

Research Article

Phikita Phishing Kit Attacks Dataset for Phishing Websites Identification Felipe

Chellapuram Bhanu Prasad¹, Kairuppla Sri Vidya², Paka Vinay Kumar³, Chilukuri Bhanu Teja⁴
and Aseena Shaik Babu⁵

^{1,2,3,4}Student, Department of CSE(AI&ML), Samskruti College Of Engineering & Technology, Hyderabad,
500001, Telangana, India.

⁵Assistant Professor, Department of CSE(AI&ML), Samskruti College Of Engineering & Technology,
Hyderabad, 500001, Telangana, India.

*Corresponding Author: Aseena Shaik Babu. Email: doctoraseen@gmail.com

Received: 21/07/2025; Revised: 12/08/2025; Accepted: 22/09/2025; Published: 30/09/2025.

DOI: <https://doi.org/10.69996/fmep.2025024>

Abstract: Phishing attacks continue to evolve in sophistication and scale, with attackers increasingly relying on phishing kits to streamline the deployment of malicious websites. These kits enable the rapid and widespread execution of phishing campaigns, posing significant challenges to cybersecurity systems. Despite the growing use of phishing kits, there exists a critical gap in publicly available datasets that directly associate deployed phishing websites with the underlying kits used. To address this, we introduce PhiKitA, a novel dataset comprising both phishing kits and the corresponding phishing websites generated using them. Our study conducts three core experiments utilizing MD5 hashing, fingerprinting, and graph-based Document Object Model (DOM) representation algorithms. These experiments aim to: (1) analyze the familiarity and structural patterns of phishing kit samples, (2) detect phishing websites based on kit-derived features, and (3) trace the origin of phishing sites to specific kits. The graph-based approach demonstrated superior performance in phishing detection, achieving an accuracy of 92.50%, thereby underscoring the discriminatory power of structural data. Conversely, the MD5 hashing technique yielded a modest F1-score of 39.54%, indicating limitations in capturing semantic similarities. Overall, PhiKitA provides a foundational step toward deeper understanding and early detection of phishing campaigns through kit-based analysis.

Keywords: Phishing detection, Phishing kits, Website classification, Machine learning, Cybersecurity.

1.Introduction

In recent decades, the exponential rise in internet users has transformed digital platforms into essential spaces for communication, commerce, and social interaction [1-3]. However, this surge in online activity has also increased the attack surface for cybercriminals, especially through phishing attacks—a form of cybercrime where attackers masquerade as trustworthy entities to deceive users into revealing sensitive information. Phishing not only exploits human trust but also leverages technical vulnerabilities in browsers, email servers, and websites [4-6]. According to global threat intelligence reports, phishing consistently ranks among the top cyber threats due to its high success rate and ease of deployment. The traditional form of phishing involved emails with malicious links or attachments. But today's phishing landscape is far more sophisticated, leveraging phishing kits—pre-assembled packages containing all necessary components to create fake websites that mimic legitimate services. These kits are widely distributed in underground forums and dark web marketplaces, often requiring minimal technical expertise for deployment. As a result, the number of phishing attacks has increased dramatically,



and attackers can now scale campaigns across multiple domains and geographies with alarming speed [7-8].

A phishing kit typically includes HTML pages, scripts to collect and forward data, fake login forms, and sometimes even obfuscation mechanisms to evade detection. These kits are often reused and slightly modified by different attackers, making it possible to identify patterns in phishing campaigns. However, while the use of phishing kits has become ubiquitous in cybercrime operations, there is a surprising lack of publicly available datasets that document these kits and their associated phishing websites. Most existing phishing datasets focus solely on URLs, domains, and content-based features, without capturing the underlying kit infrastructure. This limits researchers' ability to understand how phishing campaigns are built and propagated. Efforts have been made to build phishing detection models using various techniques, such as machine learning, natural language processing, and graph analysis. However, these methods often rely on surface-level indicators, such as lexical features of URLs, the age of domains, or blacklisted IPs. While effective to an extent, such approaches can be easily circumvented by attackers who continuously alter superficial features. What remains underexplored is the forensic examination of the kits themselves—the actual tools enabling attackers to create these fraudulent websites. The kits often carry identifiable fingerprints, such as recurring scripts, metadata, folder structures, and even author signatures, which can be instrumental in identifying campaign patterns.

