International Journal of Applied Information Systems and Cybersecurity
Vol. 1 No. 1 (2026) pp. 36-59
Online ISSN: XXXX-XXXX

Original Research

An Interpretable Exploratory Framework for Analyzing Structural Characteristics of Phishing Websites

Arifah Adlina[1]✉, Nur Aina Amanina[1]
[1] International Islamic University Malaysia, Malaysia
✉ arifah.adlina@gmail.com
10.65840/ijaisc.vxix.xx

Article Info
Submitted:
09/21/2025
Revised:
11/10/2025
Accepted:
12/25/2025
Available Online:
01/15/2026

Abstract. *Phishing attacks continue to be a persistent danger to cybersecurity by exploiting technical flaws and human trust. Previous studies have primarily focused on machine learning performance, whereas relatively few have examined the structural characteristics that differentiate phishing websites from legitimate websites. In this paper, we provide an interpretable exploratory approach to analyze phishing behavior utilizing structural, symbolic and domain related website features from a publicly available phishing dataset. The analytical approach incorporates distribution analysis, non-parametric statistical testing and correlation-based investigation to explore structural anomalies associated with phishing. The study analyzes a large-scale phishing website dataset containing both phishing and legitimate instances represented by URL-based and domain-related attributes. Representative features were selected through preprocessing and redundancy filtering to retain structurally informative indicators while improving analytical interpretability. The proposed framework emphasizes statistical understanding of phishing behavior rather than predictive optimization, enabling a transparent examination of feature distributions, class-specific differences, and feature interactions. The results indicate that phishing URLs display more structural variation, and more apparent right-skewed distributions, especially for URL length and symbolic composition patterns. Statistical tests validated significant structural differences between phishing and legitimate websites ($p < 0.001$). Correlation analysis showed a moderate positive link between URL length and symbolic manipulation behavior. Overall, the results show that the interpretability and the analytical robustness of phishing analysis are improved when multiple structural indicators are evaluated jointly rather than in isolation.*

Keywords: *phishing website detection; lexical URL analysis; exploratory data analysis; explainable cybersecurity; phishing behavior analysis*

1. Introduction

Phishing continues to be one of the most common cybersecurity threats, as it manipulates technological vulnerabilities and human trust. Phishing tactics, unlike many technically advanced assaults, are generally cheap to conduct and may be implemented quickly, while nevertheless leading to credential theft, account compromise, and broader social engineering incidents affecting both individuals and organizations (ENISA, 2022, 2023). The increasing use of digital communication

channels, cloud services and online transactions has made phishing operations more scalable and accessible.

As per the latest numbers published by Statista, malicious email traffic still represents a significant fraction of the total email traffic worldwide, making phishing a scalable cybercrime method, rather than an isolated attack event (Petrosyan, 2024). Simultaneously, phishing infrastructures have become more flexible via automated deployment, dynamic URL generation, and realistic replication of trusted online services (Basit et al., 2021; Popescul and Radu, 2025). Recent studies have also shown that Large Language Models (LLMs) can be used to produce highly convincing phishing content at scale, and are also being explored as defensive tools for phishing detection and analysis (Heiding et al., 2024).

These improvements have increasingly moved phishing detection research from blacklist-based defenses to feature-driven analytical techniques. Early defenses against phishing depended mainly on pre-defined malicious URL repositories and manually vetted signatures. Blacklist-based systems are successful against known threats, but they typically fail to detect new phishing websites that are created and taken down quickly and that change their structure swiftly. Similar constraints were pointed out by Jain and Gupta (2018), who underlined that blacklist reliant methods are not suitable for real-time phishing detection and so there is a need for lightweight structural analysis on the client side. To tackle these shortcomings, Marchal et al. (2014) proposed the PhishStorm framework and showed that real-time lexical analysis of URLs can offer a higher degree of robustness against fast-changing phishing infrastructures compared to traditional blacklist approaches.

Figure 1 shows the main operational workflow of phishing attempts. Attackers usually create fake websites that seem like real digital services, sending modified URLs by email, SMS and malicious ads. Attackers can collect credentials, financial information or other sensitive data when users interact with fraudulent interfaces. The framework underscores that URL manipulation and interface imitation continue to be core elements of phishing attacks.

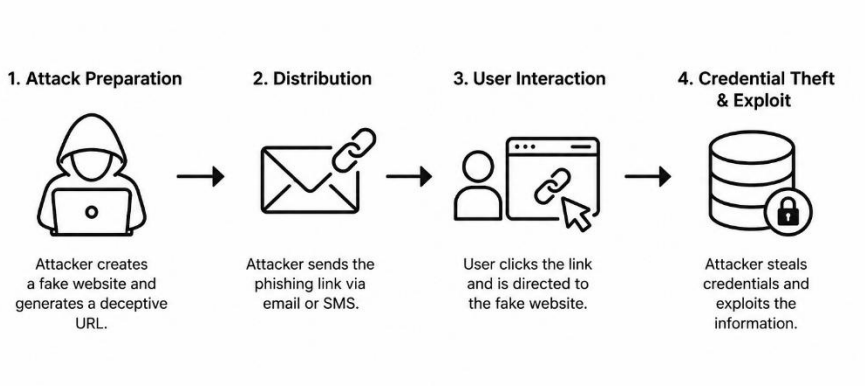

Figure 1. Conceptual Framework of a Typical Phishing Attack Lifecycle, from The Infrastructure Setup Phase to The Data Exfiltration Phase

One of the most useful behavioral signals among the visible phishing signs has been the structural alteration of URLs. Attackers often create URLs that keep a sense of superficial familiarity but introduce minor inconsistencies. Such anomalies include

over-expansion of the path, symbolic obfuscation, misleading subdomain composition, and deceptive lexical groupings. The CANTINA+ framework introduced by Xiang et al. (2011) revealed that the performance of phishing detection can be significantly improved if several structural and lexical indications are assessed simultaneously rather than individually. Sahingoz et al. (2019) later reported similar results showing that NLP-oriented lexical characteristics taken straight from URLs might yield good detection performance without using third-party services or manually maintained blacklists.

