Base Papers for AI Powered Phishing Link Identifier for Social Media DMs

Authors:
Nayana H S (nayanahs1301@gmail.com)
Harshitha R (harshi.r04@gmail.com)
Namitha Biswal (namitha05.nb@gmail.com)
Mahesh Gowda N S (maheshgowdamahesh01@gmail.com)
Guide:
Dr Guruprasad Y K (hodcy@svcengg.edu.in)

1. AI-Powered Phishing Detection and Prevention

1.1 Abstract

The purpose of this paper is to provide a comprehensive review of the way in which artificial intelligence can improve the capabilities of detecting and preventing phishing, as well as traditional approaches to phishing and why they fail to keep up with the changes in the ways people are using technology, tools, or other types of attacks to target individuals using phishing. Authors discuss how AI-based solutions can leverage machine learning, deep learning, Natural Language Processing (NLP), and ensemble-based learning models to improve the ability to detect even the smallest of phishing patterns. In the conclusion to this study, the authors stress that the adaptability, automation, and scalability of AI systems create a strong defence against the ever-evolving threats associated with phishing.

1.2 Approach

The approach taken in this paper is not to provide a proposed phishing detection system or experimental model. Instead, the paper has used an analytical/survey approach to assess the phishing detection methods currently available.

The first stage of this survey is to describe and summarize the heuristic and signature-based techniques that have historically formed the basis of security mechanisms, and the limitations that they present. The next stage of the analysis focuses on Machine Learning (ML) and Deep Learning (DL) techniques, with a specific focus on how ML and DL techniques are able to improve the accuracy of detecting phishing attempts by learning from large amounts of data, through either supervised, unsupervised, ensemble or hybrid forms of AI models.

Additionally, the paper examines how Natural Language Processing (NLP) and Sentiment Analysis can assist in identifying the use of social engineering tactics and psychological manipulation techniques employed by phishing attackers. Finally, the authors consider some of the practical challenges associated with deploying AI-based phishing detection technologies in the real world (e.g., scalability, integration with other systems, timeframes).

1.3 Contributions

The authors provide a clear and categorized breakdown of AI methods for characterizing online phishers into groups with different machine learning strategies and approaches. Their detailed analysis of the performance benefits of combining different types of classifiers into one ensemble or hybrid model has implications for the future development of robust phishing detection solutions.

They also identify emerging areas of research that present challenges for AI-based phishing detection systems. For example, adversarial machine learning is demonstrating how attackers can exploit weaknesses in existing AI detection systems through malicious modifications of the data that a system uses for training purposes, thereby defeating the AI's ability to detect phishing attacks.

Finally, the authors provide researchers and practitioners with a reference outline of the state of current research with a summary of knowledge gaps and potential directions for advancing the development of effective phishing detection systems.

1.4 Limitations

This paper's most significant weakness is that it does not contain any empirical experimental results or a working prototype for a phishing detection system, making it impossible to empirically validate any of the methods proposed in this work.

There are no quantitative performance comparisons such as accuracy, precision, recall, and false-positive rates to enable an objective evaluation of the effectiveness of the various AI-based methods. Related practical issues (e.g., cost of deployment, cost of computation, latency of operations) are addressed briefly but not analysed, which limits the applicability of these methods in real-world scenarios.

Privacy issues associated with email content analysis and the processing of user data are not adequately examined, and there are no discussions in this study regarding the ethical implications of using automated systems to make decisions on behalf of users and to build user trust.

This paper does not address cross-linguistic phishing attacks or the potential for dataset bias, which pose significant challenges when implementing AI-based systems for the detection of phishing attacks across diverse and international settings.

2. AI-Based Phishing Detection Systems: Real-Time Email and URL Classification

2.1 Abstract

This paper presents a framework based on artificial intelligence that detects phishing emails and harmful URLs in real time. The study tackles the limitations of traditional rule-based and blacklist-driven detection methods by combining machine learning, natural language processing, and image analysis. Supervised learning models are trained with publicly available phishing datasets and real-world email samples to ensure they are relevant. The results show that using a mix of textual, visual, and structural features significantly improves detection accuracy while keeping the false-positive rate low. This makes the system effective against modern and sophisticated phishing attacks.

