Phishing Detection Using Machine Learning

Raju Deb and Sadman Sakib Choudhury

Department of Computer Science, University of New Brunswick, Canada

Abstract

Phishing is a persistent and transforming enemy of cyberspace, mostly dependent on human credulity and real-world technological glitches that result in completely stealing confidential access to sensitive information such as credentials and financial data. Traditional detection methods including blacklists, rule-based systems, etc are becoming insufficient with the ever-evolving phishing techniques, such as domain spoofing and AI-generated content. Machine learning (ML) now acts as a state-of-the-art solution for adaptation, accuracy, and evaluation against both known and zero-day attacks. This article investigates the entire survey of ML techniques applied to the phishing detection domain. It defines all the methodologies, challenges, and recent advancements in this area. Leading advances have included application of deep learning, ensemble methods, and natural language processing techniques for extracting features on URLs, domains, and email content. The detailed performance evaluation section makes comparisons among top ML models through various metrics: accuracy, precision, recall, and F1 score together with their respective strengths and shortcomings. There are, however, many advances that have been made, but challenges such as dataset imbalance and the occurrence of adversarial attacks are still major hurdles in most of the phishing detection systems. Future research directions call for the exploration of using transfer learning and synthetic data generation, privacy-conserving technologies, and integration with blockchain to tackle the challenges mentioned. It shows the significant necessity of developing adaptive, scalable, and explainable systems. This survey, thus, emphasizes the crucial role of ML as a strong means by which solid defenses are built against phishing attacks and the necessity for collaboration efforts by academia and industry in keeping a secure digital landscape.

1 Introduction

Hackers use phishing as a clever technique to acquire one's sensitive information such as usernames, passwords, and credit card numbers. As one of the oldest and most significant threats that face the cybersecurity landscape, phishing continues to be a source of worry. The sheer amount of damage that can be done through phishing is due to the existence of many forms of human trust and technical vulnerabilities through which hackers impersonate good and genuine entities using email, websites, or other channels of communication. The universal penetration of digital transactions and online communications around the world has brought mushrooming increases in the size and sophistication of phishing campaigns, creating critical challenges for individuals, organizations, and countries worldwide.

1.1 Motivation

Phishing attacks have become a vital source of financial loss and loss of private data, with major consequences to individual victims. Industry reports indicate that billions are lost every year in terms of phishing-related scams. Phishing goes beyond these losses, as it creates a dent in the trust of users, damages the reputation of an organization, and puts privacy in jeopardy. For instance, spear phishing is a kind of targeted phishing, which has led to data breaches across even the most secured corporate and governmental networks.

The increasing sophistication of phishing techniques further worsens the issue. Advanced techniques including social engineering, content generated by AI, and real-time evasion strategizing have given way to making the traditional detection system obsolete. Techniques like domain spoofing, URL obfuscation, and encrypted communication channels have made it increasingly difficult to identify such forms of phishing by both human user and automated system.

Traditional anti-phishing methods like rule-based systems, blacklist-based detection, and manual monitoring could not catch up with the changing threat scenario. Their limitations include the dependency on prior knowledge and lack of adaptation to novel attacks. This leads to an urgent requirement for new innovations that can carry out real-time effective and efficient anti-phishing measures.

1.2 The Role of Machine Learning

Machine learning helps a lot to fight against the phishing detection barriers. Learning from the pattern of data, ML models can identify highly accurate as well as adaptable detection methods. Unlike the traditional methods, the ML techniques do not only identify a new phishing attack but also adapt to changing tactics and analyze enormous amounts of real-time data. These features make ML the best tool to solve all aspects of changing and evolving phishing.

Phishing detection would also use popular supervised machine learning algorithms such as decision trees, support vector machines (SVMs), and ensemble methods like random forests in order to classify phishing and legitimate URLs or emails based on their classified extracted features. But, more sophisticated methods like deep neural networks and natural language processing would make it possible to analyze complicated, unstructured data, like email content and website layouts. By incorporating machine learning into phishing detection, accuracy is increased while human dependence is reduced.

1.3 Objectives of the Survey

This survey aims to provide an exhaustive overview and understanding of those methods of phishing detection that have been developed using machine learning. The specific objectives of the survey include:

- A discussion of the motivation for adopting machine learning in phishing detection, namely, improving upon traditional detection methods.
- An overview of existing research and a classification of contemporary methods based on the features, algorithms, and datasets used.
- Research will be conducted to identify problems in applying machine learning to phishing detection, such as limitations of datasets, adversarial attacks, and real-time performance constraints.
- Further future studies will include new prospective directions meant to improve these phishing detection systems by integrating advanced machine learning techniques with privacy preservation methods.

1.4 Structure of the Paper

The remainder of this paper is organized as follows: **Section 2: Background and Key Concepts** provides an overview of phishing, its various types, and the role of machine learning in combating phishing; **Section 3: Literature Review** reviews existing research on phishing detection techniques, datasets, and algorithms; **Section 4: Current Trends and Innovations** discusses recent advancements and emerging trends in machine learning-based phishing detection; **Section 5: Performance Evaluation** defines the metrics that are used for evaluation and compares the performance of the state-of-the-art phishing detection models; **Section 6: Challenges** identifies the limitations and challenges in deploying ML-based phishing detection systems; **Section 7: Future Directions** proposes potential research areas and innovative solutions to address these challenges; and **Section 8: Conclusion** summarizes the survey findings and emphasizes the importance of continued research in this field.

2 Background and Key Concepts

Phishing attacks are a prevalent form of cybercrime that aim to deceive users into revealing sensitive information. This section provides an overview of phishing, its various types, the importance of machine learning in phishing detection, and a detailed explanation of key features commonly used for detecting phishing attempts.

2.1 Overview of Phishing and Its Types

Phishing attacks impersonate trusted organizations in order to get victims to divulge private details such as usernames and passwords, financial details, or personal data.