Additional research further highlighted the usefulness of structural feature analysis. Rao and Pais (2019) suggested that phishing behavior is typically driven by interactions between several URL and domain-level features, instead not by isolated signs alone. Similarly, da Silva et al. (2020) called attention to the fact that many old heuristic indications are becoming less effective as phishing tactics advance, especially because attackers increasingly use realistic-looking HTTPS certificates and domain naming conventions.

In addition to these improvements, machine learning and deep learning techniques have grown prevalent in phishing detection research. These surveys collectively characterize a broader shift toward AI-driven phishing detection architectures with ensemble learning, neural networks, and hybrid classification systems (Basit et al., 2021; Catal et al., 2022; Safi and Singh, 2023). While these approaches tend to produce high predictive performance, recent bibliometric analyses reveal that the current state of phishing research is still predominantly centered around optimization accuracy and developments in model architecture and less around interpretability and statistical feature behavior (Popescul and Radu, 2025).

This constraint is critical in cybersecurity situations where understanding the behaviour of suspicious features may be as crucial as getting accurate classification outcomes. Nguyen et al. (2024) claim that phishing detection systems should be able to provide interpretable explanations that reveal the structural and contextual elements that lead to harmful classification choices. Uddin et al. (2025) also pointed to the increasing significance of transparency, feature interaction analysis, and human-centric interpretation in explainable phishing detection systems.

There is a large literature on phishing detection, however there are few research on the characteristics of phishing in an exploratory and statistical perspective. Most of the current research is centered on the assessment of predictive performance, while relatively less emphasis is given to distributional asymmetry, inter-feature dependency, structural irregularity and class-specific statistical behavior. Existing works on phishing have focused on prediction optimization and classification accuracy. In contrast, the current study aims for interpretable structural analysis through exploratory statistical investigation. Instead than presenting a novel categorization architecture, this work aims at exploring the way in which structural indications related to phishing interact, distribute, and describe phishing behavior in modern web contexts.

Driven by this research gap, the current study utilizes a data-driven exploratory methodology to examine the features of phishing websites using publicly available data. The analytical process implements distribution analysis, non-

parametric hypothesis testing and correlation-based investigation to assess URL-based and domain-related structural indicators linked with phishing behaviors. This work seeks to contribute to the larger body of machine learning-oriented phishing detection research and to make phishing analysis more transparent and explainable by focusing on statistical interpretation and structural insight.

The primary contributions of this work can be summarized as:

1. In this paper, we provide an interpretable exploratory framework for assessing the behavior of phishing websites based on structural, symbolic and domain-related characteristics.
2. The study finds statistically significant structural differences between phishing and legal web sites by distributional analysis, non-parametric testing of hypothesis and correlation-based investigation.
3. The analysis indicates that signs related to phishing have inter-related structural links among lexical extension, symbolic manipulation, and domain instability, not separate suspicious conduct.
4. The results demonstrate that several traditional heuristic indicators, particularly individual symbolic features such as hyphen usage, exhibit limited discriminative capability when considered independently in modern phishing environments.
5. This paper offers an explainable statistical perspective to the examination of phishing, supplementing prior machine learning oriented research on phishing detection.

This exploratory project is guided by the following research questions:

1. **RQ1:** Do phishing websites have statistically distinct structural traits compared to legitimate websites?
2. **RQ2:** Which URL-based and domain-related features are the most significant indicators of phishing?
3. **RQ3:** How do the structural phishing indicators connect and interact in the behavior of phishing websites?

This study is different from prediction-oriented phishing studies which mostly focus on classification performance, as it focuses on interpretable explorative analysis to analyze the structural behavior of phishing websites. The goal is not to enhance forecast accuracy but rather to find statistically important structural trends that can increase transparency and explainability in phishing analysis.

2. Related Work

2.1. URL-Based Phishing Detection

Early phishing detection research relied heavily on blacklist systems and manually engineered heuristic rules. Although computationally efficient, these approaches struggled to cope with the short operational lifespan and rapidly changing structure of phishing websites. This limitation encouraged a gradual transition toward URL-centric analytical methods emphasizing lexical and structural characteristics.

A prominent early framework in this area was CANTINA by Zhang et al. (2007), followed by the feature-rich CANTINA+ framework by Xiang et al. (2011). These experiments demonstrated that the detection of phishing behavior improves significantly when numerous structural and lexical features are jointly considered

rather than in isolation. Jain and Gupta (2019) later examined similar client-side approaches and found that hyperlink-related features derived directly from HTML source code can effectively identify phishing websites without extensive dependence on external third-party services.