In addition, the framework is designed to operate in real-time environments, enabling timely identification and mitigation of phishing threats before they reach end users. The integration of multiple analysis layers allows the system to capture both technical anomalies and social-engineering indicators commonly exploited by attackers. The study also demonstrates that AI-based detection systems can adapt better to evolving phishing techniques compared to static security mechanisms. Furthermore, the proposed approach highlights the importance of automated threat detection in reducing human dependency and response delays. Overall, the findings support the adoption of intelligent, data-driven security solutions for strengthening email and web-based communication security.

2.2 Approach

The authors propose a multi-layered artificial intelligence pipeline that analyzes phishing indicators from multiple perspectives. The textual content of emails is processed using natural language processing techniques to identify suspicious language patterns, abnormal sentence structures, and social-engineering cues commonly used by attackers. URLs embedded within emails are examined

using lexical and structural feature analysis to detect obfuscation, domain spoofing, and abnormal URL patterns. Embedded images, including logos and branding elements, are analyzed using image recognition methods to identify visual impersonation and spoofed content. Supervised machine learning models are then trained on these combined features and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. Finally, the system is validated using benchmark datasets and real phishing incidents to assess its effectiveness in real-time detection scenarios.

To enhance reliability, feature preprocessing and normalization steps are applied to reduce noise and improve model generalization. The multi-modal fusion of textual, visual, and URL-based features enables the system to detect both technical and psychological aspects of phishing attacks. The framework is designed to function within real-time email filtering environments, allowing rapid classification without significant delays. Comparative evaluation against traditional detection techniques demonstrates superior performance in identifying sophisticated phishing attempts. Overall, the proposed pipeline provides a comprehensive and scalable approach to phishing detection in modern cybersecurity systems.

2.3 Contributions

The paper presents a phishing detection framework that combines natural language processing, URL feature analysis, and image recognition into one system. This design allows the framework to capture various phishing indicators across text, links, and visual elements at the same time.

The study shows that extracting and combining textual, structural, and visual features significantly boosts phishing detection accuracy. The multimodal approach is more effective than traditional methods that rely on blacklists and rules, especially for identifying complex and new phishing attacks.

The framework has been tested in real-time detection scenarios. It demonstrates its ability to classify phishing emails and malicious URLs quickly. This confirms that the system can be used in real-world settings like email gateways and enterprise security systems. By examining both technical indicators and social engineering patterns, the system effectively counters advanced phishing tactics that use psychological manipulation and technical tricks. This two-pronged strategy enhances its strength against changing attack methods.

Combining different feature types helps lower false-positive rates by providing richer context for classification. This makes it less likely for legitimate emails to be wrongly identified, boosting user trust and making the system more reliable.

The paper emphasizes how AI-driven, multimodal phishing detection aids in automated threat identification and response. This decreases the need for manual checking and allows for quicker action against phishing threats in large communication networks.

2.4 Limitations

The proposed system relies heavily on supervised learning models that need large amounts of labelled phishing and legitimate data. This reliance may limit the system's effectiveness against new or zero-day phishing attacks, where labelled examples are not easily available.

Using image recognition techniques increases computational complexity and processing time. This

can restrict deployment on low-resource environments such as mobile devices, embedded systems, or organizations with limited infrastructure.

The paper does not thoroughly examine threats from adversarial machine learning, where attackers deliberately alter email content, URLs, or images to avoid detection. This omission may affect the long-term strength of the system against adaptive attackers.

The framework primarily addresses phishing content in one language, usually English. As phishing campaigns increasingly target multilingual users, the system's effectiveness may decrease in global and diverse communication environments.

While the paper discusses real-time performance, it does not closely analyze scalability issues, such as managing high email volumes in enterprise settings. It also does not fully explore integration challenges with existing email gateways and security systems.

3. AI-Driven Phishing Detection Systems

3.1 Abstract

This paper looks at how artificial intelligence has greatly improved phishing detection systems by moving past old static and rule-based security methods. It discusses the use of machine learning, deep learning, natural language processing, and ensemble techniques to detect phishing emails and harmful websites. By using AI's capacity to analyze large datasets and find hidden patterns, the study shows better detection abilities against complicated and zero-day phishing attacks. The authors also stress the increasing need for smart security solutions to tackle the ever-changing threat landscape and list key challenges that future research must address. The paper explains how AI-based models can continuously learn from new data, allowing them to adapt to emerging phishing strategies more effectively than conventional systems. The role of automation in reducing human intervention and response time is also highlighted, particularly in large-scale communication environments.

3.2 Approach

The study uses a conceptual and comparative research approach to look at current phishing detection methods instead of suggesting a single solution. It starts by reviewing traditional phishing detection methods and points out their shortcomings in dealing with advanced and changing attack strategies. The paper then discusses supervised and unsupervised machine learning models, showing how they learn complex phishing patterns from data. It explores natural language processing techniques, including semantic and sentiment analysis, to understand how they identify linguistic cues and psychological manipulation. The authors also cover deep learning architectures like convolutional neural networks and recurrent neural networks, along with ensemble and hybrid methods, to explain how combining multiple models can improve detection accuracy and system reliability.

The study compares the strengths and weaknesses of each approach. It highlights situations where some techniques work better than others. It underscores how important feature engineering and data quality are for getting reliable detection results. The paper also talks about scalability and the challenges of using these models in real-world settings with high volumes of email and web traffic. Additionally, the authors look into issues related to model interpretability and explainability. These issues are crucial for building trust in AI-driven security systems. Overall, the approach offers a clear understanding of how various AI techniques help in effective phishing detection.

3.3 Contributions

The paper offers a detailed and well-organized review of the role of artificial intelligence in phishing detection, helping readers understand both the capabilities and limitations of existing approaches. It explains how techniques such as natural language processing and deep learning strengthen defenses against social-engineering attacks by analyzing context, intent, and user behaviour rather than relying only on static rules.

The authors also bring attention to key research challenges, including the lack of model transparency, vulnerability to adversarial manipulation, and issues related to data quality and imbalance, all of which can impact detection performance.

Furthermore, the study discusses the need to balance high detection accuracy with computational efficiency, especially for systems operating in real-time environments. It stresses the importance of interpretable AI models to ensure trust, accountability, and ease of adoption in practical cybersecurity applications.

The paper also highlights research gaps such as limited multilingual phishing analysis and insufficient testing across diverse datasets. Overall, it acts as a valuable reference for both researchers and practitioners by consolidating current knowledge and outlining future directions for building more robust, scalable, and dependable AI-based phishing detection systems.

3.4 Limitations

The paper does not include any experimental implementation or benchmark testing to validate the discussed AI-based phishing detection techniques. As a result, the effectiveness of the reviewed methods cannot be empirically verified or compared under real-world conditions.

Although the paper discusses common evaluation metrics such as accuracy, precision, and recall, these metrics are presented only at a conceptual level. Without quantitative results or dataset-based evaluations, it becomes difficult to objectively assess the relative performance of different AI models. This restricts meaningful comparison between competing approaches.

The study does not address real-time operational challenges such as detection latency, computational overhead, or system throughput. In practical deployments, phishing detection systems must process large volumes of data within strict time constraints. The absence of this discussion reduces the relevance of the proposed insights for real-time cybersecurity applications.

The paper provides minimal analysis of dataset-related issues such as class imbalance, data diversity, and sampling bias, which significantly affect model performance. Without addressing how AI models generalize across different user groups, languages, and attack types, the robustness of the reviewed techniques remains uncertain.

Operational aspects such as system scalability, integration with existing email gateways, and maintenance costs are not discussed in detail. These factors are critical for translating AI-based phishing detection research into deployable solutions. The lack of deployment-focused analysis limits the paper's usefulness for practitioners.

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

4.1 Abstract

This paper presents an approach for detecting phishing emails in real time using artificial intelligence. It employs multiple machine learning classifiers to improve decision-making accuracy. The proposed system focuses on data preprocessing. It uses techniques like text normalization, stemming, and feature extraction to reduce noise and emphasize relevant patterns in email content. By changing raw text into meaningful numerical forms, the system supports more effective learning by classification models.

The authors assess the performance of various machine learning algorithms with a publicly available Kaggle dataset that includes both phishing and legitimate email samples. This ensures consistent and repeatable results. The experimental findings show that ensemble-based classifiers, especially XGBoost and CatBoost, outperform traditional models. They effectively capture complex and non-linear relationships in phishing data, achieving detection accuracy of up to 98%. These results highlight how gradient-boosting techniques work well with imbalanced and high-dimensional datasets. Overall, the study confirms that machine learning solutions can greatly enhance email security. They provide fast, accurate phishing detection, reduce dependence on manual analysis, and improve defenses against changing cyber threats.

4.2 Approach

The study looks at how different machine learning algorithms identify phishing emails using a model-driven experimental setup. It begins by cleaning and preparing a dataset of labeled emails. The researchers use methods like text normalization, tokenization, and stemming to make sure the data is consistent and meaningful. Then, they transform the textual data into numerical features that the models can process. This helps the models detect patterns that are often linked to phishing attempts.

The research evaluates several algorithms, including Logistic Regression, Support Vector Machines, Random Forest, CatBoost, and XGBoost. Accuracy is the main measure of performance. The results show that ensemble methods perform better than individual models by combining the strengths of multiple learners. Overall, the findings indicate that machine learning can effectively detect phishing in real-time. It can also be integrated into email security systems to automatically identify and reduce phishing risks.

4.3 Contributions

The paper offers several important contributions to phishing email detection using machine learning. First, it presents a detailed comparison of different classifiers, including both traditional algorithms and ensemble-based models, all tested under the same conditions.

This approach clarifies each model's strengths and weaknesses, helping researchers and practitioners identify the most effective algorithms for detecting phishing emails. One key finding is that gradient-boosting models like XGBoost and CatBoost consistently perform better than other methods. They effectively capture complex patterns in phishing content, making them very suitable for real-world use.

Furthermore, the study emphasizes the need for solid data preprocessing and feature extraction to

boost model performance. Techniques like text normalization, tokenization, and stemming, along with effective feature representation, significantly improve detection accuracy. Besides evaluating algorithms, the research shows that it is practical to deploy machine learning-based phishing detection systems in real-time email environments.

It demonstrates how AI can serve as a dependable tool for enhancing email security. Overall, the paper provides strong evidence that ensemble learning techniques, especially gradient-boosting models, can greatly improve the effectiveness of phishing detection systems, offering useful insights for both academic research and real-world cybersecurity applications.

4.4 Limitations

The experiments are conducted using only a single publicly available dataset, which may limit the applicability of the results to different organizations, languages, or real-world scenarios with varying phishing patterns.

Ensemble models such as XGBoost and CatBoost, although highly accurate, are treated as black boxes. The study does not explain how individual features influence predictions, which can reduce transparency and confidence in practical use.

The system is not evaluated against adversarial phishing attacks, where attackers modify content to bypass detection. This leaves uncertainty about the model's effectiveness against evolving threats.

Although real-time deployment is discussed, no practical testing is performed in live email environments. This limits insights into how the system would perform under dynamic, real-world conditions.

While preprocessing improves model performance, the study does not explore advanced or domain-specific feature engineering. This may constrain the detection of more subtle or sophisticated phishing techniques.

5. AI-Driven Phishing Detection: Enhancing Cybersecurity with Reinforcement Learning

5.1 Abstract

This paper presents a new way to detect phishing attacks by using reinforcement learning (RL), specifically a Deep Q-Network (DQN) architecture. Unlike standard supervised machine learning models, which depend on fixed datasets and unchanging patterns, this system keeps learning from its interactions with the environment. This lets it adjust to new and changing phishing tactics. By treating phishing detection as a decision-making task, the DQN agent assesses incoming emails. It gets rewards for correctly identifying emails and penalties for mistakes, which helps the model get better over time. The framework is tested on both real-world datasets and synthetic phishing emails. This ensures it works well against known and new attack methods. Experimental results show high detection accuracy and a notable decrease in false positives, proving the system's dependability. These findings highlight the potential of reinforcement learning as a flexible and responsive tool for cybersecurity. It can work alongside traditional detection methods to strengthen defenses against more sophisticated phishing threats.

5.2 Approach

The authors present phishing detection as a Markov Decision Process (MDP). In this setup, the system's actions, states, and rewards are clearly defined. A DQN agent learns to classify incoming emails as either phishing or legitimate. It gets rewards for correct classifications and penalties for incorrect ones. The reward function is designed to punish false positives, which improves the system's reliability in real-world use. They train the model using real-world datasets. Additionally, they validate the model's strength with synthetic phishing scenarios to see how it performs against new and unseen threats. They compare the performance of this reinforcement learning method with traditional machine learning classifiers, showing that it is more adaptable and effective for dynamic phishing detection.

5.3 Considerations

The study develops the use of RL techniques, moving from static supervised models to a system that can learn and adapt in real time.

The RL-based system shows more flexibility and responsiveness to changing phishing strategies than conventional machine learning models. By using a reward-based learning mechanism, the framework reduces false-positive rates, which is important for keeping user trust and maintaining efficiency.

The study performs extensive tests against traditional ML classifiers and assesses the system on both real and synthetic datasets, proving its strength and practical use.

The system constantly updates its understanding based on feedback from classification results, allowing it to improve over time without needing complete retraining.

The framework is built to work in near real time, enabling early detection of phishing emails as they come in, which is crucial for strengthening cybersecurity.

5.4 Limitations

Training a reinforcement learning agent with a Deep Q-Network (DQN) takes a lot of time and computing power. Organizations with limited hardware may find it hard to carry out large-scale training or make frequent updates.

The agent's learning depends on a clear reward system. If the rewards are unclear, the agent might learn ineffective strategies or miss important phishing patterns, which can harm accuracy.

Using RL-based phishing detection in real-world situations is harder than using traditional machine learning. It requires specific infrastructure, continuous monitoring, and smooth integration with existing email platforms to keep up performance.

The current model is mainly focused on email phishing, which limits its ability to spot other types of attacks, like malicious links, attachments, or social media phishing. Changes will be needed to expand its capabilities.

Reinforcement learning and ensemble models often act like black boxes. This makes it difficult to understand why certain emails are flagged. This can reduce trust and complicate troubleshooting and auditing.

The system's effectiveness heavily depends on the quality and variety of the training data. Poor or unrepresentative datasets can limit its ability to recognize new or changing phishing threats.

Conclusion

Title and Year	Pros	Cons
AI-Powered Phishing Detection and Prevention	Comprehensive survey of AI Identifies emerging challenges like adversarial ML Highlights adaptability, automation, and scalability of AI systems	No experimental validation or prototype- Lacks quantitative metrics (accuracy, precision, recall) Limited discussion on privacy, ethics, dataset bias, cross-linguistic issues
AI-Based Phishing Detection Systems: Real-Time Email and URL Classification	Multi-layered pipeline analyzing text, URLs, and images Real-time detection capability Combines technical and social-engineering indicators	Requires large labeled datasets (supervised learning) High computational cost due to image recognition Limited adversarial ML considerations
AI-Driven Phishing Detection Systems	Detailed review of ML, DL, NLP, and ensemble methods Explains role of automation in real-time detection Highlights semantic, sentiment, and behavioural analysis	No experimental implementation No quantitative benchmark testing Real-time operational constraints not analyzed Limited discussion on scalability, system integration, and deployment costs
An AI-Powered Approach to Real-Time Phishing Detection for Cybersecurity	Model-driven experimental setup using ML classifiers Detailed comparison of Logistic Regression, SVM, Random Forest, CatBoost, XGBoost- Ensemble methods achieve high accuracy	Only one public dataset used- Black-box nature of ensemble models No evaluation against adversarial phishing Limited exploration of advanced/domain-specific feature engineering
AI-Driven Phishing Detection: Enhancing Cybersecurity with Reinforcement Learning	Introduces RL (DQN) for adaptive phishing detection Reward-based learning reduces false positives Near real-time detection capability	High computational cost and long training time Complex reward function design Difficult real-world deployment Focused mainly on email phishing Black-box interpretability issues Dependent on quality and diversity of training data