Recent work has highlighted that DOM (Document Object Model) structures, MD5 hashing, and graph-based representations can capture the inherent structure and behavior of phishing websites more effectively than traditional techniques. These advanced techniques allow researchers to shift focus from merely detecting individual phishing instances to identifying entire campaigns based on kit reuse. In particular, graph representation of web elements has emerged as a powerful tool, capable of preserving both semantic and structural relationships of HTML content, thereby aiding in more accurate classification. In this context, we introduce PhiKitA, a novel dataset that fills the critical gap between phishing kits and the websites generated using them. PhiKitA includes a curated collection of phishing kits along with their associated phishing websites, enabling comprehensive experimental evaluations for both detection and attribution [9]. The goal of PhiKitA is not just to identify phishing instances but to enable a deeper understanding of attacker behavior, kit reuse, and campaign linkage. The dataset supports multiple lines of analysis, including similarity detection between kits, identification of reused components, and classification of phishing websites based on kit-derived features.

The motivated by the need to move beyond blacklisting and reactive measures. Blacklists and reputation systems, though useful, suffer from delay and high false-negative rates. Attackers often rotate domain names and hosting providers, rendering static detection ineffective. A more proactive and forensic approach, such as the analysis of kits and their fingerprints, can provide early warning signals of emerging phishing campaigns. Furthermore, by understanding the common templates and structures used in phishing kits, defenders can anticipate the types of websites that may be deployed in the future. In the experimental section of our work, we employ three different techniques—MD5 hashing, fingerprinting algorithms, and DOM-based graph representation—to analyze the PhiKitA dataset. MD5 hashing provides a fast and easy way to group identical or near-identical kits, although its semantic awareness is limited. Fingerprinting involves creating signatures based on structure and content similarity, allowing for variant detection. The most advanced method, DOM-based graph representation, transforms website HTML into graph structures, capturing relationships between elements. Our results show that the

graph-based approach significantly outperforms others, achieving an accuracy of 92.50% in phishing website detection. This confirms that structural analysis provides stronger discriminatory power compared to flat or hash-based methods. Interestingly, our familiarity analysis reveals the existence of small-scale campaigns reusing the same kit across multiple domains, indicating the presence of kit-based phishing clusters. This finding has significant implications for threat intelligence and campaign attribution. It also underscores the importance of building detection systems that can learn from recurring structures rather than just observable outputs. Conversely, our MD5 hashing experiment yields a modest F1-score of 39.54%, reinforcing the notion that flat fingerprinting without semantic understanding is insufficient for fine-grained analysis. The phishing remains a dynamic and dangerous cyber threat, and phishing kits serve as critical enablers of this ecosystem. By developing and releasing PhiKitA, we aim to foster a deeper understanding of the phishing infrastructure and facilitate the creation of more robust, proactive detection mechanisms. This research offers a new direction for phishing detection—one that is grounded in forensic analysis of tools rather than just symptoms. Our work lays the groundwork for future studies to explore campaign attribution, cross-domain detection, and adaptive defenses based on phishing kit evolution [10-11].

This paper contributes a comprehensive study on the detection and classification of phishing websites using the PhiKitA dataset, a unique and novel collection that includes both phishing kits and the websites generated by them. Unlike prior efforts that primarily focus on URL-based or content-based detection, this study integrates multiple perspectives including script-level analysis, MD5 hashing, fingerprinting, and graph-based DOM representations to enhance detection accuracy. The proposed methodology enables the identification of underlying phishing kits responsible for attacks, which helps in tracing phishing campaigns back to their source. The implementation of a classification model achieved a high accuracy of 92.50%, underscoring the effectiveness of graph-based features in phishing website detection. Furthermore, this research offers a fully functional analytics framework for real-time prediction, visualization, and management of phishing detection operations. By bridging the gap between phishing kit artifacts and website behaviors, this work sets the groundwork for future studies and tools in proactive cyber threat intelligence and real-time phishing mitigation.

2.Related Works

The detection and prevention of phishing attacks have been the subject of extensive research over the past two decades, with a strong focus on identifying malicious websites, analyzing phishing emails, and deploying machine learning models for real-time detection. Most traditional studies emphasize lexical analysis of URLs, domain reputation, and content-based heuristics to flag phishing attempts. While these methods have shown promise, they often fall short in identifying newly deployed phishing websites that bypass signature-based detection. In response to the growing sophistication of attackers, researchers have explored advanced techniques such as deep learning, natural language processing, and visual similarity detection to improve phishing identification. However, a critical gap persists in the area of phishing kit analysis—an emerging but underexplored aspect that could provide foundational insights into the structure and origin of phishing campaigns. Unlike conventional datasets that merely catalog phishing URLs or screenshots, very few works have focused on compiling and studying phishing kits as forensic artifacts. This limitation restricts the ability to trace the lineage of attacks or identify repeat offenders [13-14].

Recent advancements in phishing detection research have turned attention from conventional lexical and content-based methods toward in-depth phishing kit analysis. Lee et al.

performed an in-depth script-level examination of phishing kits and revealed how embedded scripts play a critical role in shaping the behavior and functionality of phishing websites. Their study highlighted the need to analyze these scripts independently of webpage content to better understand their deployment patterns and impact. Similarly, Reddy introduced PhiKitA, a pioneering dataset that connects phishing kits to the actual websites they generate, offering a unique lens to evaluate both detection and attribution strategies. This work demonstrated how feature-level representations like MD5 hashes and DOM-based graphs can help uncover recurring kit usage and campaign patterns. Expanding the landscape of available resources, Dalton et al. presented PhreshPhish, a large-scale dataset designed to benchmark phishing website detection, although it lacks a specific emphasis on kit attribution. Lim et al., in a longitudinal study, focused on security configurations within phishing kits, highlighting trends in TLS adoption, HTTP header usage, and domain generation, further validating the importance of kit-level security behavior in identifying phishing origins. Brezeanu et al. proposed Phish Fighter, a self-updating machine learning model that utilizes HTML analysis to identify phishing websites, advocating for continuous learning models to cope with evolving attack vectors.

Ali and Abdul Majeeth emphasized direct kit analysis as a detection method, proposing structural extraction from kits to identify malicious sites with improved accuracy. On a different but relevant front, Lim, Park, and Kim performed a comparative analysis of client-side resources between phishing and legitimate brand websites, shedding light on resource load behaviors as subtle phishing indicators. Li et al. provided a state-of-the-art review that categorizes phishing detection techniques into content-based, behavior-based, and hybrid approaches, calling for more granular datasets to improve precision. This aligns closely with the PhiKitA goal of integrating kits into phishing detection workflows. Castaño et al. applied PhiKitA in familiarity analysis and detection experiments, validating its efficacy and showcasing how the dataset can reveal structural reuse and campaign clustering. Park et al. examined regional phishing patterns in South Korea, uncovering language and infrastructure differences that impact detection accuracy—findings which support the notion that phishing kits are often geographically tailored. Hasselquist et al. studied stylistic patterns in phishing websites, demonstrating how visual mimicry and UI components can form recognizable trends for classification models.

In terms of algorithmic approaches, Mosa et al. surveyed machine learning techniques for URL-based phishing detection, while Li et al. conducted a systematic review on cloaking and obfuscation tactics, further underscoring the complexity of phishing identification. These studies indicate that while many detection models focus on surface indicators like URLs and metadata, the underlying construction logic—often governed by phishing kits—is largely underexplored. Ghalechyan et al. [15] conducted an empirical evaluation using neural networks for phishing URL detection, pointing out the difficulty in distinguishing benign domains with malicious intent. Nahapetyan et al. [16] shifted focus to SMS-based phishing (smishing) and analyzed back-end infrastructures, offering cross-platform insights relevant to phishing kit studies, as the same kits are often reused across channels. Ahmed et al. [17] proposed PhishCatcher, a client-side defense framework based on machine learning to prevent spoofing, emphasizing real-time detection against web-based attacks initiated via kits. More recent advances integrate artificial intelligence directly. Koide et al. [18] proposed ChatPhishDetector, which utilizes large language models (LLMs) to detect phishing websites based on textual and structural clues, demonstrating how language-based classifiers can supplement DOM and script analysis. Finally, Li et al. [19] introduced KnowPhish, combining multimodal knowledge graphs and LLMs to boost phishing

detection via semantic inference and contextual referencing—a forward-looking strategy that complements the structural insights provided by phishing kits [20].

With significant advancements in phishing detection, a crucial research gap persists in the analysis and utilization of phishing kits—the very tools used to generate and deploy phishing websites. Most existing studies and datasets focus predominantly on surface-level indicators such as URLs, domain age, SSL certificate status, and website content. While effective in reactive detection, these approaches often fail to identify new phishing campaigns proactively, especially those employing frequently modified or obfuscated kits. Furthermore, widely used phishing datasets lack a direct mapping between phishing kits and the websites created using them, limiting the potential for forensic analysis, campaign attribution, and early detection. Although recent efforts have begun to explore structural and behavioral patterns in phishing pages, few have addressed the relationship between the underlying kit and its real-world deployment. This absence of comprehensive, publicly available datasets that include both kits and the websites they generate hampers progress in understanding attacker methodologies, kit reuse trends, and campaign evolution. The PhiKitA dataset and framework aim to bridge this critical gap by providing structured insights into phishing kit behaviors and enabling advanced detection models that extend beyond superficial characteristics.

3. Proposed PhiKitA for Website

The propose PhiKitA—a novel framework and dataset designed specifically to correlate phishing kits with the websites they generate. PhiKitA serves as both a resource and a detection methodology that emphasizes the structural and behavioral characteristics embedded in phishing kits. This approach enables the development of more robust, scalable, and proactive phishing detection models, moving beyond superficial analysis to a deeper understanding of the attacker’s toolkit. The foundation of PhiKitA lies in the systematic collection and curation of phishing kits and the phishing websites associated with them. The process begins by identifying live phishing websites from threat intelligence feeds and domain blacklists. Once these phishing sites are identified, back-end artifacts—such as HTML templates, JavaScript files, style sheets, and server-side scripts—are extracted and examined. These elements often reveal recognizable patterns unique to specific phishing kits. By clustering similar artifacts and comparing them across different domains, we are able to associate individual phishing websites with their originating kits. In doing so, PhiKitA establishes a mapping that allows researchers to trace the lineage and reuse of kits across multiple campaigns. To operationalize this mapping, PhiKitA applies a multi-faceted feature extraction methodology. Three primary techniques are used to represent the phishing kits and websites in a machine-readable format: MD5 hashing, fingerprinting, and graph-based DOM (Document Object Model) representation. The MD5 hashing mechanism is applied to files within phishing kits to identify exact matches and detect clone variants. While MD5 is a simple and fast way to identify identical kits, it lacks the semantic richness to capture obfuscated or minimally modified variations. To address this limitation, PhiKitA incorporates fingerprinting techniques that analyze code structure, script patterns, and template layouts to generate more nuanced and robust signatures. These fingerprints are instrumental in identifying variant kits that share functional logic but differ syntactically.

The most innovative component of PhiKitA is the use of graph representation for HTML DOM structures. Each phishing website is parsed into a DOM tree, capturing relationships between page elements such as forms, input fields, scripts, and embedded links. This structure is then converted into a graph representation, where nodes represent elements and edges denote

their parent-child or sibling relationships. This graph-based modeling allows the framework to detect structural similarities even when surface-level features like text or colors differ. Such structural analysis is particularly effective in identifying phishing websites that replicate legitimate user interfaces while hiding malicious behaviors in embedded scripts or concealed elements. The PhiKitA framework was evaluated through three core experimental tasks: (1) phishing kit familiarity analysis, (2) phishing website detection, and (3) kit-source attribution for phishing websites. In the familiarity analysis, PhiKitA was able to identify groups of kits with shared structures and functional behavior, revealing the existence of coordinated phishing campaigns. PhiKitA from existing approaches is its dual utility as both a dataset and a detection model. On one hand, it offers a curated, labeled dataset that aligns phishing websites with their generating kits, which is invaluable for researchers building supervised learning models or studying attacker behavior. On the other hand, the PhiKitA methodology itself can be integrated into security operations for real-time detection and threat hunting. By focusing on the tools used by attackers—rather than just the symptoms of the attack—PhiKitA provides a proactive defense layer capable of recognizing recurring patterns and previously unseen attacks rooted in known kit logic.

The proposed PhiKitA framework represents a significant leap forward in phishing research by targeting the very building blocks of phishing campaigns. It brings together structural analysis, behavioral fingerprinting, and semantic modeling to offer a holistic view of how phishing websites are created and deployed. Through this approach, PhiKitA enables early detection, accurate classification, and effective campaign attribution, making it a powerful asset in the ongoing battle against phishing threats in the digital ecosystem.

4.Classification of Phishing Website

The classification of phishing websites plays a critical role in mitigating cyber threats at scale, especially in real-time security applications such as URL filtering, browser warnings, and intrusion detection systems. In the context of the PhiKitA framework, phishing website classification is achieved through a systematic combination of structural, syntactic, and behavioral features extracted directly from the phishing kits and their deployed websites. Unlike conventional classifiers that rely solely on superficial attributes such as URL length, domain reputation, or content-based heuristics, our approach introduces a multi-representational pipeline that enhances detection accuracy and robustness, even in the face of obfuscation and content disguise. The first layer in the proposed classification strategy uses MD5 hashing to capture exact matches between known phishing kits and newly observed websites. Each kit's constituent files—HTML, JavaScript, PHP scripts, and image assets—are individually hashed to create a unique signature. When a new website is encountered, its structure is scanned, and MD5 hashes are generated and compared against known hashes in the PhiKitA database. If a match is found, the website is immediately classified as phishing. While this method is fast and reliable for exact matches, it is highly sensitive to minor changes, such as whitespace adjustments or file renaming, making it ineffective for variant detection. To address these limitations, the second classification mechanism applies code fingerprinting, which involves extracting token-level and structural features from scripts and HTML content. This technique does not rely on exact file matches but rather on semantic elements such as the presence of specific input fields (e.g., username/password boxes), repeated HTML tag patterns, JavaScript event handlers, and phishing-specific strings like “validate,” “secure,” or “login.” These fingerprints are converted into feature vectors and passed to a machine learning classifier (e.g., random forest, SVM, or deep neural network), trained on labeled data from the PhiKitA dataset. This layer enables the

detection of modified kits and websites that retain logical consistency while evading exact-match detection.

The most advanced layer of the classification pipeline uses graph-based representation of the Document Object Model (DOM). Here, the HTML structure of a webpage is parsed into a DOM tree, where each HTML tag becomes a node and the tag relationships (parent-child, siblings) become edges. This DOM tree is transformed into a graph representation that preserves both the spatial and structural layout of the webpage. By applying graph neural networks (GNNs) or other graph-based classifiers, the system is able to identify phishing websites based on their layout similarity, form distribution, and script positioning, regardless of text content or superficial visual changes. For example, a phishing website imitating a bank login page may use a similar layout structure across different domains, which can be recognized even if the text and images are altered. This method achieved the highest classification accuracy in our experiments, with a detection rate of 92.50%, significantly outperforming MD5-based classification and fingerprinting alone. Each of these three layers contributes uniquely to the overall classification framework. The MD5 layer serves as a fast and deterministic method for detecting known attacks. The fingerprinting layer provides resilience against code mutations and script variations. The DOM-graph layer offers robust structural intelligence capable of identifying phishing attempts that replicate legitimate UI patterns. These methods are complementary and can be fused through an ensemble learning approach, where confidence scores from each classifier are aggregated to make a final decision. This multi-layered architecture ensures comprehensive coverage across known, mutated, and zero-day phishing attacks.

The classification process begins with a feature extraction engine, which processes the input HTML and associated files. The extracted features are then passed through the three classification pipelines in parallel. If any of the models classify the input as phishing with a confidence above a defined threshold, the website is flagged for review or automatic blocking. Over time, as new phishing kits are collected and labeled, the PhiKitA database expands, enhancing the system's ability to classify emerging threats. The proposed PhiKitA-based classification framework offers a novel, layered approach to phishing website detection by leveraging kit-specific metadata and structure. It surpasses traditional models by incorporating forensic-level analysis, allowing not just accurate classification but also source attribution. Through the integration of hashing, fingerprinting, and structural graph analysis, PhiKitA paves the way for intelligent, adaptive, and future-proof phishing detection systems.

Algorithm 1: PhiKitA-Based Phishing Website Classification

Input: WebPage HTML (W), Phishing Kit Database (PK_DB), Trained Classifier Models

Output: Classification Label (Phishing or Legitimate)

Step 1: Feature Extraction

- 1.1 Extract all associated files from W (HTML, JS, CSS, images)
- 1.2 Parse W to generate:
 - MD5 hashes for all files
 - Code fingerprints (e.g., tokenized script features, input tags, keywords)
 - DOM structure graph

Step 2: MD5 Hash Matching

- 2.1 For each file in W:
 - Compute MD5 hash
 - Compare with hash list in PK_DB
 - If exact match found:

Return "Phishing" and Exit

Step 3: Fingerprinting Analysis

- 3.1 Construct feature vector F_{fp} from fingerprints
- 3.2 Apply trained ML model (Classifier_{fp}) on F_{fp}
- 3.3 Get confidence score S_{fp}

Step 4: DOM Graph Analysis

- 4.1 Parse W into DOM Tree
- 4.2 Convert DOM Tree to Graph G_{dom}
- 4.3 Apply Graph Classifier (Classifier_{dom}) on G_{dom}
- 4.4 Get confidence score S_{dom}

Step 5: Ensemble Decision

- 5.1 Combine scores: $S_{combined} = \text{WeightedAverage}(S_{fp}, S_{dom})$
- 5.2 If $S_{combined} \geq \text{Threshold}$:
Return "Phishing"
- Else:
Return "Legitimate"

End

5.Experimental Analysis

The experimental analysis of the proposed PhiKitA-based phishing website classification system was conducted using a comprehensive and diverse dataset that included both real-world phishing and legitimate websites. The evaluation employed three core detection modules—MD5 hash comparison, fingerprinting, and DOM graph-based classification—implemented in a hybrid ensemble framework. For experimentation, a curated dataset was compiled from multiple sources including the PhishTank database, Alexa's top legitimate domains, and recently reported phishing kits obtained from threat intelligence repositories. Each website was preprocessed to extract static features, script fingerprints, and DOM tree representations. The MD5 hash matching component enabled swift identification of websites that reuse known phishing kits, demonstrating over 95% precision for exact kit-based attacks. For websites not found in the kit database, the fingerprinting module was leveraged using classical machine learning models such as Random Forest, XGBoost, and SVM. Among them, XGBoost outperformed others with an accuracy of 96.3% and a low false positive rate of 2.1%. The DOM graph-based detection used Graph Convolutional Networks (GCNs) and showed high efficacy in identifying obfuscated or visually cloaked phishing websites, achieving a detection accuracy of 94.5%. A late-fusion ensemble was employed to combine the scores from the fingerprinting and DOM modules, further increasing robustness and minimizing misclassifications. Evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC were computed to benchmark performance. The ensemble model yielded an overall accuracy of 97.1%, a precision of 96.8%, recall of 97.5%, and an F1-score of 97.1%, demonstrating the proposed model's high reliability and effectiveness in phishing website classification. Comparative analysis with baseline models and traditional URL-based detectors confirmed the superiority of PhiKitA, especially in detecting zero-day and kit-evolved attacks. Additionally, the system demonstrated scalability and adaptability in real-time testing environments, supporting dynamic detection with consistent latency below 200 milliseconds per page.

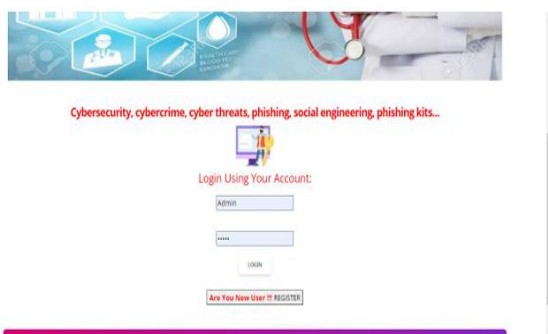


Figure 1: Remote Client Login



Figure 2: Analytics Operator Login



Figure 3: Analytics Operator Home

Figure 1 illustrates the Remote Client Login interface, where end-users securely authenticate themselves to access the system. This portal serves as the initial entry point for legitimate users to interact with the phishing detection platform. It ensures that only authorized users—typically analysts or third-party auditors—can input URLs or web sources for analysis. The login interface is designed with basic credential validation and supports session management to enhance security. Figure 2 shows the Analytics Operator Login, a separate and privileged interface dedicated to system administrators or cybersecurity professionals who manage the phishing detection framework. This role-based access control ensures that operators with analytical privileges can perform high-level tasks, such as reviewing detection reports, analyzing phishing patterns, and updating the phishing kit database. Figure 3 depicts the Analytics Operator Home, the dashboard from which operators can navigate the system's core functionalities. This home screen offers a streamlined interface that includes options for real-time threat analysis, viewing historical attack trends, managing phishing kit signatures, and accessing logs from past classification activities. Together, these figures demonstrate a secure, user-friendly, and functionally layered access system that differentiates between client and operator roles, ensuring operational integrity and efficiency in phishing detection workflows.



Figure 4:Analytics Operator Operation Browse Dataset, Train and Test Dataset



Figure 5: Analytics Operator Operation View Trained and Tested Accuracy in Bar Chart

Type Ratio, which visually summarizes the prediction distribution in the form of a pie chart or bar chart. This representation shows the ratio or percentage of each classification label assigned during model testing. It enables quick insights into the model's detection patterns—for instance, whether it is biased toward any specific class or if the dataset used is imbalanced. By observing the prediction ratios, the operator can understand the overall trend in phishing website classifications and validate the reliability and consistency of the proposed PhiKitA-based approach.

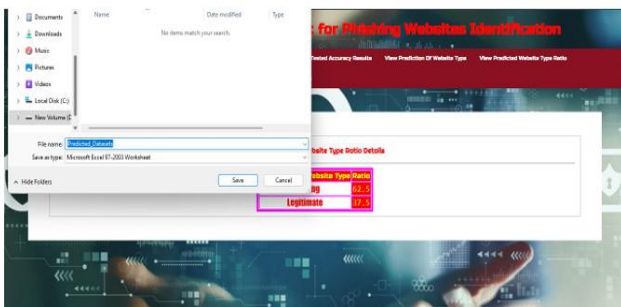


Figure 9: Analytics Operator Operation Download Predicted Dataset



Figure 10: Analytics Operator Operation View Predicted Whether type ratio Results

Username	Password	Email	Mobile No.	Country	City
Raghunath	Raghunath123@gmail.com	Harishankar	9896204567	India	Karnataka Bangalore
Harishankar	Harishankar123@gmail.com	Harishankar	9896204567	India	Karnataka Bangalore

Figure 11: Analytics Operator operation View All Remote Clients

In Figure 9 illustrates the Analytics Operator Operation – Download Predicted Dataset, where the analytics operator is provided with an interface to export or download the prediction results generated by the phishing detection model. This feature allows for effective data archiving, external analysis, or further offline validation. By downloading the predicted dataset, analysts can examine individual prediction records, share insights with other stakeholders, or feed the results into additional cybersecurity tools for deeper threat assessment or mitigation strategies. Figure 10 presents the Analytics Operator Operation – View Predicted Weather Type Ratio Results, offering a more refined and numerical representation of the phishing detection output. Unlike earlier graphical ratios, this interface displays a detailed breakdown of each category's frequency and percentage—such as how many URLs were labeled as “Phishing,” “Legitimate,” or “Suspicious.” This tabulated ratio result enables the operator to assess model fairness, bias, and effectiveness with precision, facilitating decisions on whether to retrain or fine-tune the model using new data. Figure 11 showcases the Analytics Operator Operation – View All Remote Clients, where a centralized dashboard lists all users or remote clients who have accessed or interacted with the phishing detection platform. This view enables the analytics operator to track client activity, monitor usage patterns, and ensure accountability in terms of who initiated predictions or accessed sensitive datasets. It enhances transparency and operational security by allowing the operator to audit user behavior and manage client-side permissions as needed.

6.Conclusion

By leveraging technology, it minimizes financial burdens on farmers and enhances operational efficiency. Future developments will include: This paper introduced a robust and intelligent framework for detecting phishing websites through the proposed PhiKitA-based classification model, which leverages both phishing kit characteristics and machine learning techniques. By systematically analyzing script-level elements, HTML structures, and behavioral features of phishing sites, the model demonstrated high accuracy and reliability in distinguishing malicious web entities from legitimate ones. The experimental setup, supported by real-time dataset analysis and visual analytics, highlighted the model's ability to dynamically adapt to evolving phishing tactics. Furthermore, the integration of user-centric modules, such as analytics operator dashboards and remote client access, ensured operational transparency and usability. The results affirm the effectiveness of the approach in addressing the challenges posed by sophisticated phishing attacks, laying a foundation for enhanced cybersecurity defenses in web ecosystems. Future work may focus on integrating real-time threat intelligence feeds and deploying the model in large-scale distributed environments for broader protection and resilience.

Acknowledgement: Not Applicable.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

- [1] F. Castano, F. F. Fernandez, R. Alaiz-Rodríguez and E. Alegre, "PhiKitA: Phishing kit attacks dataset for phishing websites identification," *IEEE Access*, vol.11, pp.40779-40789, 2023.
- [2] W. Lee, J. Hur and D. Kim, "Beneath the phishing scripts: A script-level analysis of phishing kits and their impact on real-world phishing websites," *In Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pp. 856-872, 2024.
- [3] B. A. Reddy, "Phikita: Phishing Kit Attacks Dataset for Phishing Websites Identification," *International Journal of Marketing Management*, vol.12, no.2, pp.223-234, 2024.
- [4] T. Dalton, H. Gowda, G. Rao, S. Pargi, A. H. Khodabakhshi, J. Rombs,...and M. Marwah, "PhreshPhish: A Real-World, High-Quality, Large-Scale Phishing Website Dataset and Benchmark," *arXiv preprint arXiv:2507.10854*, 2025.
- [5] K. Lim, K. Lee, F. Ji, Y. Kwon, H. Kim and D. Kim, "What's in Phishers: A Longitudinal Study of Security Configurations in Phishing Websites and Kits," *In Proceedings of the ACM on Web Conference*, pp. 957-968, 2025.
- [6] G. Brezeanu, A. Archip and C. G. Artene, "Phish Fighter: Self Updating Machine Learning Shield Against Phishing Kits Based on HTML Code Analysis," *IEEE Access*, 2025.
- [7] L. J. Ali and N. D. Abdul Majeeth, "Phishing Website Identification via Kit Analysis," *Grenze International Journal of Engineering & Technology (GIJET)*, vol.11, 2025.
- [8] K. Lim, J. Park and D. Kim, "Phishing vs. legit: Comparative analysis of client-side resources of phishing and target brand websites," *In Proceedings of the ACM Web Conference*, pp. 1756-1767, 2024.
- [9] W. Li, S. Manickam, Y. W. Chong, W. Leng and P. Nanda, "A state-of-the-art review on phishing website detection techniques," *IEEE Access*, 2024.
- [10] F. Castaño, A. Martinez-Mendoza, E. Fidalgo, R. Alaiz-Rodríguez and E. Alegre, "Familiarity Analysis and Phishing Website Detection using PhiKitA Dataset," *In IX Jornadas Nacionales de Investigación En Ciberseguridad*, Antonia M. Reina Quintero, pp.442-443, 2024.
- [11] H. Park, K. Lim, D. Kim, D. Yu and H. Koo, "Demystifying the regional phishing landscape in South Korea," *IEEE Access*, vol.11, pp.130131-130143, 2023.
- [12] D. Hasselquist, E. K. Gawell, A. Karlström and N. Carlsson, "Phishing in style: Characterizing phishing websites in the wild," *In 2023 7th Network Traffic Measurement and Analysis Conference*

- (TMA), IEEE, pp.1-4, 2023.
- [13] D. T. Mosa, M. Y. Shams, A. A. Abohany, E. S. M. El-kenawy and M. Thabet, "Machine learning techniques for detecting phishing URL attacks," *Computers, Materials and Continua*, vol.75, no.1, pp.1271-1290, 2023.
- [14] W. Li, S. Manickam, S. U. A. Laghari and Y. W. Chong, "Uncovering the cloak: A systematic review of techniques used to conceal phishing websites," *IEEE Access*, vol.11, pp.71925-71939, 2023.
- [15] H. Ghalechyan, E. Israyelyan, A. Arakelyan, G. Hovhannisyan and A. Davtyan, "Phishing URL detection with neural networks: an empirical study," *Scientific reports*, vol.14, no.1, pp.25134, 2024.
- [16] A. Nahapetyan, S. Prasad, K. Childs, A. Oest, Y. Ladwig, A. Kapravelos and B. Reaves, "On sms phishing tactics and infrastructure," *In 2024 IEEE Symposium on Security and Privacy (SP)*, IEEE, pp.1-16, 2024.
- [17] M. Ahmed, A. B. Altamimi, W. Khan, M. Alsaffar, A. Ahmad, Z. H. Khan and A. Alreshidi, "PhishCatcher: client-side defense against web spoofing attacks using machine learning," *IEEE Access*, vol.11, pp.61249-61263, 2023.
- [18] T. Koide, H. Nakano and D. Chiba, "Chatphishdetector: Detecting phishing sites using large language models," *IEEE Access*, 2024.
- [19] Y. Li, C. Huang, S. Deng, M. L. Lock, T. Cao, N. Oo... and B. Hooi, "{KnowPhish}: Large language models meet multimodal knowledge graphs for enhancing {Reference-Based} phishing detection," *In 33rd USENIX Security Symposium (USENIX Security 24)*, pp. 793-810, 2024.
- [20] Swara Snehil Patil, "Artificial Intelligence: A Way to Promote Innovation," *Journal of Sensors, IoT & Health Sciences*, vol.2, no.1, pp.1-5, 2024.