Finding Phishing URL’s Using Machine Learning and Future Selections Methods

**Abstract:**

As the digital world continues to expand with the growing number of internet-connected devices, phishing attacks have become a critical cybersecurity concern. These attacks, which target human vulnerabilities rather than technical flaws, trick users into providing personal information including, but not limited to, log in credentials and financial details. While traditional and machine learning (ML)-based phishing detection methods have proven effective, they often depend on a large set of features, limiting their applicability in resource-constrained environments. Furthermore, the ever-evolving strategies of cybercriminals, including the increasingly subtle and sophisticated nature of phishing websites, present additional challenges. To address these issues, recent research has shifted towards more efficient and adaptable detection methods. This review paper provides a thorough examination of these advancements, with a focus on the adoption of Explainable AI (XAI) methods such as sapley Additive explanations (SHAP) as well as Local Interpretable Model-agnostic Explanations (LIME) into phishing detection models

**Keywords:**

Phishing detection, Cybersecurity, Explainable AI (XAI), SHAP (Shapley Additive explanations), LIME (Local Interpretable Model-agnostic Explanations), Machine learning (ML).

**1. Introduction**:

As the internet the growth of the Internet of Things has made many aspects of daily life easy, but it has also contributed to an increase in phishing attacks – one of the most widespread and malevolent kinds of cybercrime nowadays. Phishing is a method that takes advantage of the weaknesses located in human nature to deceive people into handing over sensitive secrets like passwords, bank accounts, or personal data. Traditional phishing detection techniques, while effective, often rely on extensive datasets and feature-rich models, which can be difficult to implement in environments with limited computational resources. Moreover, cybercriminals are constantly evolving their methods, making phishing websites more deceptive and harder to detect.

In response to these challenges, recent research has focused on developing more efficient and adaptable phishing detection approaches. Notably, Integrating Explainable Artificial Intelligence (XAI) techniques in the form of Shapley Additive Explanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME) into machine learning based detection systems, has received numerous attentions. These methods not only boost detection of targets but also increase the explanatory power of the models so that the reasons for making such decisions become apparent to the users. This paper provides a comprehensive review of the advancements in phishing detection, emphasizing the role of XAI in improving both efficiency and transparency in combating phishing attacks.

2. **Related Work**:

The literature review on the previous work on phishing website detection methodologies shows the evolution of the approaches through the machine learning teeth according to this rising menace. One of the core references is the study by Bahaghighat et al. (2023), which developed a highly accurate phishing website detection model using machine learning algorithms, particularly XGBoost, achieving over 99.2% accuracy. This work focuses on feature extraction from websites to identify phishing attempts, offering a promising solution for real-time detection. However, its reliance on a public dataset limits its ability to generalize across all potential phishing scenarios.

Other prominent studies include Gupta et al. (2021), which leveraged lexical features for real-time phishing URL detection, and Sahingoz et al. (2019), who focused on machine learning techniques for phishing detection via URL characteristics. These studies address phishing from different angles, such as lexical analysis and feature optimization, demonstrating the versatility of machine learning in this domain. Despite notable accuracy improvements, many models still struggle with adaptability across different datasets or the sophistication of phishing tactics, like the use of embedded objects or novel obfuscation techniques.

The literature also emphasizes the ongoing need for improved generalization and feature selection methods. For instance, works by Oram et al. (2021) and Das et al. (2024) explore gradient boosting and hybrid feature-based approaches, showing significant accuracy improvements in phishing detection while stressing the importance of selecting the right feature set. The trend in recent research is toward using advanced feature selection techniques, like genetic algorithms, or integrating domain adaptation to enhance model performance across diverse datasets.

3. **Methodology**

### **3.1 Data Collection and Preprocessing**

Data collection and preprocessing are critical steps in building robust machine learning (ML) models for phishing detection. The performance and accuracy of these models largely depend on the quality and relevance of the datasets used. This section outlines the process of gathering phishing and legitimate data, preparing it for analysis, and ensuring its suitability for machine learning models, particularly those enhanced by Explainable AI (XAI) techniques.