Additional research offered more evidence of the analytic use of URL-based indicators. Sahingoz et al. (2019) showed that NLP-based lexical analysis directly on URL structures can achieve good performance in phishing detection by using token composition and symbolic anomalies. Likewise Rao and Pais (2019) stressed the point that phishing activity is often an outcome of interactions between a number of URL and domain-level features and not due to individual indicators alone. Separately, da Silva et al. (2020) have claimed that many old heuristic indications have become increasingly less accurate as phishing methods improve, and increasingly mimic legal HTTPS-enabled environments.

In addition to lexical modification, domain-related behavior has now become an essential analytical feature. Mohammad et al. (2014) underlined the significance of domain linked qualities to distinguish phishing sites from legal websites, particularly with regards to unstable or short-lived domain activity indicators. Taken together, these findings positioned structural URL and domain analysis as one of the most viable and scalable paths for phishing detection research, especially for detecting subtle phishing anomalies that would not be captured by static blacklist methods alone.

2.2. Machine Learning and Deep Learning Approaches

With the growth of phishing datasets and the development of increasingly complex feature extraction techniques, machine learning has become a prominent paradigm in phishing detection research. Early work by Abdelhamid et al. (2014) shown that associative classification algorithms can increase the efficacy of phishing detection, preserving decision structures that are generally interpretable. Ali and Ahmed (2019) further show that the quality of feature engineering frequently has a greater impact on phishing detection performance than the complexity of the model alone.

Ensemble and optimization-oriented methodologies have been widely utilized later to improve the prediction robustness and computational efficiency. Research on stacking, voting classifiers, meta-learning, feature selection, and optimization methods consistently shows improved phishing detection performance across various datasets (Alsariera et al., 2020; Li et al., 2019; Nisreen Innab Ahmed Abdelgader Fadol Osman, 2024; Orunsolu et al., 2022; Saeed M. Alshahrani Nayyar Ahmed Khan, 2022). These research jointly emphasized the significance of adaptive ensemble structures and efficient feature selection procedures to cope with the evolution of phishing behaviour while lowering false alarm rates and processing overhead. Further ensemble-oriented phishing frameworks have combined clustering and classification strategies for the improvement of predictive robustness on heterogeneous phishing datasets, reinforcing the effectiveness of hybrid ensemble architectures in phishing analytics (Alsharaiah et al., 2023).

This transformation was further expedited by the rapid expansion of deep learning. CNN, LSTM, DNN, DBN, and hybrid neural architectures have been shown

in several studies to be effective in identifying phishing-related structural patterns and hidden feature relationships that are difficult to capture using conventional machine learning pipelines (Do et al., 2021; Somesha et al., 2020; Wang et al., 2019; Yang et al., 2019; Yi et al., 2018).

Recent phishing detection systems have focused on lightweight neural architectures, feature-efficient learning processes, and automated optimization strategies. The 1D CNN models, feature ranking methods, topic modeling, Variational Autoencoders (VAE), and hyperparameter optimization frameworks have shown good prediction capabilities with improvement in scalability and computational efficiency (Barik et al., 2025; Gualberto et al., 2020; Mousavi and Bahaghighat, 2025).

Recent studies have also evolved towards intelligent and adaptive phishing detection architectures integrating reinforcement learning, swarm optimization, attention mechanisms and real-time detection. Khan and Unhelkar (2024) proposed a multilayer anti-phishing framework (Q-learning, LB-LSTM, and swarm optimization) for social media phishing environments. Asiri et al. (2024) proposed PhishingRTDS, a BiLSTM-attention framework to detect sophisticated phishing strategies including Browsers-in-the-Browser (BiTB) attacks and TinyURL-based obfuscation. Similarly, Omari (2023) studied adaptive phishing detection systems to enhance resilience against changing phishing attacks.

This progress has, however, come at the expense of an increasing dependence on complex learning systems, leading to a significant trade-off. Many of the high-performance methods for phishing detection are black box predictive systems that offer little explanation of the role played by specific criteria in the categorization decision. This constraint has raised the interest in explainable and interpretable phishing detection frameworks able to facilitate human-centered cybersecurity investigation.

2.3. Explainable and Interpretable Phishing Detection

The growing prevalence of AI-based phishing detection systems has raised questions on the interpretability, transparency, and accountability of these models. Capuano et al. (2022) in their survey of explainable cybersecurity research, state that black-box AI systems may hinder the trust of analysts, impede forensic investigation, and restrict decision transparency within security-critical domains. These constraints are particularly relevant in phishing detection, as analysts typically need interpretable evidence to support threat validation and security evaluation.

Accordingly, recent attempts on explainable phishing detection have moved focus to transparent analytical frameworks that can show how structural and contextual variables contribute to malicious classification conclusions. (Nguyen et al., 2024) argued that the interpretability is a key requirement for the phishing detection systems, which should not be used just as binary classifiers but also provide interpretable explanations. Likewise, Uddin et al. (2025) stressed the significance of feature interaction analysis and transparent behavioral interpretation to comprehend the evolving tactics of phishing.

Additional studies have investigated the use of explainable machine learning approaches in cybersecurity decision-making frameworks. Rahebi and Rahebi (2025) claimed that interpretable phishing detection frameworks can help increase analyst's

confidence, provide security audits, and reduce the uncertainty of automated categorization systems. Similarly, Asiri et al. (2024) note that many current papers on phishing detection still focus on predictive optimization with little discussion on analytical transparency and feature interpretability.

Taken together, these studies reflect a larger methodological shift in the study of phishing. Although predicted accuracy still plays a large role, there is a growing consensus that knowing the reason why a website is suspicious can also be useful for practical cybersecurity analysis.

2.4. Surveys and Emerging Research Trends

The increased sophistication of phishing attempts has resulted in a large body of survey literature analyzing methodological change, new research goals, and open difficulties in phishing detection research. Reviews by Basit et al. (2021), Mahdavifar and Ghorbani (2019), Catal et al. (2022), and Safi and Singh (2023) collectively documented the rapid shift towards machine learning, deep learning, ensemble learning and automated phishing detection architectures that can manage large-scale datasets with little manual intervention is well documented.

Recent trend-oriented and bibliometric studies have further emphasized growing research directions related to adaptive phishing infrastructures, hybrid AI architectures, and explainable cybersecurity systems. Popescul and Radu (2025) noted that current phishing research is quite concentrated on the evaluation of predictive optimization and deep learning performance, whereas Kytidou et al. (2025) noted growing worries around adversarial manipulation, dataset bias, and AI-generated phishing content. Their review also pointed out the growing importance of hybrid detection frameworks that integrate machine learning, deep learning, and Large Language Model (LLM)-oriented techniques.

Wilk-Jakubowski et al. (2025) also observed that the research on phishing detection is trending towards explainable AI, behavioral analysis, and feature interaction modeling. However, relatively few research have been performed to study phishing features by explicitly exploratory statistical analysis, which combines distributional assessment, hypothesis testing, and correlation-based feature exploration into a single analytical framework.

This gap motivates the present study, which employs a feature-centric exploratory framework, stressing statistical interpretation, structural irregularity analysis, and feature interaction exploration, rather than predictive optimization alone. This work is an attempt to enhance the landscape of research on machine learning based phishing detection by providing a better interpretable picture of the behaviour of phishing websites through the combination of distribution analysis, non-parametric hypothesis testing and correlation based inspection.

3. Methodology

This study employs a data-driven exploratory framework to investigate structural phishing characteristics using publicly available phishing website data. The analysis focuses on statistical interpretation, feature interaction, and structural behavior associated with phishing websites rather than predictive optimization alone.

The overall workflow consists of dataset preparation, preprocessing, feature selection, exploratory statistical analysis, and correlation-based structural evaluation.

3.1. Dataset and Feature Representation

This study employs a data-driven exploratory framework to analyze structural phishing characteristics using a publicly available phishing detection dataset. This approach is aimed at statistical interpretation and investigation of feature behavior rather than prediction optimization. The dataset contains cases of phishing and genuine websites described using URL-based, content-based and domain related features. Public phishing datasets provide reproducible benchmarks for feature behavior and analytical consistency in cybersecurity research (Linh and Hung, 2025; Vrbančič et al., 2020).

This study is mainly based on the lightweight structural indicators from URL and domain information, which is different from content-intensive phishing detection methods that need to generate the web site or rely on external reputation services. It has been established in previous work that lexical analysis of URLs yields computationally efficient and scalable phishing signals (Sahingoz et al., 2019; Xiang et al., 2011).

The original dataset contains 11,430 occurrences of URLs and 87 extracted features such as lexical indicators, properties of hyperlinks, HTML characteristics and domain-related factors. After preprocessing and redundancy filtering, 20 representative features were kept for statistical analysis. Preprocessing included duplicate inspection, missing value verification, feature consistency evaluation and redundancy filtering. Features with relatively sparse distributions, little variability or overlapping structural behavior were eliminated to improve interpretability and reduce analytical noise.

A general statistical summary of the dataset used in this study is presented in **Table 1**.

Table 1. Presents The Statistical Characteristics of The Dataset Used in This Study

Attribute	Value
Total Samples	11,430
Legitimate Samples	~50%
Phishing Samples	~50%
Original Features	87
Selected Features After Preprocessing	20

The retained features were selected to preserve structurally informative indicators while reducing redundancy and low-variance attributes. **Table 2** presents the representative features retained after preprocessing along with their analytical relevance in phishing structure interpretation.

Table 2. Presents Representative Features Retained After Preprocessing

Feature	Category	Retained Reason
length_url	URL-based	Captures abnormal URL expansion
nb_hyphens	URL-based	Indicates symbolic obfuscation
nb_dots	URL-based	Represents subdomain complexity
nb_at	URL-based	Detects suspicious symbolic insertion

Feature	Category	Retained Reason
domain_age	Domain-related	Reflects infrastructure stability
domain_registration_length	Domain-related	Measures domain persistence
https_token	URL-based	Detects deceptive HTTPS usage
ratio_digits_url	URL-based	Captures lexical irregularity
longest_word_path	URL-based	Indicates path manipulation
shortest_word_host	URL-based	Measures host inconsistency

To keep the paper concise, only some typical features are shown here while the whole processed feature matrix and the preprocessing outputs are supplied in the supplemental materials for reproducibility and transparency.

3.2. Feature Engineering and Extraction

Our analytical approach uses handmade feature extraction methods focused on phishing, based on past works on phishing detection, specifically the framework presented by Hannousse and Yahiouche (2021). The extracted indications are URL-based features, content-based features, and domain-related features. URL oriented features are used to model lexical composition, symbolic manipulation, suspicious subdomain structures, anomalous redirection patterns, digit ratios, shortening services, and phishing related lexical cues. Content-based traits include hyperlink structure, iframe activity, popup behavior, properties of dangerous anchors, and suspicious form actions from the webpage HTML content.

Domain-related indicators include WHOIS registration information, DNS behavior, domain age, traffic indicators and registration length, as phishing infrastructures often exhibit unstable or short-lived domain behavior. The feature extraction has been performed utilizing a set of automated Python scripts that involve url parsing, HTML structural analysis, WHOIS lookup and metadata retrieval.

3.3. Data Preprocessing

Prior to statistical analysis, several preprocessing processes were performed to increase the consistency of the dataset and the reliability of the study. Missing numerical values were imputed by median and categorical inconsistencies were corrected using mode substitution. Median-based imputation was adopted since several of the variables had skewed distribution and probable outliers; so, the median is more robust than the mean-based imputation. After the data cleaning, the dataset was divided into classes of phishing and authentic websites for comparative statistical analysis. Then, feature refining was undertaken by looking at the variance, mutual information and correlation to eliminate redundancy and increase interpretability. Features with very low variance or too much multicollinearity were removed from the analysis.

This preprocessing method is consistent with previous phishing research that highlight the importance of the quality of the features and the structural relevance for phishing analytics (Chiew et al., 2019; Rao and Pais, 2019). Twenty selected features were kept in the final analysis dataset. Records with unresolved missing values and substantially redundant attributes were eliminated during preprocessing.

3.4. Exploratory and Statistical Analysis Framework

The analytical approach was intended to analyze phishing activity from an exploratory and interpretative standpoint, rather than for predictive classification.

This approach integrates distribution analysis, statistical comparison, and correlation-based investigation to investigate structural differences between classes of phishing and authentic websites.

The initial part of the investigation consisted of the exploratory distribution analysis through the use of density plots and boxplots to understand the central tendency, variability, skewness and outlier behavior. This type of visualization is particularly beneficial in the case of phishing analytics since the malicious websites usually have asymmetric and uneven structure patterns.

The second step was dedicated to a statistical comparison between the phishing and legal website categories. Due to the violations of normality assumptions on various variables, the Mann–Whitney U test was conducted for the non-parametric comparison and the rank-biserial correlation (RBC) was utilized to evaluate the magnitude of effect size.

In the last step, we explored correlations using Spearman rank correlation to uncover monotonic links between parameters associated with phishing. We then visualized the associations between features and between structurally similar phishing signs using correlation heatmaps.

Previous research has demonstrated that phishing behaviour is often the result of the interplay of lexical, structural and domain-level irregularities, rather than individual feature anomalies (Rao and Pais, 2019; Xiang et al., 2011). The suggested approach combines exploratory visualization, non-parametric statistical comparison and correlation-based analysis, aiming at enabling more interpretable and explainable phishing investigation.

By integrating exploratory visualization, non-parametric statistical comparison, and correlation-based analysis, the present framework aims to provide a more interpretable understanding of phishing website behavior while supporting explainable cybersecurity analysis.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i represents the rank difference between paired observations and n denotes the number of observations.

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

where n_1 and n_2 denote the sample sizes of the compared groups, while R_1 represents the rank sum of the first group. A statistical significance threshold of $p < 0.05$ was used throughout the analysis.

Unlike black-box phishing detection frameworks emphasizing predictive optimization, the proposed workflow prioritizes interpretable structural behavior analysis through statistically explainable feature relationships.

3.5. Implementation Environment

All preprocessing, feature extraction, statistical analysis and visualization was done using Python 3.11 in a Jupyter Notebook Environment. Data preprocessing and manipulation were done using Pandas and NumPy. Statistical analyses were performed with SciPy and Pingouin packages. Visualization processes were created using Matplotlib and Seaborn. The analytical workflow integrated URL parsing,

HTML structure analysis, WHOIS inspection, DNS-based feature retrieval, and domain-related metadata extraction through customized Python extraction scripts adapted from prior phishing detection studies.

The workflow additionally included missing value handling, feature filtering, distribution analysis, non-parametric hypothesis testing, and correlation-based exploration. The Mann–Whitney U test and Spearman rank correlation were applied using a statistical significance threshold of $p < 0.05$.

To improve reproducibility and analytical consistency, all preprocessing and statistical procedures were executed within a unified computational environment, and all figures were exported in publication-quality format.

Figure 2. Illustrates The Overall Analytical Workflow Applied in This Study

4. Results

4.1. Distributional Characteristics of URL-Based Features

The exploratory analysis showed significant distributional disparities between phishing and authentic websites, especially in relation to URL-related variables. Several variables showed asymmetric distributions with extended right tail patterns, suggesting the presence of structural abnormalities. Among the features analyzed, URL length showed one of the most evident differences between phishing and legitimate classes. **Figure 3** shows that both classes were positively skewed, but the distribution of phishing URLs had far larger tails with more extreme outliers. The skewness for the phishing class is 1.90 which indicates that the distribution is heavily right skewed with some outliers of long URLs. Legitimate websites likewise displayed positive skewness but with significantly less structural heterogeneity.

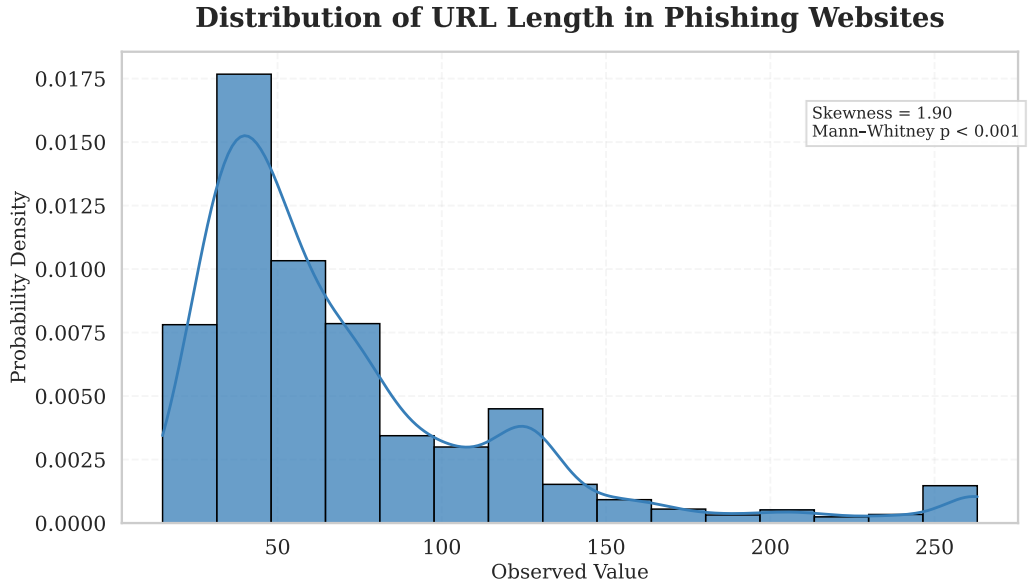

Figure 3. Distribution of Url Length Across Phishing and Legitimate Website Classes, Illustrating Positively Skewed Distributions and Heavier Structural Dispersion Among Phishing URLs

Phishing URLs also showed higher dispersion and greater variability than legal URLs. Some phishing samples had abnormally long URL structures while real websites were more centered around moderate URL lengths. Distributional analysis also showed that symbolic indications like special characters and encoded symbols were more frequently connected with structurally sophisticated phishing URLs. In contrast, valid URLs tended to have smaller symbolic variability and lower structural dispersion.

Overall, the results present a stronger asymmetry, a greater lexical expansion and a higher variability in the phishing-related URL structures than in the legal online entities.

4.2. Statistical Comparison of Structural Features

To investigate the structural differences between phishing and legitimate websites, non-parametric statistical analysis was performed using the Mann–Whitney U test. The analysis focused on selected lexical and domain-related features commonly associated with phishing behavior. In addition to statistical significance testing, rank-biserial correlation (RBC) was employed to quantify the practical magnitude of observed differences. The statistical comparison results are summarized in Table 3 and visualized in Figure 4.

Table 3. Statistical Summary of Selected Features.

Feature	Phishing Mean	Legitimate Mean	Phishing Skewness	Legitimate Skewness	p-value	RBC	Effect Size
URL Length	74.87	47.38	1.90	1.42	<0.001	0.329	Medium
Number of Hyphens	0.79	1.21	3.84	2.91	0.557	0.006	Negligible
Number of "@" Symbols	0.04	0.00	4.72	0.00	<0.001	0.043	Negligible
Domain Age	3031.15	5093.94	0.48	-0.22	<0.001	-0.358	Medium
Domain Registration Length	360.77	624.29	2.94	4.81	<0.001	-0.212	Small

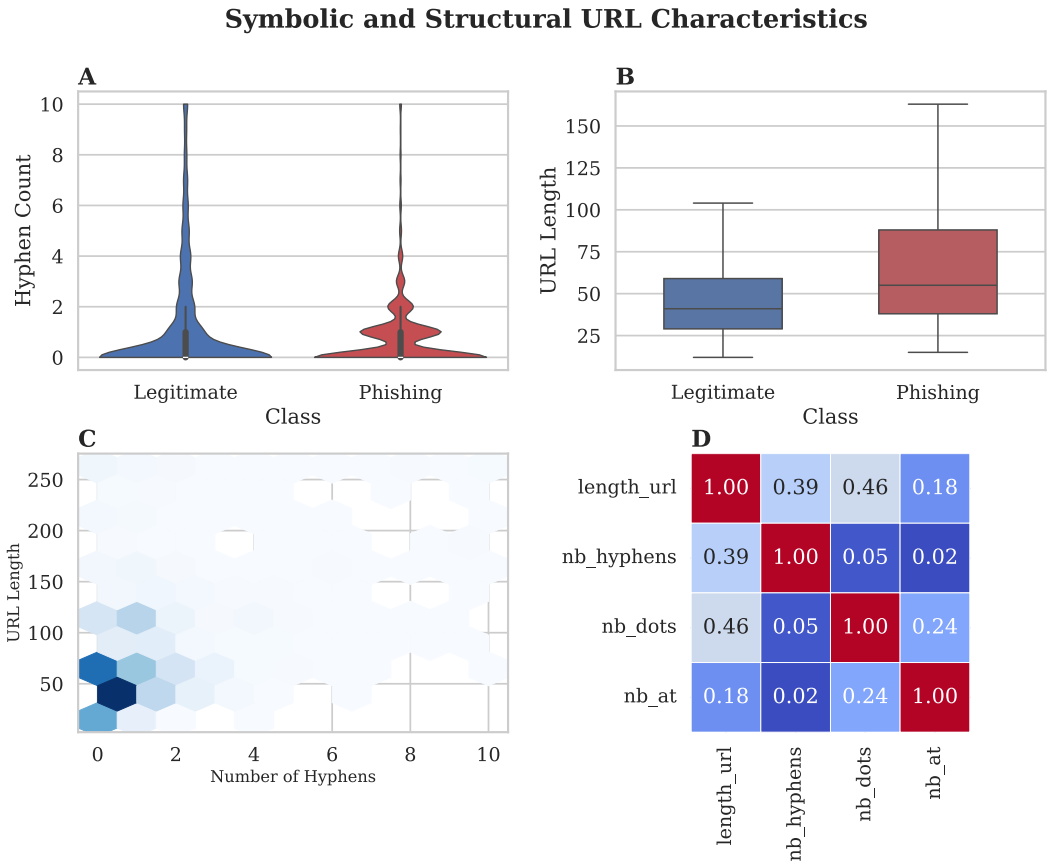

Figure 4. Comparative Visualization of Selected Structural Features Across Phishing and Legitimate Website Classes

Panel A shows a substantial overlap in hyphen counts between phishing and legitimate websites. This result is consistent with the non-significant Mann–Whitney U test ($p = 0.557$) and negligible effect size (RBC = 0.006), indicating that hyphen usage alone is not a strong discriminator between the two classes.

Panel B highlights clear differences in URL length. Phishing URLs tend to be longer and more variable than legitimate URLs, producing one of the strongest statistical differences observed in this study ($p < 0.001$, RBC = 0.329).

Panel C illustrates the joint distribution of URL length and hyphen usage in phishing websites. Most observations are concentrated at low hyphen counts and moderate URL lengths, suggesting that phishing URLs do not necessarily rely on extensive hyphen usage despite their longer structures.

Panel D presents the correlations among selected lexical features. URL length shows moderate positive correlations with both the number of dots ($\rho = 0.46$) and hyphens ($\rho = 0.39$), indicating that longer URLs generally contain more structural components. In contrast, correlations involving the "@" symbol remain weak.

Effect size analysis further supports these findings. URL length (RBC = 0.329) and domain age (RBC = −0.358) exhibited medium effect sizes, while domain registration length showed a small effect (RBC = −0.212). Hyphen usage (RBC = 0.006) and "@" symbol frequency (RBC = 0.043) produced negligible effect sizes, indicating limited practical contribution to class separation.

Based on the magnitude of the rank-biserial correlation values, domain age (|RBC| = 0.358), URL length (RBC = 0.329), and domain registration length (|RBC| = 0.212) emerged as the most discriminative indicators among the features examined. These findings suggest that domain maturity and URL structural expansion provide stronger phishing-related signals than individual symbolic indicators, thereby directly addressing RQ2 concerning the most informative URL-based and domain-related characteristics associated with phishing behavior.

Domain-related features also differed substantially between classes. Phishing domains were generally younger and registered for shorter periods than legitimate domains ($p < 0.001$), reinforcing the importance of domain maturity as a phishing indicator.

Overall, the results suggest that phishing websites are more strongly characterized by URL length and domain-related attributes than by individual symbolic indicators.

4.3. Correlation Structure and Feature Interaction Analysis

Then we did a correlation-based analysis to check the structural links among phishing-related characteristics. Several variables exhibited asymmetric and non-normal distributions and the Spearman rank correlation was used to quantify monotonic feature correlations. The resulting correlation structure is presented in **Figure 5**.

Figure 5. Spearman correlation heatmap showing relationships among selected phishing-related features

Several modest positive associations were found for URL length, symbolic indicators, and encoded character properties. The URL length was positively related with the amount of hyphens and the number of "@" symbols, suggesting structurally long URLs have often patterns of symbol manipulation. Negative relationships additionally appeared between domain age and multiple structural complexity indicators. Domain age exhibited largely negligible correlations with the structural indicators examined. Although weak associations were observed with several symbolic features, the overall relationship between domain age and structural complexity remained limited.

To further examine the interaction between structural expansion and symbolic manipulation, scatter-based regression analysis was conducted between URL length

and the number of hyphens across phishing and legitimate website classes. The resulting visualization is shown in **Figure 6**.

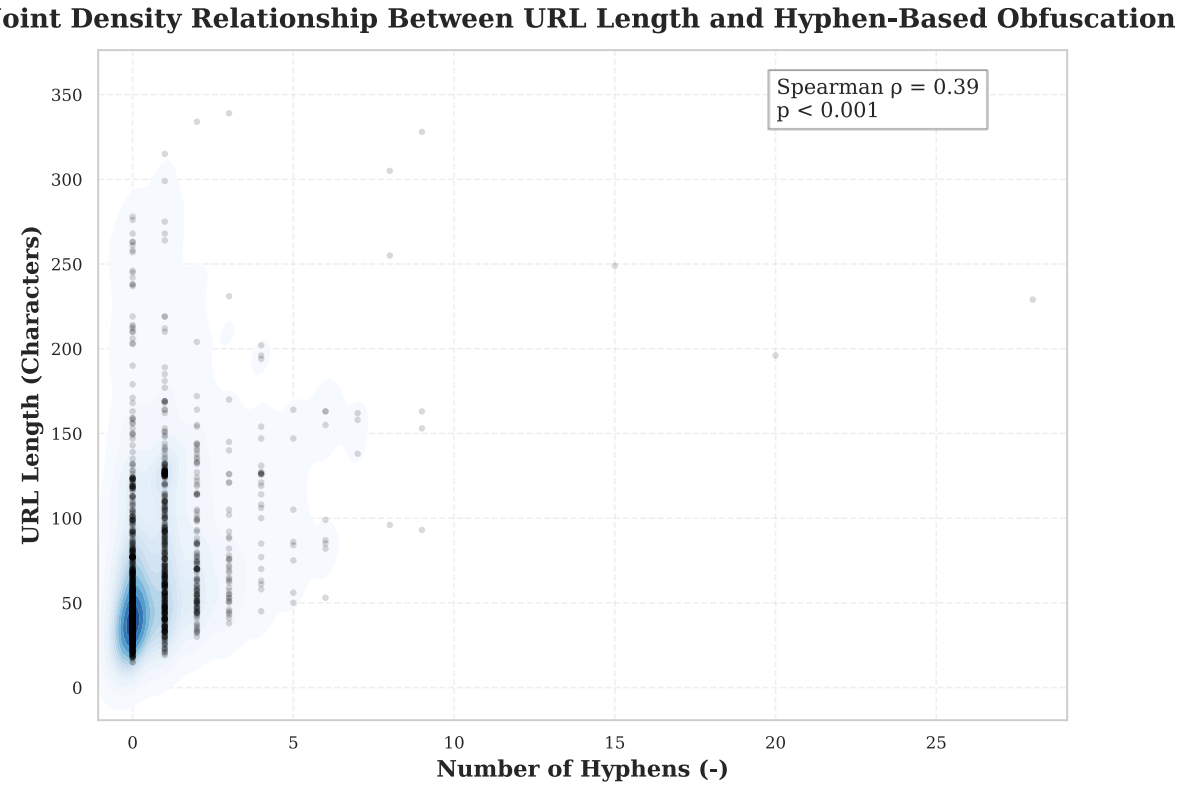

Figure 6. Relationship between URL length and hyphen-based obfuscation in phishing websites. The analysis reveals a moderate positive association between URL length and hyphen usage (Spearman ρ = 0.39, p < 0.001), suggesting that lexical expansion and hyphen-based obfuscation frequently co-occur in phishing-related URL structures

Figure 6 further illustrates the relationship between URL length and hyphen usage. A moderate positive association (ρ ≈ 0.39) suggests that longer URLs tend to contain a higher number of hyphens. This pattern supports the interpretation that lexical expansion and hyphen-based obfuscation often co-occur within phishing-related URL structures. Rather than acting independently, these indicators appear as interconnected structural mechanisms that increase URL complexity.

5. Discussion

5.1. Structural Characteristics of Phishing URLs

The results show that the structure of phishing websites is far more irregular than that of legal websites, especially in terms of lexical expansion and symbolic manipulation. The observed substantially right-skewed distributions of phishing URLs show that attackers often use extended URL structures with many path components, encoded strings, and syntactic padding methods. These findings are similar with earlier studies on phishing, which have found that URL complexity and linguistic irregularity are major indications of phishing (Sahingoz et al., 2019; Xiang et al., 2011). Similar findings have also been published by Rao and Pais (2019). They

stated that phishing activity often results from combinations of many structural anomalies, instead of standalone suspicious traits.

Our results further highlight the analytical utility of lightweight URL-based analysis for phishing identification. In contrast to webpage-rendering or content-intensive techniques, lexical URL indicators give interpretable structure information and relatively low processing overhead. In this work, we focus on URL-based and domain-related data because such features may be extracted efficiently without using expensive webpage rendering or dynamic behavior analysis. Lightweight structural analysis is especially relevant for real-time phishing screening environments, browser-side protection systems and scalable cybersecurity monitoring apps.

The observed distributional asymmetry shows that phishing URLs often adopt structural expansion tactics that include path elongation, symbolic insertion and lexical padding processes.

5.2. Declining Reliability of Traditional Heuristic Indicators

A significant discovery is that a lot of the classic heuristic indications cannot reliably differentiate phishing sites from authentic environments any longer. In particular, use of hyphens did not exhibit statistically significant class disparities, despite having been historically seen as a suspicious linguistic signal. This research indicates that several traditional heuristic assumptions are losing discriminative power as modern web environments get more complicated. Now legitimate sites often use multi-level pathways, tracking parameters, dynamic routing systems and marketing-oriented naming standards that echo lexical patterns previously associated with phishing activity.

The findings are consistent with previous studies suggesting that modern phishing infrastructures increasingly mimic legitimate web environments through realistic domain naming conventions and other trust-inducing design strategies, thereby reducing the effectiveness of some traditional heuristic indicators. As such, the efficiency of phishing detection systems that rely solely on separate heuristic criteria may be diminished in modern cybersecurity contexts.

5.3. Feature Interaction and Explainable Phishing Analysis

The correlation research showed that signs related to phishing often act as interrelated behavioral mechanisms, rather than isolated anomalies. Longer URLs were consistently linked to greater symbolic manipulation, more encoded characters, and more structural irregularity. This interaction-oriented behavior supports recent explainable cybersecurity research emphasizing the importance of feature relationships and contextual interpretation in phishing analysis (Nguyen et al., 2024; Uddin et al., 2025). These findings reinforce the importance of interpretable phishing analysis frameworks capable of explaining structural phishing behavior beyond opaque predictive classification outcomes. Rather than appearing through single suspicious attributes, phishing behavior appears to emerge through coordinated structural patterns involving lexical complexity, symbolic manipulation, and domain instability.

From an interpretability perspective, these findings highlight the limitations of purely black-box predictive approaches that prioritize classification performance without explaining structural phishing behavior. The present exploratory framework

instead emphasizes transparent statistical interpretation capable of supporting explainable phishing intelligence analysis.

5.4. