Engineering and Technology*, vol. 11, no. 9, pp. 82-86, 2024. [Online].
- [5] Federal Bureau of Investigation, "2022 Internet Crime Report," FBI, 2022. [Online]. Available: https://www.ic3.gov/AnnualReport/Reports/2022_ic3report.pdf. [Accessed: Feb. 03, 2025].
- [6] O. Perera and J. Grob, "Generative AI in Phishing Detection: Insights and Research Opportunities," *2024 Cyber Awareness and Research Symposium (CARS)*, Grand Forks, ND, USA, 2024, pp. 1-5, doi: 10.1109/CARS61786.2024.10778758.
- [7] B. V. Pavani, D. Mahitha, and B. U. Maheswari, "Enhancing Online Safety: Phishing URL Detection Using Machine Learning and Explainable AI," *2024 15th International Conference on Computing Communication and Networking Technologies (ICT)*, Kamand, India, 2024, pp. 1-6, doi: 10.1109/ICCCNT61001.2024.10723976.
- [8] Lavanya, R. R. Kumaran, M. Ashiq, M. Kumar. K, and V. Vishal, "Phishing Site Detection using Machine Learning," *2024 International Conference on System, Computation, Automation and Networking (ICSCAN)*, Puducherry, India, 2024, pp. 1-5, doi: 10.1109/ICSCAN62807.2024.10894084.
- [9] R. Ferdaws and N. E. Majd, "Phishing URL Detection Using Machine Learning and Deep Learning," *2024 IEEE World AI IoT Congress (AIoT)*, Seattle, WA, USA, 2024, pp. 0485-0490, doi: 10.1109/AI-IoT61789.2024.10579005.
- [10] S. Jain and C. Gupta, "A Support Vector Machine Learning Technique for Detection of Phishing Websites," *2023 6th International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10111968.