Phishfinder: Phishing Detection Plugin

Joyal Devassy[1*], Nikhil Ravi[2†], Vijay Bhaskar[3],
Vishal C Varghese[4], Sruthy Suresh[5†]

[1,2,3,4,5*]Dept. of CSE, FISAT, Angamaly, Kochi, 683577, Kerala, India.

*Corresponding author(s). E-mail(s): joyaldevassy14@gmail.com;
Contributing authors: nikhilravi546@gmail.com; kichu56798@gmail.com;
vishalcvarghese22@gmail.com; sruthysuresh@fisat.ac.in;
†These authors contributed equally to this work.

Abstract

Phishing is defined as mimicking a credible company's website aiming to take private information of a user. Based on the IEEE paper, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," the random forest classifier seems to outperform other techniques in detecting phishing sites. These techniques, however, use Python machine learning libraries and thus can't be used in the browser in real-time. The main objective of this project, PhishFinder: A real-time phishing website detection system, is to develop a Chrome browser plugin that detects phishing sites in real-time while the user browses the page.

The detection system utilizes URL analysis as a key component in identifying phishing websites. This involves examining various characteristics of the URL such as its length, presence of suspicious characters, and similarity to known phishing URLs. By analyzing these attributes, PhishFinder can determine the likelihood of a webpage being a phishing site.

One common approach is to make the prediction on a server and then let the plugin contact the server for each page. Unlike the old approach, PhishFinder aims to run the classification in the browser itself using the Random Forest algorithm. The advantage of classifying in the client-side browser includes better privacy (the user's browsing data need not leave their machine) and independence from network latency.

This project mainly involves implementing the above-mentioned paper in JavaScript for it to run as a browser plugin. Since JavaScript doesn't have much ML library support and considering the processing power of client machines, the approach needs to be made lightweight.

Keywords: Phishing detection Real-time detection Browser plugin JavaScript URL analysis Random Forest classifier PhishFinder Machine learning Lightweight

implementation Privacy-preserving Client-side classification Chrome extension
AI-based detection Cybersecurity Web security

1 Introduction

In the realm of cybersecurity, the detection of phishing websites poses a significant challenge due to their deceptive nature and evolving tactics. Phishing, characterized by the impersonation of legitimate entities to acquire sensitive information from unsuspecting users, continues to be a prevalent threat in the digital landscape. Traditional approaches to phishing detection often rely on server-side processing or complex machine learning models, which may not be suitable for real-time detection within web browsers.

This literature review focuses on the development of a novel solution, PhishFinder, a real-time phishing website detection system implemented as a lightweight browser plugin. Inspired by the paper "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," this project leverages URL analysis and the Random Forest classifier to identify phishing websites directly within the client's browser environment. By analyzing key attributes of URLs and employing machine learning techniques, PhishFinder aims to provide users with immediate protection against phishing threats without compromising their privacy or introducing network latency.

The integration of JavaScript and browser plugins enables PhishFinder to offer real-time detection capabilities, empowering users to browse the web with confidence while mitigating the risk of falling victim to phishing attacks. This literature review explores the technical implementation, effectiveness, and implications of PhishFinder in enhancing web security and combating phishing in the modern digital landscape.

2 Problem Statement

Phishing, a prevalent cyber threat, targets sensitive user information through fraudulent means, often via spoofed emails or websites. Traditional methods like directory lookups struggle with the transient nature of phishing sites. Machine learning, particularly the random forest classifier, shows promise in detection. Existing browser plugins rely on external servers for classification, raising privacy concerns and latency issues. Our solution is a Chrome plugin that performs real-time classification locally, preserving user privacy and ensuring instant warnings upon encountering phishing sites. This project aims to empower users with timely alerts without compromising their browsing data, enhancing overall cybersecurity.

3 Objectives

The primary objectives of MANVS can be summarized as follows:

- Develop a Chrome browser plugin for real-time phishing detection
- Ensure the plugin functions independently without relying on external servers.

- Provide instant warnings to users upon encountering phishing websites.
- Preserve user privacy by avoiding transmission of browsing data to external sources.
- Enhance overall cybersecurity by empowering users with timely alerts during web browsing.

4 Proposed Work

The development of the PhishFinder browser extension entails creating a seamless user experience while integrating essential functionalities to detect phishing websites in real-time. This involves crafting a user-friendly interface that seamlessly integrates with popular browsers like Chrome and Firefox. Core functionalities such as URL extraction, querying the Phishtank database, and integrating a pre-trained machine learning model must be implemented to ensure accurate detection. Efficient querying of the Phishtank database is crucial for swiftly determining the legitimacy of accessed websites. To achieve this, mechanisms will be developed to optimize database queries and minimize latency in identifying potential phishing threats. Integration of a pre-trained machine learning model, trained on a dataset of phishing and legitimate URLs, forms the backbone of PhishFinder's detection capabilities. Various classification algorithms such as Random Forest, Decision Tree, or XGBoost will be considered to enhance the robustness and accuracy of the detection process. In terms of user interface design, a clear dialogue box or notification system will be implemented to promptly inform users about the phishing status of visited websites. Additionally, user awareness features will be integrated to provide visual cues or warnings when accessing suspicious URLs, enhancing overall user safety. Compatibility with popular web browsers and seamless integration with their interfaces are paramount to ensure a smooth user experience. Performance optimization measures will also be implemented to minimize resource usage and latency, particularly on devices with limited processing power. Privacy considerations will be prioritized, with measures in place to safeguard users' browsing data throughout the detection process. Comprehensive testing and evaluation will validate the extension's accuracy and effectiveness across various browsing scenarios, with user feedback guiding iterative improvements. Documentation and support resources will be provided to assist users in understanding the extension's features and troubleshooting any issues they encounter. Additionally, channels for user feedback and support will be established to address queries and concerns in a timely manner.

5 Literature review

5.1 PDGAN: Phishing Detection With Generative Adversarial Networks[8]

The paper addresses the pressing issue of phishing attacks, emphasizing the increased demand for high-accuracy detection tools amid the growing prevalence of online services and payment systems. Traditional phishing detection methods relying on webpage content features often result in high false detection rates. The study highlights the emergence of deep learning, particularly generative adversarial networks (GANs),

as a promising approach. The proposed phishing detection model, PDGAN, stands out by solely utilizing a website's uniform resource locator (URL) for reliable performance. PDGAN incorporates a long short-term memory network (LSTM) as a URL generator and a convolutional neural network (CNN) as a discriminator to differentiate between phishing and legitimate URLs. The dataset, comprising nearly two million URLs from PhishTank and DomCop, demonstrates the model's efficacy with a remarkable 97.58% detection accuracy and 98.02% precision. Importantly, PDGAN achieves these results without relying on third-party services, showcasing its superiority over existing state-of-the-art models in terms of accuracy and independence from external sources.
The methodology of the study involves proposing a phishing detection model named PDGAN, which focuses exclusively on a website's uniform resource locator (URL) for accurate and efficient performance. The model employs a dual-network architecture, comprising a long short-term memory network (LSTM) as a generator for synthetic phishing URLs and a convolutional neural network (CNN) as a discriminator to differentiate between phishing and legitimate URLs. The dataset used for training and evaluation consists of nearly two million URLs sourced from PhishTank and DomCop. Unlike traditional phishing detection methods, PDGAN eliminates dependence on third-party services and instead relies solely on the URL itself. The experimental results showcase the model's robustness, achieving an impressive 97.58

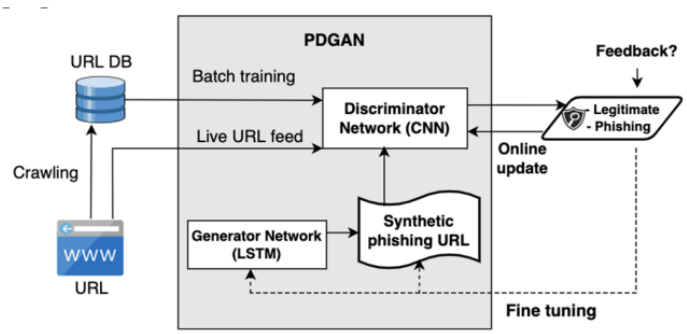


Fig. 1 System architecture of PDGAN

The paper contributes significantly to the development of a phishing website detection extension by introducing a novel approach that achieves high accuracy and precision. PDGAN relies exclusively on a website's uniform resource locator (URL), eliminating the need for third-party services and reducing privacy concerns. The integration of a generative adversarial network (GAN) enhances the extension's capabilities by generating synthetic phishing URLs, introducing diversity to better identify evolving tactics. The use of a long short-term memory network (LSTM) and convolutional neural network (CNN) further enhances the extension's ability to differentiate between phishing and legitimate URLs. Leveraging a large and diverse dataset, the model proves effective in training and validation. Most notably, PDGAN's focus on URL-based detection reduces false positives associated with content-based approaches,

aligning well with the efficiency goals of a browser extension. In summary, the paper's contributions pave the way for an advanced, autonomous, and adaptable phishing detection extension with improved accuracy and reduced dependency on external services.

5.2 PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System[10]

This paper introduces PhishHaven, an innovative ensemble machine learning-based detection system designed to combat phishing attacks generated by both deep neural networks and human attackers. Addressing the emerging threat posed by DeepPhish, a deep neural network-based phishing URL generator, PhishHaven employs lexical analysis for feature extraction. The system introduces URL HTML Encoding for on-the-fly classification and utilizes a URL Hit approach to tackle the challenge of tiny URLs. The final classification is determined through an unbiased voting mechanism, aiming to prevent misclassification in cases of equal votes. To ensure real-time detection, PhishHaven employs a multi-threading approach for parallel execution of ensemble-based machine learning models. Theoretical analysis demonstrates the system's capability to detect tiny URLs and AI-generated phishing URLs with 100% accuracy. Experimental results, based on a benchmark dataset of 100,000 phishing and normal URLs, showcase PhishHaven's impressive 98.00% accuracy, surpassing existing lexical-based phishing detection systems focused on human-crafted URLs.
The proposed methodology in this paper, named PhishHaven, addresses the emerging threat of phishing attacks facilitated by a deep neural network-based phishing URL generator called DeepPhish. To counter both AI-generated and human-crafted phishing URLs, PhishHaven employs an ensemble machine learning-based detection system. The methodology utilizes lexical analysis for feature extraction and introduces innovations such as URL HTML Encoding to enhance lexical analysis and a URL Hit approach to tackle tiny URLs—a previously unresolved challenge. The final classification of URLs is determined through an unbiased voting mechanism to prevent misclassification in case of tied votes. To expedite the ensemble-based machine learning models, PhishHaven incorporates a multi-threading approach for parallel execution, enabling real-time detection. Theoretical analysis asserts the system's ability to consistently detect tiny URLs and achieve 100% accuracy in identifying future AI-generated phishing URLs based on selected lexical features. Experimental evaluation on a benchmark dataset of 100,000 phishing and normal URLs demonstrates PhishHaven's high accuracy of 98.00%, surpassing existing lexical-based detection systems for human-crafted phishing URLs.

This paper makes a significant contribution to the field of phishing website detection by introducing PhishHaven, a novel ensemble machine learning-based system designed to combat both AI-generated and human-crafted phishing URLs. The methodology offers several key contributions. Firstly, it addresses the evolving threat landscape by specifically countering phishing attacks orchestrated through deep neural network-based systems like DeepPhish. Secondly, PhishHaven leverages lexical analysis for feature extraction and introduces innovative techniques such as URL HTML Encoding and URL Hit for more robust detection, especially in handling tiny URLs. The unbiased voting mechanism ensures reliable classification, mitigating the risk of misclassification. Additionally, the multi-threading approach enables real-time detection, enhancing the system's practical applicability. The theoretical analysis underscores its efficacy in consistently detecting tiny URLs and achieving 100% accuracy in identifying future AI-generated phishing URLs. The impressive experimental

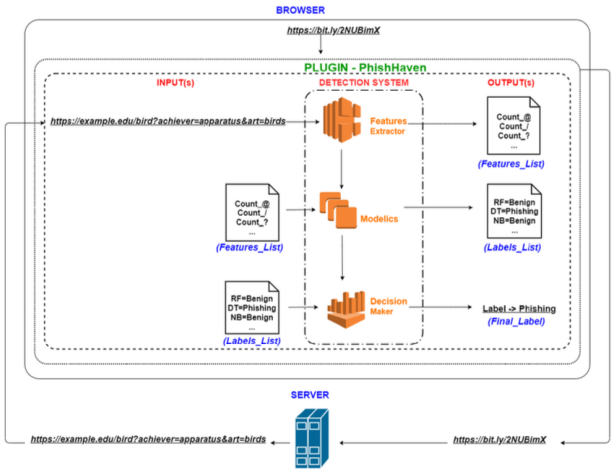


Fig. 2 System architecture of PhisHaven

results, showcasing a 98.00% accuracy on a substantial dataset, position PhishHaven as a notable advancement in the realm of phishing website detection, offering a comprehensive and effective solution for combating diverse phishing threats.

5.3 Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection[5]

This paper addresses the escalating threat of web phishing attacks, which have evolved significantly in recent years, eroding customer trust in e-commerce and online services. Traditional approaches relying on blacklists of known phishing websites struggle to keep pace with the rapid evolution of sophisticated phishing methods, including the emergence of zero-day phishing websites. To overcome these limitations, the paper proposes an intelligent phishing website detection system that leverages particle swarm optimization (PSO) for feature weighting.

The key innovation lies in utilizing PSO to effectively weigh different website features, enhancing the accuracy of phishing website detection. Unlike previous methods that often rely on human experience or frequency analysis to select important features, the proposed PSO-based approach dynamically adjusts the weights of features based on their significance in distinguishing between phishing and legitimate websites.

The experimental results demonstrate the effectiveness of the PSO-based feature weighting method, showcasing substantial improvements in classification accuracy, true positive and negative rates, as well as reductions in false positive and negative rates. Notably, these enhancements are achieved while using a reduced set of website features, underscoring the efficiency and efficacy of the proposed approach in the detection of phishing websites. The study contributes to the ongoing efforts to fortify cybersecurity measures against evolving online threats, providing a promising avenue for the development of more resilient phishing detection systems.

The proposed methodology in this paper introduces an intelligent phishing website detection system that employs particle swarm optimization (PSO) for feature weighting, aiming to enhance the accuracy of phishing website identification. The primary motivation stems from the inadequacies of traditional blacklist-based approaches in keeping pace with the evolving sophistication of web phishing attacks, especially in detecting zero-day phishing websites. The core of the methodology involves utilizing PSO, a nature-inspired optimization algorithm, to dynamically assign weights to various website features. Unlike conventional methods that often rely on human expertise or frequency analysis to determine the importance of features, PSO optimizes the feature weights based on their effectiveness in discriminating between phishing and legitimate websites. The optimization process involves a population of potential solutions (particle swarm) that iteratively adjusts the weights to minimize a predefined objective function, which, in this context, is linked to the accuracy of the phishing detection model. The significance of each feature is thus adaptively determined by the PSO algorithm during the optimization process.

To implement the proposed approach,we collect a dataset comprising instances of both phishing and legitimate websites, incorporating a diverse range of features characterizing each website. These features may include URL structure, HTML content, SSL certificate information, and more. The dataset is then split into training and testing sets, with the training set used to fine-tune the PSO-based feature weights. Machine learning models, such as decision trees or support vector machines, are employed to build the phishing detection system using the optimized feature weights. The performance of the system is evaluated on the testing set, and key metrics such as

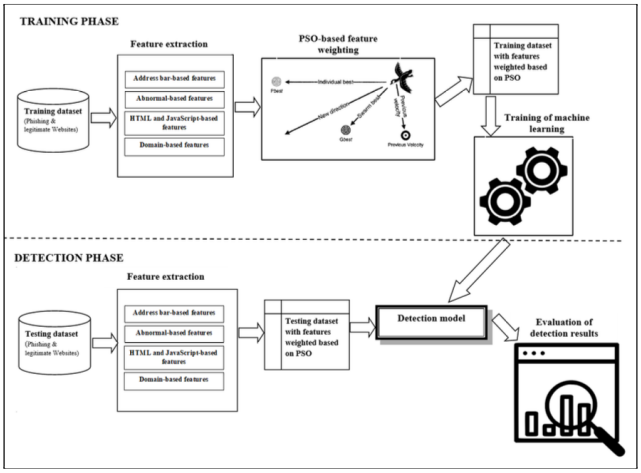


Fig. 3 System architecture of PSO-based feature weighting detection system.

classification accuracy, true positive and negative rates, and false positive and negative rates are analyzed to assess the efficacy of the proposed methodology. The results demonstrate significant improvements in the phishing detection model's performance, validating the effectiveness of the PSO-based feature weighting approach in enhancing the accuracy and efficiency of phishing website detection.

The contributions from this paper offer valuable enhancements to a phishing website detection extension project. By incorporating the intelligent phishing detection methodology utilizing particle swarm optimization (PSO)-based feature weighting, the extension project can significantly improve its accuracy in identifying phishing websites. The dynamic adaptation of feature weights through PSO allows the system to autonomously prioritize and adjust the significance of various website features, ensuring a more resilient defense against evolving phishing tactics, including zero-day threats. This contributes to the extension project's effectiveness in staying ahead of sophisticated phishing techniques that traditional blacklist-based approaches struggle to address. Moreover, the streamlined use of a reduced set of website features, as demonstrated in the paper, not only optimizes computational resources but also enhances the efficiency of the phishing detection system within the extension. The integration of this intelligent methodology provides a valuable layer of defense, reinforcing the extension project's capability to safeguard users from the evolving landscape of online threats, thereby fostering increased trust in e-commerce and online services.

5.4 Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism

This paper addresses the significant and escalating threat of phishing emails, which have led to substantial financial losses. Acknowledging the limitations of current confrontation methods, the authors propose a novel phishing email detection model named THEMIS. The approach involves a detailed analysis of email structures and employs an improved recurrent convolutional neural networks (RCNN) model with multilevel vectors and an attention mechanism.

The THEMIS model simultaneously models various aspects of emails, including the email header, email body, character level, and word level. This comprehensive approach aims to enhance the detection accuracy of phishing emails. The evaluation of THEMIS utilizes an unbalanced dataset with realistic ratios of phishing and legitimate emails. The experimental results demonstrate an impressive overall accuracy of 99.848%, with a notably low false positive rate (FPR) of 0.043%. The high accuracy and low FPR indicate that THEMIS can effectively identify phishing emails with a high probability while minimizing the misclassification of legitimate emails.

The promising results of THEMIS surpass existing detection methods, confirming its effectiveness in detecting phishing emails. The paper emphasizes the need for more robust phishing detection technology to mitigate the growing threat of phishing emails, and THEMIS stands out as a highly accurate and efficient solution in this regard.

The proposed methodology in this paper centers around the development of a novel phishing email detection model named THEMIS. The approach begins with a meticulous analysis of email structures. Subsequently, an advanced recurrent convolutional neural networks (RCNN) model is introduced, enhanced with multilevel vectors and an attention mechanism. This model, THEMIS, is designed to concurrently capture various facets of emails, including the email header, email body, character-level features, and word-level features. By leveraging the strengths of RCNN along with the multilevel vectors and attention mechanism, THEMIS aims to comprehensively model the nuanced patterns and characteristics associated with phishing emails. This holistic approach contributes to the model's effectiveness in discerning between phishing and legitimate emails. The evaluation of THEMIS employs a realistic, unbalanced dataset, and the experimental results showcase an impressive overall accuracy of 99.848%, coupled with a remarkably low false positive rate (FPR) of 0.043%. These outcomes underscore the efficacy of the proposed methodology in enhancing phishing email detection, outperforming existing methods, and providing a promising solution to the escalating threat of phishing.

This paper makes significant contributions to the development of a phishing website detection system, extending its impact on cybersecurity measures. The foremost contribution lies in the introduction of THEMIS, a novel phishing email detection model that exhibits exceptional accuracy and efficiency. By proposing an advanced recurrent convolutional neural networks (RCNN) model enriched with multilevel vectors and an attention mechanism, THEMIS offers a comprehensive solution for modeling various aspects of phishing emails, spanning email headers, bodies, character-level features, and word-level features simultaneously. This methodology is adaptable and can be extended to enhance the detection capabilities of phishing websites within the context

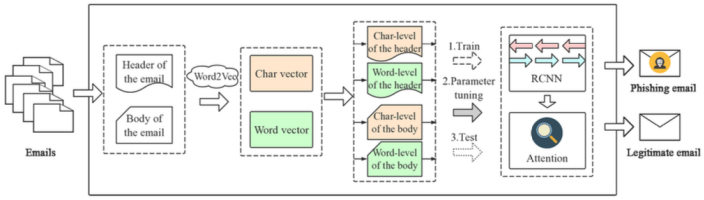


Fig. 4 The framework for classifying phishing emails and legitimate emails in this paper.

of a broader cybersecurity project.

The integration of THEMIS into a phishing website detection system presents a transformative advancement. Phishing websites often leverage email communication for their operations, and by effectively detecting phishing emails, THEMIS contributes directly to the identification and mitigation of associated phishing websites. The model's multilevel and attention-based features allow for a nuanced understanding of phishing email patterns, making it particularly adept at recognizing sophisticated phishing techniques employed by malicious websites. Consequently, the THEMIS-enhanced phishing website detection system benefits from an elevated level of accuracy and reliability in distinguishing between legitimate and malicious web entities.

Furthermore, the paper's emphasis on utilizing a realistic, unbalanced dataset in the evaluation of THEMIS aligns with the challenges faced by real-world cybersecurity systems. This ensures that the proposed model is robust and practical, providing a valuable foundation for the integration into a phishing website detection system. The impressive results, with an overall accuracy of 99.848% and a minimal false positive rate (FPR) of 0.043%, bolster the model's credibility and instill confidence in its applicability to a broader cybersecurity project.

5.5 Phishing URL Detection: A Real-Case Scenario Through Login URLs[7]

In this paper, the authors address the challenge of phishing detection by comparing machine learning and deep learning techniques, specifically focusing on URL analysis. Unlike prevailing methods that define the legitimate class solely based on homepages and exclude login forms, the proposed approach considers URLs from both homepages and login pages for both classes. This adjustment aims to better reflect real-world scenarios and exposes the high false-positive rates associated with existing techniques when confronted with URLs from legitimate login pages. The study also investigates the temporal aspect of model accuracy by training on old datasets and testing on recent URLs, highlighting a decline in performance over time. The authors introduce a new dataset, Phishing Index Login URL (PILU-90K), consisting of 60K legitimate URLs and 30K phishing URLs. A Logistic Regression model, coupled with Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction, achieves an impressive 96.50% accuracy on the presented login URL dataset, underscoring the effectiveness of the proposed method in detecting phishing websites. Additionally, the paper conducts a frequency analysis of current phishing domains, shedding light on various techniques employed by phishers in their campaigns. The methodology proposed in the paper involves a comprehensive approach to phishing detection through the comparison of machine learning and deep learning techniques, with a primary focus on URL analysis. The authors address a limitation in existing solutions, where the legitimate class is typically comprised of homepages, omitting login forms. In contrast, the proposed methodology includes URLs from both homepages and login pages in both classes to better mimic real-world scenarios.

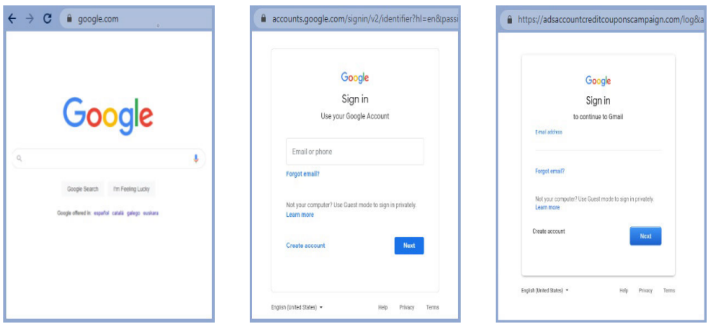
The study begins by highlighting the vulnerability of existing techniques to false positives when tested with URLs from legitimate login pages. This motivates the need for a more representative dataset and a nuanced approach to feature inclusion. To this end, the authors introduce the Phishing Index Login URL (PILU-90K) dataset, containing 60K legitimate URLs encompassing both index and login websites, along with 30K phishing URLs.

Temporal analysis is conducted to explore the evolving nature of phishing attacks over the years. A base model is trained using older datasets and then tested with more recent URLs, revealing a decrease in model accuracy over time. This temporal perspective adds a valuable dimension to the evaluation of phishing detection models.

The authors employ machine learning techniques, specifically Logistic Regression, and deep learning techniques for their comparative analysis. The chosen Logistic Regression model, combined with Term Frequency - Inverse Document Frequency (TF-IDF) feature extraction, emerges as a standout performer, achieving an impressive 96.50% accuracy on the PILU-90K dataset. This underscores the efficacy of the proposed approach in accurately identifying phishing websites.

Furthermore, the paper delves into a frequency analysis of current phishing domains to discern patterns and techniques employed by malicious actors. This provides additional insights into the evolving strategies of phishers in their campaigns.

This paper makes several significant contributions to the field of phishing website



(a) Legitimate homepage (b) Legitimate login page (c) Phishing web page

Fig. 5 A demonstration

detection, providing valuable insights and methodologies that can enhance the effectiveness of detection projects.

Firstly, the adjustment in defining the legitimate class by incorporating URLs from both homepages and login pages represents a crucial improvement. By considering login forms in the legitimate class, the proposed methodology aligns more closely with real-world scenarios, where phishing attacks often target login credentials. This modification addresses a common limitation in existing detection methods, which tend to exhibit high false-positive rates when confronted with URLs from legitimate login pages. As a result, the proposed approach offers a more robust and realistic representation of the phishing landscape.

The introduction of the Phishing Index Login URL (PILU-90K) dataset is another significant contribution. This dataset, comprising 60K legitimate URLs (including index and login websites) and 30K phishing URLs, provides a comprehensive and diverse set of examples for training and testing detection models. The dataset not only supports the evaluation of the proposed methodology but also serves as a valuable resource for future research and benchmarking in the field of phishing detection.

The temporal analysis conducted in the paper adds a novel dimension to phishing detection projects. By investigating how models trained on older datasets perform when tested with more recent URLs, the study highlights the evolving nature of phishing attacks over time. This temporal perspective is crucial for developing detection models that can adapt to emerging threats and maintain accuracy in the face of evolving phishing tactics.

Lastly, the demonstrated success of the Logistic Regression model combined with TF-IDF feature extraction, achieving an accuracy of 96.50% on the PILU-90K dataset, offers a practical and effective solution for phishing detection projects.

5.6 SPWalk: Similar Property Oriented Feature Learning for Phishing Detection[9]

The paper introduces SPWalk, an innovative unsupervised feature learning algorithm designed for phishing detection, addressing the challenge posed by recent phishing attacks that closely mimic the visual and functional aspects of legitimate webpages.

SPWalk leverages a weblink network representation of webpages, where nodes represent webpages and edges signify reference relationships through hyperlinks or similar textual content. The algorithm employs network embedding techniques, specifically a biased random walk procedure, to map nodes into a low-dimensional vector space.
Key to SPWalk's effectiveness and robustness are three critical points. First, phishing attackers lack complete control over reference relationships, providing an intrinsic difficulty for them. Second, the structural regularities emerging from diverse reference relationships can be exploited to discern between phishing and legitimate webpages. Lastly, incorporating node URL information enhances the suitability of learned node representations for phishing detection.
The proposed algorithm is evaluated through experiments where nodes serve as numeric features for classifying webpages as legitimate or phishing. SPWalk demonstrates superior performance compared to state-of-the-art techniques in phishing detection, particularly excelling in precision with results consistently exceeding 95%. Notably, even in scenarios where phishing webpages are adeptly disguised to evade detection, SPWalk exhibits superior classification efficacy. This highlights the algorithm's ability to effectively leverage structural information and URL details to discern phishing attempts, making it a promising and robust solution for the crucial task of detecting phishing webpages.
The proposed methodology in the paper introduces SPWalk, an unsupervised feature learning algorithm tailored for the detection of phishing webpages, which are increasingly adept at resembling legitimate sites. The approach relies on the construction of a weblink network, where nodes represent webpages, and edges signify reference relationships established through hyperlinks or similar textual content.
The process begins with the creation of the weblink network, capturing the intricate relationships between webpages. Notably, the edges in this network represent reference relationships that connect webpages, incorporating both hyperlink connections and shared textual content. This comprehensive network structure aims to encapsulate the complex interactions among webpages, providing a holistic representation of the web environment.
SPWalk then employs a network embedding technique, specifically a biased random walk procedure, to map the nodes (webpages) into a low-dimensional vector space. This embedding process efficiently integrates both structural information derived from the weblink network and URL information associated with each node. The biased random walk considers the diverse reference relationships, contributing to the robustness of the learned representations.
Three key aspects contribute to the effectiveness and robustness of SPWalk. Firstly, phishing attackers have limited control over reference relationships, introducing inherent difficulty for them in mimicking legitimate webpages. Secondly, the structural regularities arising from diverse reference relationships are leveraged to discriminate between phishing and legitimate webpages. Finally, the incorporation of node URL information enhances the suitability of the learned node representations for the specific task of phishing detection.
In the evaluation phase, numeric features derived from the nodes are utilized for webpage classification as either legitimate or phishing. The experiments demonstrate the

superior performance of SPWalk compared to state-of-the-art techniques, particularly excelling in precision with results consistently exceeding 95%. This underscores the efficacy of SPWalk in effectively leveraging both structural information and URL details to discern phishing attempts, even in cases where attackers employ sophisticated techniques to evade detection. Overall, the proposed methodology presents a holistic and robust approach to phishing detection in the evolving landscape of web-based security threats.

The proposed SPWalk algorithm makes significant contributions to the field of phishing website detection by addressing the challenges posed by the increasing sophistication of phishing attacks. Firstly, SPWalk introduces an innovative approach by leveraging weblink network construction, where nodes represent webpages and edges capture reference relationships through hyperlinks or similar textual content. This unique representation allows for a nuanced understanding of the structural intricacies between phishing and legitimate webpages.

Secondly, the algorithm employs a network embedding technique during a biased random walk procedure, efficiently integrating both structural information and URL details of each node. This integration enhances the robustness of feature extraction, making SPWalk adept at discerning phishing webpages from legitimate ones. The reliance on diverse reference relationships adds a layer of complexity that reflects real-world scenarios, where attackers may not have full control over these relationships.

Lastly, by utilizing node URL information as numeric features, SPWalk enhances the learning of node representations, making them particularly well-suited for phishing detection. The experimental results demonstrate the algorithm's superiority over state-of-the-art techniques, showcasing its precision, especially exceeding 95%. Even when phishing webpages are adeptly camouflaged, SPWalk consistently exhibits better classification efficacy. In summary, SPWalk's contributions lie in its novel network-based approach, effective integration of structural and URL information, and its ability to outperform existing techniques in the challenging task of phishing website detection.

Algorithm 1 Feature Learning

Require: graph $G(V, E, W)$
neighborhood size k
embedding size d
walks per node r
walk length l
Ensure: matrix of node representations $f \in R^{|V| \times d}$;

$T = Preprocess(G, sm_{v,x})$
$G' = (V, E, T)$
Initialize walks to empty
for iter = 1 to r **do**
    O = Shuffle(V)
    for each $v_i \in O$ **do**
        $walk = SPWalk(G', v_i, l)$
        Append walk to walks
    end for
end for
$f = StochasticGradientDescent(k, d, walks)$

Algorithm 2 SPWalk

Require: graph $G'(V, E, T)$
start node v_i
walk length l
Ensure: walk

Initialize walk to $[v_i]$
for step = 1 to l **do**
    $V_{curr} = walk$[step-1]
    N_{curr} = GetNeighbors($curr$, G')
    x = AliasSample($Ncurr$, T)
    Append x to walk
end for

5.7 The Answer Is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering[6]

This project proposes a comprehensive multi-stage approach for detecting phishing email attacks, leveraging natural language processing (NLP) and machine learning techniques. The multi-stage methodology involves feature engineering, lemmatization, feature selection, and extraction, along with improved learning techniques such as resampling and cross-validation, and hyperparameter configuration. Two methods are presented: the first utilizes Chi-Square statistics and Mutual Information to enhance dimensionality reduction, while the second combines Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA).

Both methods address challenges like the 'curse of dimensionality,' sparsity, and information extraction from the Vector Space Model (VSM) representation. The reduced feature sets obtained from these methods, when coupled with XGBoost and Random Forest machine learning algorithms, consistently achieve a remarkable F1-measure of 100% success rate in validation tests using the SpamAssassin Public Corpus and the Nazario Phishing Corpus datasets. Notably, even when considering only the text within email bodies, the proposed approach outperforms state-of-the-art schemes for accredited datasets, demonstrating higher accuracy with a smaller number of features and lower computational cost. This signifies the effectiveness and efficiency of the multi-stage phishing detection approach in countering email-based phishing threats.

The proposed methodology in this paper introduces a multi-stage approach for the detection of phishing email attacks, integrating advanced techniques from natural language processing (NLP) and machine learning. The entire process is designed to enhance feature representation and classification accuracy, addressing key challenges in phishing detection.

The first stage involves feature engineering within the realm of natural language processing. This includes techniques to preprocess and transform the textual content of emails. Lemmatization is employed to reduce words to their base or root form, aiding in the normalization of the text data. Subsequently, feature selection techniques are applied to identify and retain the most relevant features, contributing to the reduction of dimensionality and mitigating the 'curse of dimensionality.'

The second stage introduces two distinct methods for dimensionality reduction. The first method utilizes Chi-Square statistics and Mutual Information to refine the feature set, emphasizing the extraction of salient information while addressing sparsity issues inherent in the data. The second method combines Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA), leveraging their capabilities to capture latent semantic structures in the textual data.

In the third stage, machine learning techniques are employed for model training and classification. XGBoost and Random Forest, two powerful algorithms, are utilized to build robust models capable of distinguishing between phishing and legitimate emails. The models are fine-tuned through hyperparameter configuration to optimize their performance.

The final stage focuses on evaluating and validating the proposed approach. Cross-validation techniques are applied to ensure the generalizability of the models, and resampling methods are employed to address potential imbalances in the dataset. The

effectiveness of the approach is assessed using two widely recognized datasets: the SpamAssassin Public Corpus and the Nazario Phishing Corpus.
The results showcase the remarkable success of the proposed multi-stage methodology, achieving a perfect F1-measure of 100% in phishing detection. This indicates the robustness and efficiency of the approach in accurately identifying phishing emails while maintaining a reduced feature set and computational cost.

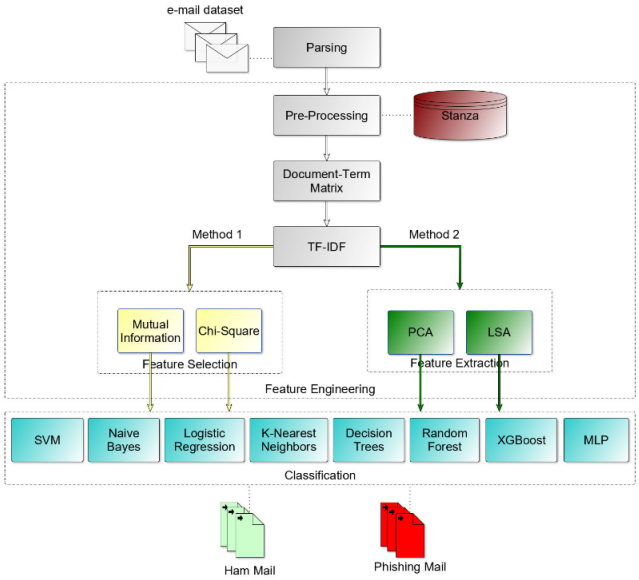


Fig. 6 The main architecture of the proposed Multi-Stage approach for Phishing Detection.

The paper makes several significant contributions to phishing website detection systems through its proposed multi-stage approach, leveraging natural language processing (NLP) and machine learning techniques.
Firstly, the paper addresses the critical issue of phishing attacks by focusing on email communication, acknowledging that emails are a common vector for phishing attempts. By employing advanced techniques in NLP, the proposed approach engages in feature engineering, lemmatization, and feature selection to extract highly representational features from the text of phishing emails. This is crucial as it enables the creation of robust machine learning models capable of distinguishing between phishing and legitimate messages.
The multi-stage approach further contributes by introducing two distinct methods. The first method utilizes Chi-Square statistics and Mutual Information to enhance dimensionality reduction, addressing the challenges of the 'curse of dimensionality,' sparsity, and information extraction from the context in the Vector Space Model. The second method combines Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA), offering an alternative approach to tackle the same challenges.

Additionally, the paper showcases the efficacy of these methods by achieving a remarkable F1-measure of 100% success rate during validation tests using the SpamAssassin Public Corpus and the Nazario Phishing Corpus datasets. This demonstrates the practical effectiveness of the proposed approach in accurately identifying phishing emails, even outperforming existing state-of-the-art schemes. Importantly, the approach achieves these results with a reduced number of features, indicating efficiency and a lower computational cost, which is crucial for real-time phishing detection systems. In summary, the contributions of this paper lie in its innovative and effective methods for phishing email detection, providing a valuable advancement in the field of cybersecurity.

5.8 A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites[3]

This paper addresses the persistent threat of phishing attacks, which involve deceptive tactics to obtain personal information. To counteract this, the authors propose a sophisticated approach: a boosting-based multi-layer stacked ensemble learning model that incorporates a hybrid feature selection technique to identify relevant features for classification.

The proposed model operates in multiple layers, with each layer employing various classifiers. The dataset, preprocessed with the hybrid feature selection technique, is sequentially fed into these classifiers. Importantly, predictions from lower layers serve as input for upper layers, enhancing the phishing detection process.

Through extensive experimental analysis, the authors report impressive results. The proposed model consistently achieves high accuracy, ranging from 96.16% to 98.95% without feature selection and 96.18% to 98.80% with feature selection across diverse datasets. This indicates the model's robust performance across different scenarios.

In comparison to baseline models, the proposed model stands out significantly, outperforming existing approaches. The project demonstrates the effectiveness of the boosting-based multi-layer stacked ensemble learning model, emphasizing its potential as a powerful tool in the ongoing battle against phishing attacks.

The proposed methodology in this paper introduces a comprehensive strategy for combating phishing attacks through a boosting-based multi-layer stacked ensemble learning model. The approach begins with the utilization of a hybrid feature selection technique to identify and extract pertinent features from the dataset, enhancing the model's efficiency in distinguishing relevant patterns associated with phishing. This initial preprocessing step is crucial for optimizing the model's performance.

The model architecture consists of multiple layers, each employing different classifiers. The preprocessed dataset is sequentially passed through these layers, and the predictions from lower layers serve as input for subsequent higher layers. This hierarchical arrangement allows for a more nuanced and sophisticated analysis, with each layer building upon the insights gained from the preceding one. The ensemble nature of the model, incorporating diverse classifiers at each layer, contributes to a robust and comprehensive phishing detection mechanism.

The boosting technique is a key element in the model's learning process, as it focuses on improving the performance of weak classifiers by assigning higher weights to misclassified instances. This adaptive learning approach enhances the model's overall accuracy and effectiveness in discerning phishing attempts.

The experimental analysis conducted on various datasets demonstrates the proposed model's remarkable performance. Without feature selection, the model consistently achieves accuracy levels ranging from 96.16% to 98.95%, showcasing its robustness. Even with feature selection, the accuracy remains high, ranging from 96.18% to 98.80%. The comparison with baseline models highlights the significant superiority of the proposed methodology, signifying its potential as an advanced and effective tool in the ongoing battle against phishing threats. Overall, the paper's methodology

combines feature selection, boosting, and ensemble learning to create a powerful and adaptive system for phishing detection.

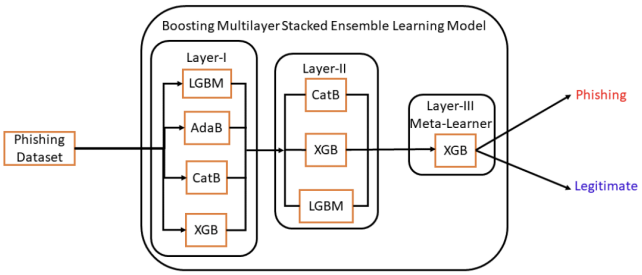


FIGURE 4. Architecture of boosting based multi-layer stacked ensemble learning model.

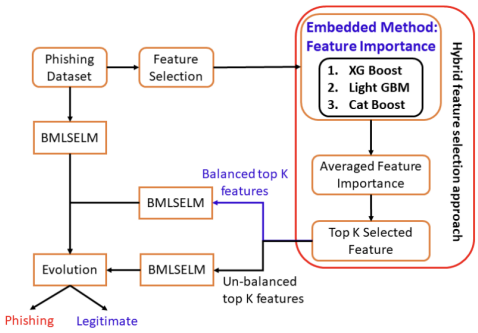


FIGURE 5. Different phases in boosting based multi-layer stacked ensemble learning model.

This paper makes substantial contributions to the field of phishing website detection by proposing an advanced and effective model that significantly enhances the accuracy and reliability of detection systems. One major contribution lies in the introduction of a boosting-based multi-layer stacked ensemble learning model. This model's innovative architecture, employing multiple layers of diverse classifiers, allows for a more intricate and nuanced analysis of phishing patterns. The hierarchical structure, where predictions from lower layers inform the decision-making process in higher layers, enhances the model's ability to discern subtle and complex phishing strategies.
Another significant contribution is the incorporation of a hybrid feature selection technique in the preprocessing phase. This technique optimizes the selection of relevant features from the dataset, thereby improving the model's efficiency and reducing the risk of overfitting. The careful curation of features ensures that the model focuses on the most discriminative aspects of phishing attempts, enhancing its overall performance.
The experimental results presented in the paper underscore the practical impact of these contributions. The proposed model consistently achieves high accuracy levels ranging from 96.16% to 98.95%, showcasing its robustness across diverse datasets. Even with feature selection, the accuracy remains impressive, ranging from 96.18% to 98.80%. The comparison with baseline models further highlights the superiority of

the proposed approach, indicating its potential to significantly advance the capabilities of phishing website detection systems. In summary, the paper's contributions lie in the novel model architecture, the integration of boosting and ensemble learning, and the effective use of feature selection to create a sophisticated and high-performing phishing detection system.

5.9 Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning[2]

The paper proposes an innovative phishing webpage detection model aimed at addressing challenges in current detection methods, particularly the complexity of manual feature collection and the potential correlation issues between features. The model employs representation learning to automatically derive representations from multiple aspects, including URL, HTML page content, and DOM structure. These representations are then fed into a hybrid deep learning network consisting of a convolutional neural network (CNN) and a bidirectional long short-term memory network (Bi-LSTM). The use of an attention mechanism enhances the influence of critical features. The final classification prediction is obtained by fusing the outputs from multiple channels. The model is validated through four experiments, demonstrating superior classification effectiveness compared to existing classic phishing webpage detection methods. The proposed approach achieves a notable accuracy of 99.05% and an impressively low false positive rate of 0.25%. The study establishes that the strategy of extracting webpage features from various aspects through representation learning and a hybrid deep learning network significantly enhances the efficacy of phishing webpage detection.

The proposed methodology in this paper presents a novel and comprehensive approach to enhancing phishing website detection through the introduction of a boosting-based multi-layer stacked ensemble learning model. The methodology begins with a crucial preprocessing step, where a hybrid feature selection technique is employed to meticulously identify and extract relevant features from the dataset. This step ensures that the model focuses on the most discriminative aspects of phishing attempts, improving its efficiency and reducing the risk of overfitting.

The core of the model lies in its multi-layer stacked ensemble architecture. The dataset, preprocessed with the selected features, is sequentially passed through different layers, each equipped with various classifiers. What sets this approach apart is the interplay between the layers, where predictions from lower layers are used as input for higher layers. This hierarchical arrangement enables the model to capture and build upon the intricate patterns associated with phishing attacks. The boosting technique is strategically integrated into the learning process, enhancing the performance of weak classifiers by assigning higher weights to misclassified instances. This adaptive learning approach contributes to the model's overall accuracy and effectiveness in discerning phishing attempts.

The experimental analysis conducted on diverse datasets demonstrates the robustness of the proposed methodology. Without feature selection, the model consistently achieves high accuracy, ranging from 96.16% to 98.95%, across different datasets. Even with feature selection, the accuracy remains impressive, ranging from 96.18% to 98.80%. This consistent high performance underscores the effectiveness of the model in detecting phishing websites. Moreover, the comparison with baseline models reveals a significant improvement, highlighting the superiority of the proposed methodology over existing approaches.

The contributions of this paper to the field of phishing website detection are multifaceted and significant. Firstly, the introduction of a boosting-based multi-layer

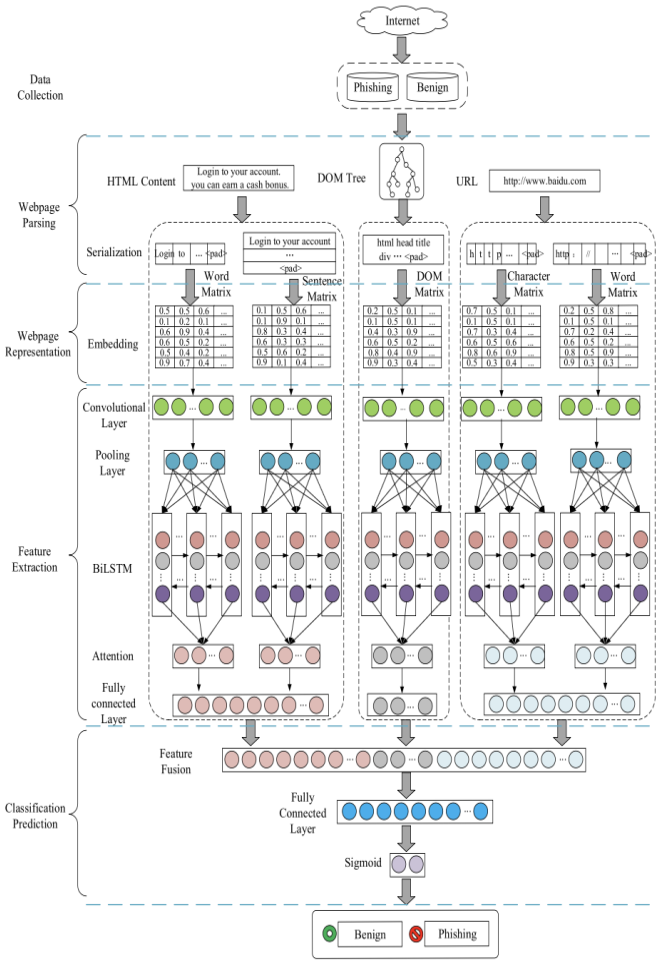


Fig. 7 Model Framework.

stacked ensemble learning model represents a novel and sophisticated approach. This model's architecture, featuring multiple layers with diverse classifiers, facilitates a more nuanced analysis of phishing patterns. The sequential processing of the dataset through these layers, with predictions from lower layers informing higher layers, enhances the model's ability to discern intricate phishing strategies. This contributes to an elevated level of accuracy in identifying malicious websites.

A second major contribution is the incorporation of a hybrid feature selection technique during the preprocessing phase. By carefully selecting and extracting relevant features from the dataset, this technique optimizes the model's focus on discriminative aspects of phishing attempts. This not only improves the efficiency of the model but also mitigates the risk of overfitting, making the detection system more robust and generalizable across diverse datasets.

The experimental results presented in the paper serve as a third contribution, demonstrating the practical impact of the proposed methodology. The consistently high accuracy levels ranging from 96.16% to 98.95%, even across different datasets, underscore the effectiveness of the model. Furthermore, the comparison with baseline models reveals a significant improvement, showcasing the superiority of the proposed approach over existing methods.

5.10 Combining Long-Term Recurrent Convolutional and Graph Convolutional Networks to Detect Phishing Sites Using URL and HTML[1]

The paper proposes an innovative phishing webpage detection model aimed at addressing challenges in current detection methods, particularly the complexity of manual feature collection and the potential correlation issues between features. The model employs representation learning to automatically derive representations from multiple aspects, including URL, HTML page content, and DOM structure. These representations are then fed into a hybrid deep learning network consisting of a convolutional neural network (CNN) and a bidirectional long short-term memory network (Bi-LSTM). The use of an attention mechanism enhances the influence of critical features. The final classification prediction is obtained by fusing the outputs from multiple channels. The model is validated through four experiments, demonstrating superior classification effectiveness compared to existing classic phishing webpage detection methods. The proposed approach achieves a notable accuracy of 99.05% and an impressively low false positive rate of 0.25%. The study establishes that the strategy of extracting webpage features from various aspects through representation learning and a hybrid deep learning network significantly enhances the efficacy of phishing webpage detection.

The proposed anti-phishing model, PhishDet, introduces a novel approach leveraging Long-term Recurrent Convolutional Network (LRCN) and Graph Convolutional Network (GCN) for detecting phishing websites. The methodology integrates URL and HTML features, capitalizing on the capabilities of deep learning techniques to enhance anti-phishing tools' detection efficacy.

PhishDet stands out by incorporating Graph Neural Network (GNN) technology, specifically GCN, a powerful analytical tool in the anti-phishing domain. The model achieved a notable 96.42% detection accuracy and a low false-negative rate of 0.036, showcasing its effectiveness in identifying phishing websites. Notably, PhishDet is designed to combat zero-day attacks, addressing emerging threats in real-time. The system's average detection time of 1.8 seconds further emphasizes its practicality for swift responses.

The automatic feature selection process within PhishDet is a key strength, as the model progressively learns URL and HTML content features. This adaptive learning capability is essential for handling the constantly evolving nature of phishing attacks. The methodology's effectiveness is underscored by its superior performance, achieving a remarkable 99.53% F1-score when evaluated against a public benchmark dataset.

However, PhishDet acknowledges the necessity for periodic retraining to maintain its performance over time. Despite this requirement, the model demonstrates robustness in its ability to consistently outperform similar solutions. If facilitated effectively, the periodic retraining process ensures that PhishDet remains a formidable tool in the ongoing battle against phishing threats, offering extended protection for Internet users. Overall, the proposed methodology showcases a sophisticated integration of LRCN and GCN, demonstrating promising results in the detection of phishing websites and emphasizing adaptability to evolving cyber threats.

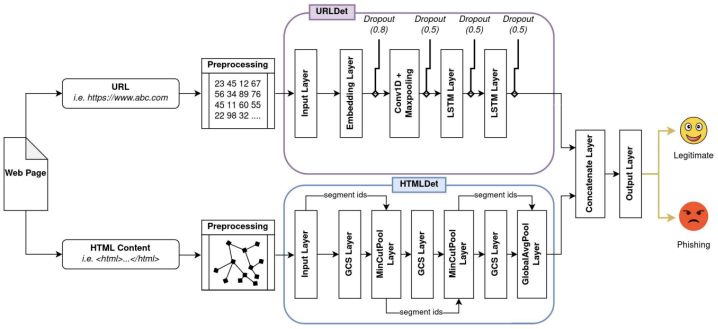


Fig. 8 The architecture of the PhishDet model

The contributions of this paper, PhishDet, to the field of phishing website detection are substantial and transformative. Firstly, the integration of Long-term Recurrent Convolutional Network (LRCN) and Graph Convolutional Network (GCN) architectures represents a novel and effective approach. By combining these two powerful neural network models, PhishDet comprehensively analyzes both the sequential data from URLs and the relational patterns within HTML content. This innovative combination enhances the model's ability to capture the temporal dependencies inherent in phishing attacks and extract semantic information from webpage structures.

Secondly, PhishDet introduces the groundbreaking utilization of Graph Neural Network techniques in the anti-phishing domain. The incorporation of GCN allows the model to leverage graph-structured data analysis, providing a more holistic understanding of the complex relationships within HTML content. This not only improves the accuracy of phishing detection but also signifies a pioneering contribution to the advancement of anti-phishing tools.

Furthermore, PhishDet's automatic feature selection mechanism is a notable contribution. The model dynamically learns and refines URL and HTML content features, enabling adaptability to evolving phishing strategies. This autonomous feature selection enhances the system's robustness and ensures its effectiveness against emerging threats.

The impressive performance metrics of PhishDet, including a 96.42% detection accuracy, a low false-negative rate of 0.036, and the ability to counter zero-day attacks with an average detection time of 1.8 seconds, underscore its practical impact. Finally, while the paper acknowledges the need for periodic retraining, PhishDet's achievements, particularly with a 99.53% f1-score on a public benchmark dataset, showcase its superiority over existing solutions. Overall, PhishDet significantly advances the capabilities of phishing website detection systems, offering a novel framework with improved accuracy and adaptability to counter evolving cyber threats.

5.11 Comparison Table

Paper Name	Model Used	Accuracy	Features & Novelty
Long-Term Recurrent Convolutional and Graph Convolutional Networks[1]	Long-term Recurrent Convolutional Network and Graph Convolutional Network	96.42%	Uses URL and HTML features. Claims to be the first to use Graph Neural Network in anti-phishing with high accuracy. Requires periodic retraining.
Web2Vec: Phishing Webpage Detection Method Based on Multi-Dimensional Features Driven by Deep Learning [2]	Hybrid deep learning network with CNN and bidirectional LSTM	99.05%	Utilizes URL, HTML page content, and DOM structure. Emphasizes representation learning from multi-aspect features.
A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites[3]	Boosting based multi-layer stacked ensemble learning	96.16%-98.95% without feature selection, and 96.18%-98.80% with feature selection	Hybrid feature selection technique. Uses boosting and ensemble learning for phishing detection.
Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism [4]	Improved recurrent convolutional neural networks (RCNN) with attention mechanism	99.848% overall accuracy with a low false positive rate (FPR) of 0.043%	Considers email header, body, character level, and word level simultaneously. Uses multilevel vectors and attention mechanism in RCNN.
Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection [5]	Particle Swarm Optimization (PSO)-based feature weighting	Outstanding improvements in classification accuracy	Uses PSO to weight various website features effectively. Focuses on intelligent phishing website detection using PSO-based feature weighting.

Table 1 Comparison of Phishing Detection Models

Paper Name	Model Used	Accuracy	Features & Novelty
The Answer Is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering[6]	Logistic Regression with TF-IDF feature extraction	96.50% accuracy on the introduced login URL dataset	Uses URLs from login pages in both classes for training. Considers the inclusion of login pages in the legitimate class.
Phishing URL Detection: A Real-Case Scenario Through Login URLs[7]	XGBoost and Random Forest with multi-stage approach	F1-measure of 100% success rate	Involves feature engineering, lemmatization, and feature extraction within natural language processing. Proposes a multi-stage approach with Chi-Square statistics, Mutual Information, PCA, and LSA.
PDGAN: Phishing Detection With Generative Adversarial Networks[8]	PDGAN model using GAN, LSTM, and CNN	97.58% detection accuracy	Depends only on the website's URL. Proposes a phishing detection model based on GAN without relying on webpage content.
SPWalk: Similar Property Oriented Feature Learning for Phishing Detection[9]	SPWalk - an unsupervised feature learning algorithm	Demonstrates superiority over state-of-the-art techniques, especially in terms of precision	Constructs a weblink network with structural information between nodes and URL information. Focuses on a network embedding technique for feature extraction.
PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System[10]	Ensemble machine learning-based detection system	98.00% accuracy, outperforming existing lexical-based systems	Uses lexical analysis, URL HTML Encoding, and URL Hit approach. First study to consider detecting phishing attacks by both AI and human attackers.

Table 2 Comparison of Phishing Detection Models

6 Conclusion

The PhishFinder project is a groundbreaking initiative in online security, utilizing ad- vanced machine learning techniques for swift and efficient detection of phishing attempts. Its primary goal is to deliver a precise and user-friendly system that addresses the dy- namic nature of the digital landscape. The project emphasizes speed and accuracy in phishing threat detection, employing a dual-method approach with immediate Phishtank database checks and comprehensive feature extraction and model-based analysis. The system prioritizes user convenience with an intuitive interface, recognizing the need for ease of use in navigating online platforms. PhishFinder contributes significantly to cybersecurity by safeguarding individuals and organizations from the growing sophistica- tion of phishing attacks. The project's iterative development, informed by user feedback, highlights its adaptability and commitment to continuous improvement. Privacy and transparency measures have been implemented to ensure user data secu- rity, enhancing the system's overall reliability and trustworthiness. PhishFinder not only represents a technological advancement but also serves as a proactive response to the es- calating challenges posed by cyber threats. By providing a real-time defense mechanism against phishing, the project marks a significant stride toward fostering a more secure and resilient online environment globally.

7 References

[1] S. Ariyadasa, S. Fernando and S. Fernando, "Combining Long-Term Recurrent Convolutional and Graph Convolutional Networks to Detect Phishing Sites Using URL and HTML," in IEEE Access, vol. 10, pp. 82355-82375, 2022, doi: 10.1109/ACCESS.2022.3196018. genes

[2] J. Feng, L. Zou, O. Ye and J. Han, "Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 8, pp. 221214-221224, 2020, doi: 10.1109/ACCESS.2020.3043188.

[3] L. R. Kalabarige, R. S. Rao, A. R. Pais and L. A. Gabralla, "A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites," in IEEE Access, vol. 11, pp. 71180-71193, 2023, doi: 10.1109/ACCESS.2023.3293649.

[4] Y. Fang, C. Zhang, C. Huang, L. Liu and Y. Yang, "Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism," in IEEE Access, vol. 7, pp. 56329-56340, 2019, doi: 10.1109/ACCESS.2019.2913705.

[5] W. Ali and S. Malebary, "Particle Swarm Optimization-Based Feature Weighting for Improving Intelligent Phishing Website Detection," in IEEE Access, vol. 8, pp. 116766-116780, 2020, doi: 10.1109/ACCESS.2020.3003569.

[6] E. S. Gualberto, R. T. De Sousa, T. P. De Brito Vieira, J. P. C. L. Da Costa and C. G. Duque, "The Answer is in the Text: Multi-Stage Methods for Phishing Detection Based on Feature Engineering," in IEEE Access, vol. 8, pp. 223529-223547, 2020, doi: 10.1109/ACCESS.2020.3043396.

[7] M. Sánchez-Paniagua, E. F. Fernández, E. Alegre, W. Al-Nabki and V. González-Castro, "Phishing URL Detection: A Real-Case Scenario Through Login URLs," in IEEE Access, vol. 10, pp. 42949-42960, 2022, doi: 10.1109/ACCESS.2022.3168681.

[8] S. Al-Ahmadi, A. Alotaibi and O. Alsaleh, "PDGAN: Phishing Detection With Generative Adversarial Networks," in IEEE Access, vol. 10, pp. 42459-42468, 2022, doi: 10.1109/ACCESS.2022.3168235.

[9] Liu and J. Fu, "SPWalk: Similar Property Oriented Feature Learning for Phishing Detection," in IEEE Access, vol. 8, pp. 87031-87045, 2020, doi: 10.1109/ACCESS.2020.2992381.

[10] item M. Sameen, K. Han and S. O. Hwang, "PhishHaven—An Efficient Real-Time AI Phishing URLs Detection System," in IEEE Access, vol. 8, pp. 83425-83443, 2020, doi: 10.1109/ACCESS.2020.2991403.