#### **1. Data Collection**

Phishing detection models require datasets containing both phishing and legitimate examples to effectively differentiate between malicious and benign websites or emails. For this review, commonly used publicly available datasets were considered, including:

* Phish tank: A well-known, crowd-sourced database of verified phishing URLs. Phish tank provides a continuously updated list of URLs flagged as phishing, making it a reliable resource for training models to identify phishing patterns.
* UCI Repository Phishing Dataset: This dataset includes features extracted from phishing and legitimate websites. It is widely used in phishing detection research due to its comprehensive nature, providing features like URL length, domain registration details, and the presence of suspicious keywords.
* Alexa Top 1 million: This dataset lists the most frequently visited websites globally. It is often used as a source of legitimate URLs, providing a balanced comparison against phishing URLs in the dataset.

Other specialized datasets from journals and research papers, such as phishing email datasets and datasets containing hybrid URL and HTML features, were also reviewed to cover various types of phishing attacks.

#### **2. Feature Engineering**

Feature engineering plays a important role in phishing detection, as the features used by a model determine its ability to accurately classify phishing and legitimate instances. Phishing detection models typically rely on three primary types of features:

* URL-Based Features: These include characteristics like URL length, the number of subdomains, the use of IP addresses in URLs, and suspicious lexical patterns (e.g., misspelled words or unusual symbols). Studies have shown that phishing URLs often exhibit distinct patterns, such as longer lengths and the inclusion of misleading words (e.g., "secure" or "login").
* Content-Based Features: This category involves analysing the content of a webpage or email, such as the presence of suspicious HTML elements, the number of hyperlinks, the inclusion of forms requesting sensitive data, or embedded JavaScript. While content-based features can provide significant insights, they may be computationally expensive, making their use less feasible in real-time systems.
* Network-Based Features: These features include domain registration details (e.g., WHOIS data, domain age), the use of HTTPS, and the IP addresses geographic location. These features are particularly useful in identifying phishing websites that impersonate legitimate domains but have suspicious registration histories.

#### 3. **Data Cleaning**

Before feeding the data into phishing detection models, it undergoes a thorough cleaning process to ensure quality and relevance. The cleaning process typically involves:

* Handling Missing Data: Datasets may contain missing values, especially in fields like domain registration information or WHOIS data. Missing data can lead to inaccurate predictions, so various imputation techniques, such as filling in with mean/mode values or removing incomplete records, are used depending on the dataset's characteristics.
* Removing Duplicates: Duplicate records, especially in large datasets like Phish tank, can skew model training by introducing bias. All duplicate URLs or records are removed to ensure that each entry in the dataset is unique.
* Filtering Out Irrelevant Data: Some datasets may contain URLs or content that does not meet the phishing or legitimate criteria. For instance, URLs flagged as inactive or expired are excluded to focus only on active, relevant data.

### **3.2 Feature Selection and Optimization**

Feature selection and optimization play a crucial role in enhancing the efficiency and accuracy of phishing detection models, particularly in resource-constrained environments. Machine learning (ML)-based phishing detection often involves a wide array of features, such as URL characteristics, domain attributes, and HTML content. However, not all features contribute equally to the model’s performance, and selecting the most relevant ones can significantly reduce computational overhead while maintaining high detection accuracy. This section delves into the strategies employed for feature selection and optimization, with a special focus on leveraging Explainable AI (XAI) techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are discussed.

#### 1. **Challenges in Feature Selection**

Phishing detection models traditionally rely on a large set of features to differentiate between phishing and legitimate websites or emails. These features may include:

* URL-based features: Such as length, the presence of special characters, suspicious keywords, or subdomains.
* Domain-based features: Such as domain age, DNS records.
* Content-based features: Including HTML tags, embedded JavaScript, and hyperlink structures.
* Network-based features: Like IP address origin and SSL certificate validity.

While these features improve model accuracy, they also introduce several challenges:

* Computational Overhead: Using too many features can lead to high computational costs, making it difficult to deploy phishing detection models in real-time systems or on devices with limited processing power, such as IoT devices and mobile applications.
* Overfitting: Models trained on excessive or irrelevant features may become overfitted to the training dataset, reducing their generalizability to unseen data, especially as phishing tactics evolve.
* Data Redundancy: Some features may provide overlapping information, leading to redundancy that does not significantly enhance model performance but increases processing time.

#### **2. Feature Selection Methods**

To address these challenges, we employ selection mechanisms to focus on the most informative features while rejecting the ones that help very little if at all in the detection process. Feature selection methods are typical of the following:

* Filter Methods: These methods use statistical techniques like correlation analysis, Chi-squared tests, and mutual information to evaluate the relationship between every feature and the target label. Highly correlated or redundant features are filtered out before the model is trained.
* Wrapper Methods: Wrapper methods, such as Recursive Feature Elimination (RFE), iteratively remove or add features to evaluate their contribution to the model’s performance. These methods often yield better results than filter methods but are computationally expensive, as they require training multiple models during the feature selection process.
* Embedded Methods: Algorithms like Lasso (L1 regularization) and Ridge (L2 regularization) select features as part of the model training process by penalizing less important features. These methods are more efficient than wrapper methods and tend to yield good results with less computational cost.

#### 3. **Optimization Using XAI Techniques (SHAP and LIME)**

Recent advancements in Explainable AI (XAI) have introduced novel methods for feature selection and optimization, with SHAP and LIME leading the way. These techniques not only enhance the interpretability of phishing detection models but also aid in selecting the most influential features.

* SHAP (Shapley Additive Explanations): SHAP is a method based on principles of game theory, which quantifies the value or importance of each feature with respect to the prediction of the model. In phishing detection, SHAP can be used to rank features by their importance, providing a clear explanation of which features are most influential in identifying phishing websites or emails. By analysing SHAP values, less significant features can be pruned without sacrificing accuracy, resulting in a more streamlined and efficient model.

### **3.3 Explainability Integration (SHAP and LIME)**

Explainability in phishing detection models has become increasingly important, especially as machine learning (ML) techniques evolve from traditional rule-based systems to more complex, opaque algorithms like Random Forests, Gradient Boosting Machines (GBMs), and Neural Networks. While these advanced models offer high accuracy and better detection rates, they often lack transparency, making it difficult for cybersecurity experts to understand why a certain prediction was made. This is where Explainable AI (XAI) techniques such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have come into play.

These methods provide valuable insights into how phishing detection models make decisions, enhancing both trust and usability. SHAP and LIME not only improve interpretability but also help optimize feature selection and reduce the complexity of models. This section details how these XAI techniques are integrated into phishing detection models, improving both transparency and efficiency.

#### **The Importance of Explainability in Phishing Detection**

Phishing detection systems are often deployed in environments where decision-making needs to be swift and accurate. However, the high-stakes nature of cybersecurity means that the terms false positives, which refer to instances of legitimate websites being wrongfully categorized as phishing therefore causing damage to the economy, and false negatives, which refer to instances when phishing websites are not detected, pose great risks. Traditional machine learning models, while accurate, act as "black boxes" where predictions are made without explaining the reasoning behind them. This lack of transparency makes it difficult to:

* Understand Model Behaviour: Without explanations, security analysts are left guessing which features the model relied on to classify a website or email as phishing.
* Build Trust: End-users and security teams are more likely to trust AI-based systems when the decision-making process is clear and interpretable.
* Refine Models: Explainability helps identify which features contribute most to model predictions, enabling further model optimization and enhancement.

By integrating SHAP and LIME, phishing detection models provide clearer reasoning behind their decisions, making it easier for users and developers to trust and improve the models.

#### **2. Shapley Additive Explanations (SHAP)**

SHAP is a unified framework based on game theory that assigns a contribution value to each feature, explaining the influence of each feature in the model's output. In this context of phishing detection, SHAP helps identify the key features that drive a model’s decision to classify a URL, email, or website as phishing or legitimate.

* Global Explainability: SHAP provides a global understanding of the model by showing how each feature contributes to predictions across the entire dataset. This global view allows analysts to see which features (e.g., URL length, suspicious keywords, domain age) consistently have the most impact on phishing classification.
* Feature Importance: SHAP assigns SHAP values (importance scores) to each feature, indicating how much each feature contributes to increasing or decreasing the likelihood that a given instance (such as a URL) is phishing. This enables the selection of the most relevant features for model optimization.

#### **Local Interpretable Model-agnostic Explanations (LIME)**

LIME is one such technique which attempts to explain individual predictions by fitting an interpretable model to the vicinity of the prediction rather than the whole model. This ability to interpret model predictions in the local sense is important when looking at phish detection systems because such systems often rely on making decisions including those that would result in a false positive or a false negative.

* Local Explainability: LIME clarifies singular predictions via the construction of a local surrogate model with respect to a single instance. (e.g., a URL or email). It approximates the original model’s behaviour in this small region, making it easier to understand how the model arrived at a specific classification.

### **3.4. Performance Evaluation**

Performance evaluation is a critical aspect of phishing detection systems, especially when machine learning (ML) models are applied. Evaluating the performance of these models ensures that they are not only accurate but also efficient and robust in detecting phishing attacks while also aiming to reduce the occurrence of false positives and false negatives. In this subsection, we will outline the main metrics and evaluation paradigms used to evaluate the performance of phishing detection systems. In addition to this evaluation, we can deploy a new set of evaluation techniques thanks to the Explainable Artificial Intelligence (XAI) methods such as SHAP, LIME which are incorporated in these models, focusing not just on prediction accuracy but also on model interpretability and transparency.

#### **1. Evaluation Metrics for Phishing Detection Models**

To assess the efficacy of different phishing detection models, many performance measures are applied. These indicators help to evaluate the model's performance in terms of distinguishing between phishing and non-phishing cases, as well as in practical deployment of the model outside a laboratory environment.

#### **Resource Efficiency**

In addition to accuracy and other classification metrics, resource efficiency is a key factor in evaluating the performance of phishing detection models. This is especially important in resource-constrained environments like mobile devices, IoT systems, and edge computing platforms.

##### **2.1 Computational Cost**

Phishing detection models, especially those that rely on many features, can be computationally expensive. Evaluating computational efficiency involves measuring the model’s memory usage, processing time, and energy consumption.

##### **2.2 Scalability**

Phishing detection models need to scale effectively as the volume of incoming data increases. Scalability is measured by how well the model performs under increased workloads and whether it can maintain high detection accuracy in real-time applications.

### **3.5. Comparative Analysis**

Comparative analysis plays an important role in determining the effectiveness of phishing detection models by evaluating different machine learning (ML) algorithms, feature selection methods, and Explainable AI (XAI) techniques across multiple metrics. It provides a comprehensive view of how well each model performs, considering factors such as accuracy, computational efficiency, scalability, and interpretability. This section delves into the key aspects of comparative analysis, demonstrating its importance in selecting the most suitable phishing detection model for specific use cases.

#### **1. Purpose of Comparative Analysis**

The primary goal of comparative analysis is to assess various phishing detection models based on their strengths and weaknesses. No single model is universally optimal, and performance often depends on specific datasets, features, and implementation environments. By conducting comparative studies, researchers and developers can:

* Identify the most accurate models for detecting phishing attacks.
* Evaluate the trade-offs between model complexity and resource efficiency.
* Compare the level of interpretability and explainability provided by XAI techniques such as SHAP and LIME.
* Analyse the adaptability and scalability of models when exposed to evolving phishing tactics.

Comparative analysis is particularly useful for identifying models that balance high detection accuracy with low computational cost, which is crucial for deploying phishing detection systems in real-time or resource-constrained environments.

#### **2. Model Comparison Based on Detection Accuracy**

In most cases, and in this research, the performance of the model in terms of classification accuracy, precision, recall, and F1 score is one of the aspects of comparative analysis in phishing detection. Different machine learning algorithms often yield different results based on the characteristics of the dataset and the feature set used.