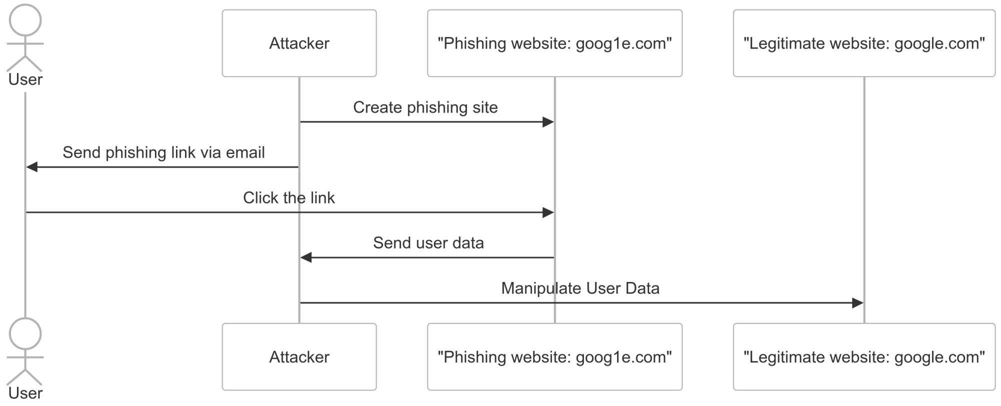
User
Attacker
"Phishing website: goog1e.com"
"Legitimate website: google.com"
Create phishing site
Send phishing link via email
Click the link
Send user data
Manipulate User Data
User
Attacker
"Phishing website: goog1e.com"
"Legitimate website: google.com"

Figure 1: Sequence Diagram of Phishing Attack

Figure 1 shows a representation of a typical sequence of actions that occur during a phishing attack: a user, an attacker, a phishing website, and a legitimate website. The attacker first creates a phishing website (e.g. "goog1e.com") which would look at least superficially to function like a real website (e.g. "google.com"). This phishing link is the bait, sent to the victim using phishing email, or via other communication means. The victim, thinking it is a legitimate link, clicks on it and gets sent to the already established phishing website, making the victim believe it is authentic. The user then unknowingly feeds the most sensitive information, such as records of his login credentials to the phishing website. The attacker would be able to steal all those details, which could be used to enter the original website illegally or to make illegal transactions. These go to show the devious nature of phishing attacks while underlining the importance of detection and prevention techniques.

Over time, phishing has evolved into several types, each targeting victims through different communication channels:

- **Email Phishing:** The most common form, where attackers send fraudulent emails mimicking legitimate organizations.
- **Spear Phishing:** A targeted attack aimed at specific individuals or organizations, often using personal information to increase credibility.
- **Smishing:** Phishing attempts conducted via SMS messages, directing users to malicious websites or prompting them to download malware.
- **Vishing:** Voice-based phishing attacks, where attackers impersonate trusted entities to extract sensitive information over the phone.
- **Clone Phishing:** Attackers replicate legitimate emails or messages, replacing valid links or attachments with malicious ones.
- **Whaling:** A highly targeted form of spear phishing aimed at high-profile individuals such as executives or government officials.

These various phishing techniques demonstrate how dynamic the danger is and how adaptive detection systems are required.

2.2 Importance of Machine Learning in Phishing Detection

The existing phishing detection schemes, such as blacklisting, rule-based systems, and manual monitoring, are all critically limited in scope within today's ever-changing threat landscape and urgently call for machine learning.

- **Zero-Day Attack**: Blacklists are made from databases of known phishing URLs and therefore cannot be applied for semi-zero-day attacks or to handle new phishing techniques. In contrast, machine learning (ML) models generalize patterns in the data and thus capture previously unseen cyber threats.
- **High False Positives**: Static rules in rule-based systems often flag legitimate sites as malicious, resulting in high false alarm incidences and overwhelming security teams with innumerable false positives. In contrast, ML algorithms reduce false positives by analyzing complex patterns and improving decision-making through varied datasets.
- **Scalability**: Manual monitoring and maintenance of rule-based systems become increasingly impractical given the sheer scale of internet traffic today. Machine learning models, on the other hand, can handle huge datasets in real-time, analyzing behaviors across billions of URLs or emails effectively.
- **Adaptability**: Phishing techniques evolve rapidly, employing methods like URL obfuscation and domain spoofing to bypass traditional countermeasures. ML models can adapt to new environments by retraining on updated data, maintaining their effectiveness against emerging threats.

Unlike traditional methods, machine learning goes beyond identifying cause-effect patterns and instead learns reactive behavior to new problems in dynamic environments by evaluating general patterns. ML models can trace link patterns across extensive data collections, enabling real-time detection of phishing attempts. Adaptability is a hallmark of these systems, making machine learning a crucial component in combating phishing in today's modern world.

2.3 Common Machine Learning Techniques in Phishing Detection

Various machine learning techniques are employed in phishing detection, including:

- **Supervised Learning:** Algorithms like Decision Trees, Support Vector Machines (SVMs), and Random Forests classify URLs or emails as phishing or legitimate based on labeled data.
- **Unsupervised Learning:** Clustering techniques group similar URLs or emails, identifying anomalies that may indicate phishing.
- **Deep Learning:** Neural networks and Natural Language Processing (NLP) models analyze unstructured data such as email content or website design.
- **Ensemble Learning:** Combining multiple models (e.g., Gradient Boosting, Random Forests) improves accuracy and reduces the likelihood of false positives.

2.4 Key Features for Phishing Detection

Feature extraction is critical in training ML models for phishing detection. These features can be grouped into three categories:

2.4.1 URL-Based Features

- **Domain of URL:** Evaluates the legitimacy of the domain name.
- **IP Address in URL:** Presence of an IP address in the URL is flagged as suspicious.
- **@ Symbol in URL:** Detects the "@" symbol, which can obscure the actual domain.
- **Length of URL:** Longer URLs (e.g., over 54 characters) are more likely to be phishing.
- **Depth of URL:** Calculates the number of subdirectories to determine complexity.
- **Redirection "//" in URL:** Flags unexpected positions of "//" within the URL.
- **HTTP/HTTPS in Domain Name:** Checks for "http" or "https" in the domain part of the URL.
- **URL Shortening Services:** Identifies shortened URLs, which are often used in phishing.
- **Prefix or Suffix "-" in Domain:** The presence of a dash in the domain name is flagged as suspicious.

2.4.2 Domain-Based Features

- **DNS Record:** Absence of a DNS record indicates a potentially malicious domain.
- **Website Traffic:** Low traffic domains (ranked above 100,000 on Alexa) are often phishing.
- **Age of Domain:** Domains less than 12 months old are flagged as phishing.
- **End Period of Domain:** Domains with less than 6 months of validity are considered suspicious.

2.4.3 HTML and JavaScript-Based Features

- **IFrame Redirection:** Detects the use of invisible iframes to redirect users.
- **Status Bar Customization:** Identifies JavaScript modifications to the status bar.
- **Disabling Right Click:** Checks if right-click functionality is disabled.
- **Website Forwarding:** Flags websites with more than four redirections.

These features enable machine learning models to differentiate between phishing and legitimate websites or emails with high accuracy.

3 Literature Review

Phishing detection has been an important field of research in cybersecurity in light of ever-increasing sophisticated phishing attacks and their frequencies. Numerous studies performed in this regard explored the provisioning of ML or DL models with components toward improving the phishing detection system's accuracy and robustness. The focus around such works includes feature extraction from URLs, email content, and website characteristics toward identifying patterns that could help deduce attempts of phishing attacks. Some of the advanced algorithms, such as Random Forest, Decision Trees, and Convolutional Neural Networks, along with their hybrid counterparts of deep learning frameworks, have loudly shown noteworthy improvements in their actual detection rates. Nevertheless, research in this area has been beset with challenges such as dataset imbalances, compute optimization limitations, and generalization to real-world scenario applications. This segment encapsulates an extensive review of the most advanced approaches in phishing detection, citing the methodologies, corresponding achievements, and limitations, thus identifying gaps that can inspire further advancements in this field.

3.1 Related Works

As part of their research on detecting phishing websites through machine learning techniques, the authors Alswailem et al. (2019) [1] focus on the Random Forest (RF) algorithm. Their study comprises feature engineering as they extracted 36 features derived from URL, page content, and page rank to distinguish between genuine and phishing activity patterns. Their systematic feature selection resulted in a minimal combination of 26 features, which provides a maximum classification accuracy of 98.8% through final RF testing. The high performance demonstrates how robust RF can be in phishing detection. The study also discussed some of the computational constraints and time limitations that restricted all feature combinations from comprehensive testing. It has a possibility of losing generalizability for various phishing attack scenarios since the features are dependent on a curated dataset.

The study by Alam et al. (2020) [2] is about using machine learning for phishing attack detection, particularly Decision Tree (DT) and Random Forest (RF) algorithms. To take a Kaggle dataset that is standardized into research, 32 properties derived from URL contents, email content, and website characteristics were analyzed. Principal Component Analysis (PCA) was another aspect selected and used in dimensionality reduction. The highest accuracy obtained was 97% using the RF classifier, which also proved robust against overfitting for scalability in large datasets. Again, the study limited the performance evaluation parameters to the specific dataset. This could also be a possible factor that would limit generalizability to other phishing attack scenarios in addition, the more likely that the RF model becomes dependent on the training data-labeled datasets, the lower its efficiency for detection of zero-day phishing attacks.

The study by Alsubaei et al. (2024) [3] proposes a novel hybrid deep learning framework, ResNeXt-GRU (RNT), enhanced with the Jaya optimization algorithm (RNT-J), for phishing detection. The framework merged various ensemble learning techniques, including autoencoders and ResNet models (EARN), for capturing robust features by addressing class imbalances in datasets using SMOTE. The RNT-J model attained a reach of 98% in its detection accuracy, competing among several state-of-the-art methods, including ResNet and DenseNet. Another important aspect of this study includes real-time adaptability and interpretability and advanced feature engineering techniques such as PCA and ensemble feature integration. Although this model shows high accuracy, it also has some computational needs and is highly curate-dependent for real-world scalability and generalization. Complicated deployment is also added by ensemble architecture dependence, which is a concern when going about low-resource settings.

The study by Lakshmi et al. (2021) [4] introduces a smart phishing detection system using Deep Neural Networks (DNNs) optimized with the Adam optimizer. The system uses 30 HTML-based features, including URL length, HTTPS tokens, and subdomains, extracted from the UCI Phishing Websites Dataset. The train evaluated against two architectures DNN model with the second model obtaining a better accuracy of 98.44% than the first model at 96.25. The Adam optimizer even outperforms SGD and RMSProp optimizers regarding training and testing accuracy, thus demonstrating the proposed scheme's competence. However, it may not be very accurate for phishing detection since it was made with static datasets and defined features. The static dataset and prescribed features do not respond to the dynamic behavior of evolving phishing strategies. Additionally, the DNN model's computational requirements may hinder its deployment in real-time detection systems.

Dutta (2021) [5] proposed a novel framework utilizing a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) for phishing detection. Temporal dimensions were considered using sequential patterns in URLs, improving predictions for phishing sites against legitimate ones. The dataset consisted of 7,900 phishing URLs and 5,800 legitimate URLs, which formed a good training and verification pool for the model resource. The RNN-LSTM model was superior in analyzing sequential URL data but severely limited in computational resources, rendering the model unsuitable for real-time detection. The framework furthermore is not concerned on the domain- and content-based features that have limited its application to pure URL-based phishing. This made the model ineffective to some extent against sophisticated phishing attacks using obfuscation techniques.

Odeh et al. (2021) [6] conducted a comprehensive review of traditional and deep learning models for phishing detection. The study reviewed included Naïve Bayes, Random Forests, and hybrid CNN-LSTM architectures and made an emphasis on deep learning capabilities for modeling complex patterns. Furthermore, the study aimed to know the integration of features derived from URLs and HTML with the goal of increasing classification accuracy. The researcher also presented issues related to overfitting and increased computation overhead, against which training is performed on small datasets. The review gives good insight into the URL-content feature combination but notes that real-time detection cannot be achieved because of the computational expense involved in deep learning models. This requirement for large labeled datasets further poses a challenge in scalability with dynamic environments.

The survey by Rastogi et al. (2021) [7] provides an in-depth analysis of phishing website detection and prevention using machine learning. It reviews the effectiveness of traditional ML algorithms such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbor (KNN), focusing on their use in supervised classification tasks. The traditional methods simply carry out feature extraction from URLs and content of websites, aimed to achieve detection from phishing attempts. An improvement of accuracy was also addressed through feature engineering, preprocessing, and model tuning. The literature review highlighted various challenges for above-named methods like dealing with types of phishing attack vectors, dependence on labeled dataset, inefficiency in zero-day attack detection, as well as the issue of computational complexity and overfitting in ensemble techniques such as RF. The study calls for robust frameworks of clustering and classification to be combined to extend research beyond these limitations.

Mandadi et al. (2022) [8] explored the integration of multiple feature types, including URL-based, domain-based, and HTML/JavaScript-based features, to enhance phishing detection. Using Random Forest and Decision Tree models, the study achieved an accuracy of 87% with the Random Forest model. The highest accuracy achieved was 87% for a random forest model with this dataset of 10,000 URLs (5,000 phishing and 5,000 legitimate). Iframe use, modification of the status bar, and DNS records among others were features that contributed to the appropriate classification. However, despite such a rich feature set, the work did not delve into advanced machine learning models like ensemble or deep learning architectures, which can add some more efficiencies. Further, there was no provision for real-time or behavioral feature integration to further enhance the efficacy of the model against sophisticated phishing methods such as dynamically generated URLs or targeted user attacks.

The study by Otieno et al. (2023) [9] explores phishing URL detection using the Bidirectional Encoder Representations from Transformers (BERT) model. The study emphasizes on the lexical and host-based features of URLs thereby relieving it from the manual feature engineering or domain expertise. The research carried out using a balanced corpus of 70,000 URLs (35,000 phishing and 35,000 benign) from ISCX-URL2016 and PhishTank datasets achieved the accuracy of 96%. Evaluation metrics included precision, recall, and F1-score which showed the model's performance consistency between phishing and benign URLs with the F1-score remaining at 0.96. The study argues that BERT captures the semantic and contextual patterns behind phishing URLs that permit effective classification. However, the authors of this study note that the model is subjected to overfitting during training and needs further exploration into the features that depend solely on URLs. It might also present deployment challenges as the computational intensity associated with BERT might pose difficulty during the real-time application in comparison to other models.

The study by Alnemari and Alshammari (2023) [10] investigates the detection of phishing domains using machine learning models, including Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Decision Trees (DTs), and Random Forests (RFs). They developed a model based on the data set of UCI Phishing Websites containing 11,055 instances and 31 attributes on which the models are assessed on their performance based on standardized metrics of accuracy, precision, recall, and F1 score. The model that performed this task best was random forest, which scored 97.3% in terms of detection accuracy. The obtained results also highlighted feature normalization through MinMax scaling for model performance improvement. The models are static and based on a data set, limiting their generalization to newer phishing approaches. Thus, while RF has a specific detection accuracy, it will not be easy to put it into real-time application as it is highly resource intensive.

The study by Nishitha et al. (2023) [11] focuses on phishing detection using both traditional machine learning models and deep learning techniques. Models such as Logistic Regression (LR), k-Nearest Neighbors (KNN), Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and Convolutional Neural Networks (CNN) were evaluated. The dataset consisted of 549,346 URLs, which were split into legitimate and phishing URLs and processed by tokenization, stemming, and vectorization techniques. CNN achieved an accuracy of 96%, closely followed by Logistic Regression with 94%. The study, meanwhile, underscored the significance of the preprocessing phase to enhance the performance. The current piece of research involves static datasets, which limit a study from adapting to new phishing patterns. At the same time, CNN has achieved high accuracy; however, its computation needs may be too high for real-time applications.

This study by Aldaham et al. (2024) [12] evaluates four machine learning models Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN) and Random Forest (RF) as solution to detection of phishing websites. The evaluation was based on the dataset collected from PhishTank.org with parameters such as URL length, presence of special symbols, HTTPS usage, and domain-related features. Among the machine learning models tested, the Decision Tree achieved the highest validation accuracy of 96.7%, followed by the Random Forest model at 95.75%. The ANN model recorded an accuracy of 86.85%, whereas the SVM model's performance was at 83.8%. The study indicates that preprocessing, normalization, and handling data inconsistencies is vital to improving classification performance. Overfitting in the Decision Tree and ANN model calls for a more efficient regularization method. An unvarying dataset is a hindrance to keep pace with the evolving nature of phishing strategies. Moreover, ANN and RF models are computationally intensive, which restricts their application to real-time situations with limited resource endowments.

The study by Chy (2024) [13] explores machine learning's role in predicting and preventing phishing attacks. The research analyzes supervised algorithms such as Logistic Regression, Decision Trees (DT), Random Forests (RF), and Gradient Boosting, as well as unsupervised methods including clustering and anomaly detection. The features under consideration include URL structure, domain registration details, and SSL/TLS certificates, all geared toward real-time detection capabilities. The Random Forest model is highly effective and provides additional insight regarding adaptive learning to combat evolving phishing strategies. The paper also refers to the concept of embedding such detection systems with web browsers and email services. However, challenges exist with respect to data imbalance, overfitting, and the requirement of large labeled datasets in supervised methods. Real-time detection using machine learning can also suffer bottlenecks as regard complex ensemble methods. Future work will include using deep learning and natural language processing from sophisticated phishing attacks.

4 Current Trends and Innovations

The rapid evolution of phishing techniques has driven significant advancements in phishing detection methodologies. This section reviews current trends and innovations in the field and demonstrates how ML and AI will transform the field. It discusses deep-learning applications, ensemble learning processes, real-time detection techniques, adversarial machine learning, and AI integration with cybersecurity facilities all in the Phishing detection requirements. These advances highlight how increasingly sophisticated and adaptive such systems are becoming in the battles against dynamically evolving threats.

4.1 Use of Deep Learning and Neural Networks

Deep learning has turned into a strong base for phishing detection because it can catch subtle patterns in a range of data sources. With the help of different neural network architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), unstructured data such as URLs, email content, and websites can all be analyzed.

One of the neural networks, Autoencoder Neural Network, consists of two processes: encoding input data into a compacted representation and decoding it back to its original form. It has the same number of input and output neurons while having hidden layers with fewer neurons that force the network to learn a more efficient representation of the data. The model minimizes the reconstruction error between input and output while capturing the most important features of the data.

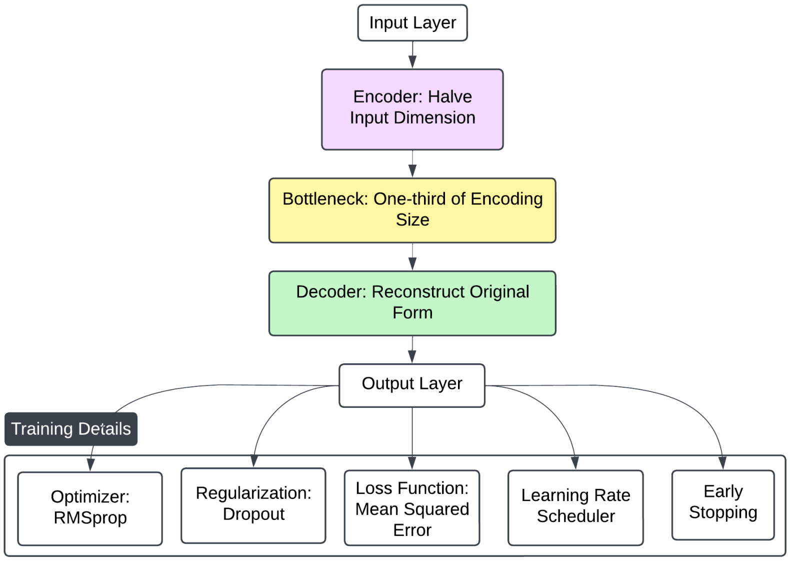
Input Layer
Encoder: Halve Input Dimension
Bottleneck: One-third of Encoding Size
Decoder: Reconstruct Original Form
Output Layer
Training Details
Optimizer: RMSprop
Regularization: Dropout
Loss Function: Mean Squared Error
Learning Rate Scheduler
Early Stopping

Figure 2: Autoencoder Neural Network Architecture of Phishing Detection

ANN can be employed for phishing website detection where the model is trained to reconstruct the input feature set. Figure 2 shows that the encoding layer is designed to halve the input dimension, while the bottleneck reduces it further to one-third of the encoding size. The model is trained with the RMSprop optimizer and a learning rate scheduler to minimize the mean squared error between the input and the reconstructed output. Early stopping is used to avoid overfitting, while dropout layers ensured regularization. The autoencoder identified phishing URLs from legitimate ones by analyzing the reconstruction error, where it capitalized on the learned representation of the data to spot anomalies.

Convolutional neural networks, meanwhile, provide the optimum performance at feature extraction, spotting patterns related to the elements of a webpage: HTML, JavaScript, and graphics. CNN models can find subtle signs of phishing, such as unusual position of buttons and hyperlinks. The temporal identification of phishing signatures is rather accomplished by Long Short-Term Memory RNNs that focus on the URL level by setting traffic flow of URLs in order to emphasize recognition on the headers of the email inside the URLs. They understand the temporal patterns of phishing since these elements are characterized by the progressions of URL segments.

Deep learning is also used in NLP approaches for dealing with the investigation of text content in phishing emails. Deceitful messages may be found by exploiting sentimental discrepancies and tonal inconsistencies, which the NLP models consider analyzing language patterns. Phishing detection systems are increasingly being integrated with pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) to make them state-of-the-art.

4.2 Ensemble Learning Approaches

Numerous ensemble learning approaches combine predictions from models in order to improve not only the robustness of the combined predictions but also their accuracy against phishing attacks. These include Random Forest, Gradient Boosting Machines (GBMs), and Voting Classifiers. Ensemble methods, in fact, are very efficient in tackling issues such as overfitting and leveraging the advantages of different kinds of algorithms.

Both Random Forest and GBMs synthesize the insights from quite a number of decision trees so that they can work effectively on higher dimensional data. Such models integrate features such as URL structures, DNS records analysis, and HTML contents evaluation to examine the possibility of phishing. Bagging and boosting techniques are the basis of ensemble methods recognized for improvement in generalization.

In reality, there are many sectors of the enterprise that rely on models such as ensemble learning for their phishing detection solutions. Using this model, millions of URLs, email attachments, and metadata can be handled at almost real time by several supervised ML models all together, thus ensuring accurate and scalable phishing protection.

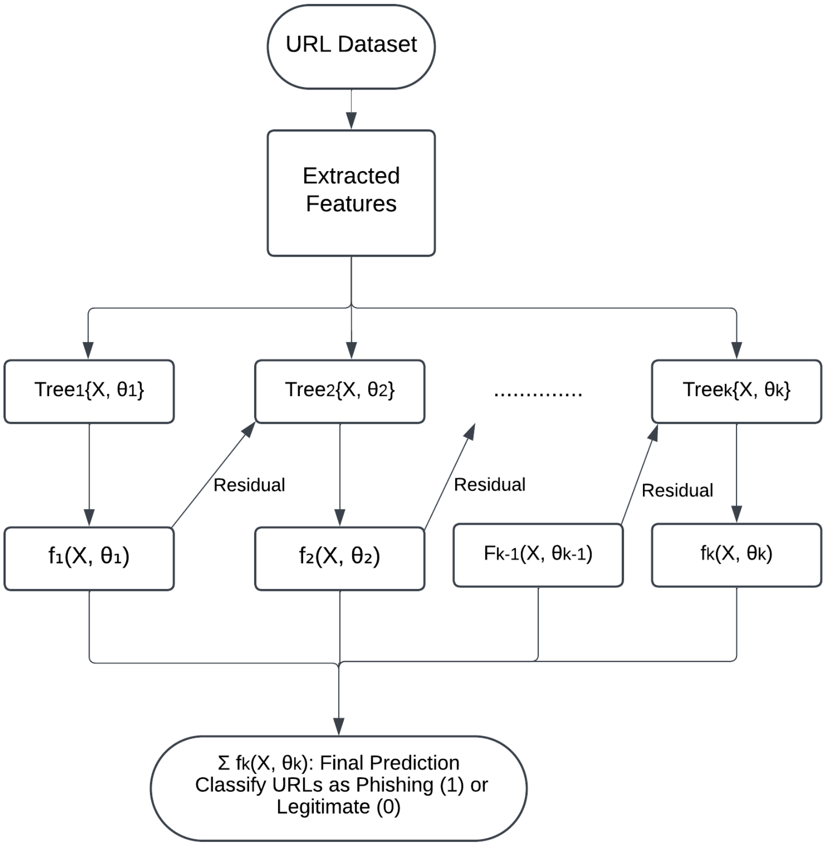

Figure 3: Flowchart of Phishing Detection using XGBoost

One of the most effective ensemble learning techniques is XGBoost, a Gradient Boosting Machine (GBM) optimized for high performance and scalability. As illustrated in Figure 3, XGBoost operates by building a series of decision trees sequentially, where each tree corrects the residual errors of its predecessor. The process starts with a dataset of extracted features (e.g., URL length, presence of special characters, and DNS-based attributes). An initial decision tree is trained, generating predictions $f_1(X, \theta_1)$. The residual errors from this tree are used as input to the next tree, $f_2(X, \theta_2)$, and this process continues until k trees are built. Each tree contributes to minimizing the overall loss function.

The final prediction is obtained by aggregating the predictions of all trees, represented as $\sum f_k(X, \theta_k)$. This cumulative approach ensures that XGBoost captures complex relationships between features while maintaining efficiency. XGBoost also incorporates techniques such as regularization to prevent overfitting and parallel processing to improve computational speed, making it ideal for phishing detection in high-dimensional datasets.

By leveraging the strengths of ensemble learning, particularly XGBoost, phishing detection systems can achieve high accuracy and scalability, effectively addressing the challenges of real-world deployment.

4.3 Real-Time Detection Methods

Real-time phishing detection is critical to thwarting attacks before users are compromised. This requires systems capable of analyzing data streams within milliseconds. Unlike traditional models, real-time detection employs ML models and big data frameworks, including Apache Kafka and Spark Streaming, to provide high-volume and high-velocity data processing.

Real-time systems also analyze some other factors, such as structures, DNS behaviors, and user interactions on websites. So, there are light-weight ML models in browser plugins and security gateways, which can identify the indices of phishing obfuscated URLs, suspicious redirections, and the rest while users access different pages. People's attraction towards using a single and commonly hybridized threat analysis mechanism has also increased with these systems.

Additionally, an online phishing detection system can also be implemented in industries where business is of utmost importance, such as the banking industry because the financial loss from phishing in these institutions is considerable. These fraud detection platforms use machine learning techniques to analyze and flag suspicious login attempts, unauthorized or unusual transactions, or systemic phishing emails from the customers, providing multi-tiered security.

4.4 Adversarial Machine Learning

There is a growing challenge for phishing detection systems posed by adversarial techniques into which an attacker designs input deceptively so that it fools the ML model. Phishing emails or HTML content to evade classifiers, or obfuscated URLs were a few examples of inputs.

The field of adversarial machine learning is concerned with overcoming such a challenge, whereby the introduction of adversarial examples during training makes models robust. For example, adversarial training involves modifying real data to seem like phishing attempts to enable learning to distinguish between authentic and fake input. There are also generic countermeasures like using Generative Adversarial Networks (GAN) to generate an imitation of a realistic phishing attack.

Many organizations have started adopting these approaches as part of their cybersecurity strategy. By continually exposing the models to various evasion attempts, they can maintain their efficacy under advanced threats for these phishing detection systems. This is particularly vital for industries such as e-commerce, primarily because such attacks exploit adversarial techniques for intrusion.

4.5 Integration of AI with Cybersecurity Platforms

Phishing fraud detection has been transformed with the introduction of artificial intelligence in complete cybersecurity platforms that now give total protection to organizations. AI-based platforms merge disparate detecting layers such as URL analysis, email filtering, and behavior monitoring into a unified system.

These systems use machine learning, where AI correlates data from different sources such as user behavior analysis, network logs, and threat intelligence feeds. For instance, AI-based phishing gateways not only stop phishing emails from reaching their destination but also analyze patterns of attachment behavior and redirected URLs. These systems evolve with the changing face of phishing tactics, so their defenses against it continue.

AI-based platforms also employ user behavior analytics solutions to identify anomalies which are characteristic of phishing. Such cases may include unusual access attempts or access patterns and might trigger phishing alerts that would need to be investigated. This proactive approach drastically reduces response times and lessens damage potential.

4.6 Emerging Trends in Automation and Threat Intelligence

By way of automation, phishing detection currently relies less on human operations and can speed up action to threats. Automated systems mainly rely on AI to triage phishing alerts, prioritize critical incidents, and trigger remediation actions. For instance Security Orchestration, Automation, and Response (SOAR) platforms can integrate phishing detection into an automated workflow, thus freeing the security teams to focus on the hard and high-priority tasks.

Threat intelligence company is also innovating on the ideal platform for phishing threats. This platform aggregates all global threat data worldwide using machine learning to find and identify patterns across phishing campaigns. It shapes meaningful outcome evidence from multiple industries and geographies, thus giving organizations actionable insights that can be used to augment organizational defenses.

5 Performance Evaluation

Performance evaluation metrics are crucial for assessing the effectiveness of machine learning models in phishing detection. This section defines common metrics and presents a comparison of various state-of-the-art phishing detection models.

5.1 Accuracy

Accuracy is the most common metric for evaluating classification models. It measures the proportion of correctly classified instances among the total number of instances. It is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Where:

- TP: True Positives (phishing websites correctly classified as phishing),
- TN: True Negatives (legitimate websites correctly classified as legitimate),
- FP: False Positives (legitimate websites incorrectly classified as phishing),
- FN: False Negatives (phishing websites incorrectly classified as legitimate) [14].

5.2 Precision

Precision, also known as Positive Predictive Value, measures the proportion of true positive predictions among all positive predictions. It focuses on the reliability of the positive predictions made by the model. Precision is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

High precision indicates that the model has a low false positive rate, which is crucial for phishing detection to minimize the misclassification of legitimate websites as phishing.

5.3 Recall

Recall, also known as Sensitivity or True Positive Rate, measures the proportion of actual positives that are correctly identified by the model. It is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

High recall ensures that the model effectively identifies most phishing websites, minimizing the risk of overlooking phishing attacks. However, improving recall often comes at the cost of reduced precision.

5.4 F1-Score

The F1-Score is the harmonic mean of Precision and Recall, providing a single metric that balances the trade-off between these two measures. The F1-Score is defined as:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

A higher F1-Score indicates a better balance between Precision and Recall, making it a critical metric for phishing detection.

5.5 Comparison of State-of-the-Art Models

Table 1 summarizes the performance of state-of-the-art phishing detection models, highlighting their algorithms, datasets, challenges, limitations, and achieved accuracy.

Table 1: Comparison of state-of-the-art phishing detection models.

Model or Algorithm	**Type**	**Dataset**	**Challenges**	**Limitations**	**Accuracy**
Grey Wolf Optimizer + SVM[15]	Hybrid	Websites (PhishTank, Yahoo): 1353 instances; 805 phishing URLs, 548 legitimate URLs; 30 rule-based features	Optimizing SVM parameters using nature-inspired optimization techniques	Small dataset; no comparison with other datasets	90.38%
Genetic Algorithm (GA) + DNN[16]	Deep Learning	UCI	Uses GAs for feature selection and weights, improving accuracy	Computationally expensive; limited accuracy compared to other methods	89.50%
Random Forest + Neural Network + Bagging[17]	Hybrid	UCI	No prior research using feedforward NN with ensemble learners for phishing detection	Insufficient data sources and feature extraction process	97.40%
Adam Optimizer + Deep Neural Network (DNN)[4]	Deep Learning	UCI	Large dataset supports semantic analysis of URLs	Requires relevant semantics; fails if URL semantics are absent	96.00%
Convolutional Neural Network (CNN)[18]	Deep Learning	Websites (Alexa, PhishTank, Spamhaus); 157,662 legitimate URLs, 161,016 phishing URLs	Uses feature extraction for character embedding	Limited URL length to 200 characters; training time high	95.02%
Random Forest[19]	Single Model	ISCXURL-2016; 2500 phishing URLs, 2723 legitimate URLs	Achieved high accuracy without third-party services	Small dataset; no evaluation across multiple datasets	99.57%
PSL + PART[20]	Hybrid	Websites (PhishTank, Relbank); 30,500 instances	Extracted 3000 features for comparison across multiple experiments	Limited to e-banking URLs; not generalizable	99.30%
Bootstrap Aggregating + Logistic Model Tree[21]	Hybrid	UCI	Trained on 10-fold cross-validation to reduce bias and variance	Insufficient data sources and feature extraction process	97.42%
ResNeXt-GRU (RNT-J) + Jaya Optimization Algorithm[3]	Hybrid Deep Learning	Multiple datasets; SMOTE applied to handle class imbalance	High accuracy achieved by leveraging robust feature extraction and advanced optimization techniques	Computationally expensive; limited scalability and generalization due to curated datasets	98.00%
BERT[9]	Deep Learning	ISCX-URL2016 and PhishTank datasets; 70,000 URLs (35,000 phishing, 35,000 benign)	Captures semantic and contextual patterns, eliminating manual feature engineering	Computationally intensive; potential overfitting during training; challenges for real-time deployment	96.00%

6 Challenges in Phishing Detection

Phishing detection systems have made significant strides, but they still present several roadblocks before their use in the real world. This section discusses important challenges, with examples that demonstrate their impact.

6.1 Dataset Limitations

6.1.1 Imbalanced Datasets

Fraudulent websites are largely skewed by datasets being used because most authentic samples drown out the number of phishing samples. For example, the typical UCI Phishing Domains dataset is more than 90% URLs which is legitimate compared to phishing URLs [10]. This thus causes classifiers to lean towards a majority class and hence produce a high false negative ratio. Generally, even though SMOTE is one of the techniques used for correcting this imbalance, it often overfits and limits the generalization ability of the model.

6.1.2 Lack of Labeled Data

In supervised learning models, high-quality labeled datasets are severely important, and it is extremely difficult to find such datasets. New and emerging types of phishing attacks cause the whole dataset to be frequently updated, which is both tedious and resource-heavy when it comes to managing the same. For instance, a model trained on older datasets is less capable of identifying the most recent phishing URLs, especially when all real scenarios are ultimately reduced [7]. Automated labeling techniques have been proposed, but they typically cause overheads and introduce errors in the training process.

6.2 Generalization Issues

6.2.1 Difficulty in Detecting Zero-Hour Phishing Attacks

Zero-hour phishing attacks take advantage of new techniques or vulnerabilities that are not represented in training data. For instance, attackers are sometimes found using dynamic URLs or generating AI-craft phishing emails, eluding detection by static feature-based models. Models developed based on old data would typically not have these patterns of new attacks, necessitating adaptive learning approaches [8].

6.2.2 Overfitting to Training Data

Machine learning models, especially the deep learning architectures, overfit rather quickly when trained on very small or cohesive datasets. For example, hybrid CNN-LSTM models can give excellent results on training data, yet they cannot generalize on new phishing samples [6]. This overfitting acts against using those models to effectively detect diverse and evolving techniques in phishing.

6.3 Adversarial Attacks on Machine Learning Models

Phishing detection is increasingly threatened by adversarial attacks. These consist of fake inputs purposely crafted to fool classifiers that include things such as obfuscated URLs and even homoglyph attacks (for example ”g00gle.com” instead of ”google.com”). For instance, phishing detection systems built specifically to detect such disguised URLs are typically unable to detect them since they are too clever. Adversarial training is the introduction of adversarial examples into the training of the model; that is the reason it is said to incur extra computational cost and difficult deployment of the model [6].

6.4 Scalability and Real-Time Processing Challenges

Real-time data for millions of interactions should be processed by these systems in milliseconds to almost instantaneity when countering any attacks. In sequential data, RNNs use deep learning models and find analysis very powerful. However, they require high computational costs and are unable to provide real-time detection. In fact, some financial services report the detection of fraud, which delays detection because of such high latency models, leading to missed opportunities to block phishing attacks [5].

Integrating such systems to be able to scale to millions of interactions a day is also challenging. Organizations in most cases don't have the capabilities to deploy these systems at such scales, all while maintaining low latency and working with other existing cybersecurity tools. [10].

6.5 Privacy Concerns in Feature Collection and Model Deployment

The use of often require the collection and processing of sensitive data such as email contents, user behavior logs, and web browsing history in phishing detections. For example, analyzing content of the e-mail messages to identify phishing may conflict with privacy regulations as GDPR or CCPA [7]. Privacy preserving methods such as federated learning allow decentralized modeling, training without sharing the raw data. However integrating federated learning comes with challenges of communications overhead and keeping data consistent across devices.

Phishing detection challenges at present, such as limited datasets, generalization issues, adversarial attacks, scalability, and privacy issues, have made machine learning models inefficient. The advanced identification of these challenges demands new approaches - adaptive learning, robust adversarial defenses, scalability-supported architecture, and privacy-preserving techniques, without which the future of reliability in systems of phishing detection may not be ensured.

7 Future Directions

These persistent challenges in phishing detection such as dataset restrictions, generalization issues, and privacy concerns require new efforts to realize the models' productivity and efficiency. This section will analyze emerging research directions that researchers and practitioners are taking in the field, with real-life case examples.

7.1 Use of Transfer Learning for Better Generalization

Transfer learning allows a model to be retrained to complete a new related task, thereby reducing the need for acquiring huge amounts of labelled data. In phishing detection, transfer learning can significantly improve generalisation across datasets and attack types. For example, language models such as BERT could be pre-trained and then fine-tuned to target specific phishing email features exploiting their deep insight into language structures [9]. Real-world implementation scenarios would be those found within email security platforms where transfer learning has been involved to learn the differences in patterns particular people use within their communication, thereby improving the detection of advanced multi-lingual phishing emails.

7.2 Synthetic Data Generation to Handle Imbalances

Generative adversarial networks and data augmentation are some artificial data generating techniques addressing the poorly imbalanced datasets. For example, they generate realistic phishing samples. Typically, GANs create approximated URLs, email headers, and website content to realistic phishing attempts to enrich the datasets for training. Most recently, cybersecurity companies began using synthetic data to create modeling for rare phishing occurrences so as not to overfit but to make sure the model can accurately identify behavioral edge cases. The approach can considerably aid in preparing for zero-day attacks for which no real-day examples are available.

7.3 Privacy-Preserving Machine Learning Techniques

Privacy-friendly methods such as federated learning, wherein users would collaborate without sharing raw data, allow joint learning of models. Federated learning includes the aggregation of insights obtained from decentralized datasets and provides compliance with privacy regulations such as GDPR and CCPA. For example, Google has introduced federated learning as a part of its Safe Browsing program to provide enhancements for phishing detection across devices without user internal data compromise. With this approach, models learn from user behavior whereby sensitive data is kept local, thereby ensuring trust and safety.

7.4 Exploration of Hybrid Approaches Combining ML and Rule-Based Methods

Hybrid systems integrate machine learning and rule-based models into either a new or existing general framework, leading to a good compromise on adaptability and interpretability. A rule-based approach is more effective for known specific patterns, while an ML approach is particularly useful against newly developed threats. One example is the case of such hybrid systems for suspicious transactions. They usually take fixed rules, such as very high or low anomaly transaction amounts, and then introduce the ML analysis of contextual data in terms of user behavior to understand the anomaly better. In this way, false positives and false negatives are going down as much as possible in a dual way.

7.5 Improved Feature Extraction Leveraging NLP for Email Content Analysis

Natural Language Processing (NLP) techniques are becoming integral to phishing detection, particularly for analyzing email content and headers. Advanced NLP models, such as GPT-based transformers, enable systems to detect subtle linguistic cues, such as impersonation tactics or persuasive language often used in phishing emails. This action applies such as AI email filters that analyse extracted textual data for its sentiment and design a machine learning view of the intent to better identify possible phishing attempts. Still, it is a typical trend that such systems - more so the ones used in enterprises - are now increasingly being deployed to protect such employees from public targeting in the cases of phishing campaigns.

7.6 Development of Explainable AI (XAI) for Phishing Detection

The conversion of complex machine learning models into explainable AI (XAI), is concerned with the identification of strategies that make the internal workings of such models interpretable. Some examples of XAI techniques are SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). According to these techniques, explanation is provided on how models make predictions, which makes it easier for an organization to trust and refine their systems [22]. For example, a cybersecurity platform integrated with an XAI algorithm would present an analyst with clear reasoning for flagging an email or URL as phishing rather than treat the outcome as a predetermined "black box." This provides better decision-making capabilities while reducing the reliance on the developed systems. This transparency is really useful in certain compliance-heavy sectors, such as healthcare and finance

The foreseeable future of phishing detection seems to be bold, innovative ideas that meet the aforementioned dimensions in regard to existing shortcomings and adapt to changing threats as they emerge. Transfer learning, synthetic data generation, privacy-preserving techniques and hybrid systems help to improve performance and scalability in models. At the same time, the various emerging advancements in NLP, explainable AI and blockchain technology offer further areas for increasing detection accuracy, transparency, and security. Organizations will benefit from adopting these methodologies when maintaining a defense mechanism against sophisticated phishing campaigns.

8 Conclusion

Phishing is indeed a very serious and still very dangerous cyber attack. This requires updating techniques for the detection of phishing attacks for users and organizations. The survey discussed the improvement of phishing detection technologies with machine learning as considering current and future trends of its application, including deep learning, ensemble methods, and real-time detection, while also addressing persistent issues: proper datasets, zero-day vulnerability detection, adversarial vulnerability, privacy issues, among others. The survey also featured an exhaustive performance evaluation of some of the most popular machine learning models, discussing the main evaluation metrics such as accuracy, precision, recall, and F1-score. The comparison of these approaches also showed their usefulness for different situations of phishing detection, presenting their strengths, weaknesses, and real applicability. Emerging solutions such as transfer learning, generation of synthetic data, privacy-preserving schemes, hybrid models, and integration of blockchain seem to promise great enhancements regarding the resolutions of these issues. Explainable Artificial Intelligence (XAI) and Natural Language Processing (NLP) add further gains to the accuracy and reliability of such systems. As such, concerted efforts in research, collaboration between academia and industry, and adaptive and scalable systems are equally necessary for combatting defense against phishing attacks. Creating a psychologically defensive and creative barrier against such threats in cyberspace goes a long way in achieving the complete defeat of the cyber community.

References

1. Alswailem A, Alabdullah B, Alrumayh N, and Alsedrani A. Detecting phishing websites using machine learning. In: *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*. IEEE. 2019:1–6.
2. Alam MN, Sarma D, Lima FF, Saha I, Hossain S, et al. Phishing attacks detection using machine learning approach. In: *2020 third international conference on smart systems and inventive technology (ICSSIT)*. IEEE. 2020:1173–9.
3. Alsubaei FS, Almazroi AA, and Ayub N. Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. IEEE Access 2024.
4. Lakshmi L, Reddy MP, Santhaiah C, and Reddy UJ. Smart phishing detection in web pages using supervised deep learning classification and optimization technique ADAM. Wireless Personal Communications 2021;118:3549–64.
5. Dutta AK. Detecting phishing websites using machine learning technique. PloS one 2021;16:e0258361.
6. Odeh A, Keshta I, and Abdelfattah E. Machine learningtechniquesfor detection of website phishing: A review for promises and challenges. In: *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE. 2021:813–18.
7. Rastogi M, Chhetri A, Singh DK, et al. Survey on detection and prevention of phishing websites using machine learning. In: *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE. 2021:78–82.
8. Mandadi A, Boppana S, Ravella V, and Kavitha R. Phishing website detection using machine learning. In: *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*. IEEE. 2022:1–4.
9. Otieno DO, Abri F, Namin AS, and Jones KS. Detecting Phishing URLs using the BERT Transformer Model. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023:2483–92.
10. Alnemari S and Alshammari M. Detecting phishing domains using machine learning. Applied Sciences 2023;13:4649.
11. Nishitha U, Kandimalla R, Vardhan RMM, and Kumaran U. Phishing detection using machine learning techniques. In: *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*. IEEE. 2023:1–6.

12. Aldaham SSM, Ouda O, and Abd El-Aziz A. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING.

13. Chy MKH. Securing the web: Machine learning's role in predicting and preventing phishing attacks. International Journal of Science and Research Archive 2024;13:10–30574.

14. Vakili M, Ghamsari M, and Rezaei M. Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv preprint arXiv:2001.09636 2020.

15. Anupam S and Kar AK. Phishing website detection using support vector machines and nature-inspired optimization algorithms. Telecommunication Systems 2021;76:17–32.

16. Ali W and Ahmed AA. Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. IET Information Security 2019;13:659–69.

17. Zamir A, Khan HU, Iqbal T, et al. Phishing web site detection using diverse machine learning algorithms. The Electronic Library 2020;38:65–80.

18. Aljofey A, Jiang Q, Qu Q, Huang M, and Niyigena JP. An effective phishing detection model based on character level convolutional neural network from URL. Electronics 2020;9:1514.

19. Gupta BB, Yadav K, Razzak I, Psannis K, Castiglione A, and Chang X. A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment. Computer Communications 2021;175:47–57.

20. Barraclough PA, Fehringer G, and Woodward J. Intelligent cyber-phishing detection for online. computers & security 2021;104:102123.

21. Adeyemo VE, Balogun AO, Mojeed HA, Akande NO, and Adewole KS. Ensemble-based logistic model trees for website phishing detection. In: *Advances in Cyber Security: Second International Conference, ACeS 2020, Penang, Malaysia, December 8-9, 2020, Revised Selected Papers 2*. Springer. 2021:627–41.

22. Hernandes PRG, Floret CP, De Almeida KFC, Da Silva VC, Papa JP, and Da Costa KAP. Phishing detection using URL-based XAI techniques. In: *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2021:1–6.