**4.Conclusion:**

In conclusion, the escalating threat of phishing attacks necessitates innovative and efficient detection methods that can adapt to the evolving tactics of cybercriminals. Traditional techniques, while valuable, face challenges in computational resource constraints and the sophistication of phishing schemes. The inclusion of Explainable Artificial Intelligence (XAI) techniques such as SHAP and LIME in phishing detection systems is likely to pave way not only to improving accuracy of such systems but also sustaining users’ trust by making it more transparent. As this field continues to develop, the collaboration between advanced machine learning models and XAI will be crucial in creating robust defences against phishing, empowering users to recognize and avoid potential threats. Ongoing research in this area will be essential to stay ahead of cybercriminals and protect sensitive information in our increasingly connected world.

**References:**

1. Bahaghighat, Mahdi, Majid Ghasemi, and Figen Ozen. "A high-accuracy phishing website detection method based on machine learning." Journal of Information Security and Applications 77 (2023): 103553.
2. Gupta, B. B., Yadav, K., Razzak, I., Psannis, K., Castiglione, A., & Chang, X. (2021). A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Computer Communications, 175, 47-57.
3. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345-357
4. Oram, Etuari, et al. "Light gradient boosting machine-based phishing webpage detection model using phisher website features of mimic URLs." Pattern Recognition Letters 152 (2021): 100-106.
5. Nagunwa, Thomas, Paul Kearney, and Shereen Fouad. "A machine learning approach for detecting fast flux phishing hostnames." Journal of Information Security and Applications 65 (2022): 103125.
6. Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. Journal of Ambient Intelligence and Humanized Computing, *10*, 2015-2028
7. Das Gupta, S., Shahriar, K. T., Alqahtani, H., Alsalman, D., & Sarker, I. H. (2024). Modeling hybrid feature-based phishing websites detection using machine learning techniques. Annals of Data Science, *11*(1), 217-242.
8. Shafin, S. S. (2024). An Explainable Feature Selection Framework for Web Phishing Detection with Machine Learning. Data Science and Management.
9. Rashid, F., Doyle, B., Han, S. C., & Seneviratne, S. (2024). Phishing URL detection generalization using Unsupervised Domain Adaptation. Computer Networks.
10. Omari (2023). Phishing Detection using Gradient Boosting Classifier. Procedia Computer Science, *230*, 120-127.
11. Abutair, H. Y, & Belghith, A. (2017). Using case-based reasoning for phishing detection. Procedia Computer Science, *109*, 281-288.
12. Routhu Srinivasa Rao1, Alwyn Roshan Pais1 "Detection of phishing websites using an efficient feature-based machine learning framework"
13. Shahrivari, V., Darabi, M. M., & Izadi, M. (2020). Phishing detection using machine learning techniques. Arxiv preprint arXiv*:*2009.11116.
14. Ubing, A. A., Jasmi, S. K. B., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Phishing website detection: An improved accuracy through feature selection and ensemble learning. International Journal of Advanced Computer Science and Applications, 10(1).
15. Butnaru, A., Mylonas, A., & Pitropakis, N. (2021). Towards lightweight url-based phishing detection. Future internet, 13(6), 154.
16. Aljofey, A., Jiang, Q., Rasool, A., Chen, H., Liu, W., Qu, Q., & Wang, Y. (2022). An effective detection approach for phishing websites using URL and HTML features. Scientific Reports, 12(1), 8842.
17. Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. Journal of King Saud University-Computer and Information Sciences, 34(2), 232-247.
18. Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing email detection using natural language processing techniques: a literature survey. Procedia Computer Science, 189, 19-28.
19. Mosa, D. T., Shams, M. Y., Abohany, A. A., El-kenawy, E. S. M., & Thabet, M. (2023). Machine Learning Techniques for Detecting Phishing URL Attacks. CMC-COMPUTERS MATERIALS & CONTINUA, 75(1), 1271-1290.
20. Babagoli, M., Agha Baba, M. P., & Solouk, V. (2019). Heuristic nonlinear regression strategy for detecting phishing websites. Soft Computing, 23(12), 4315-4327.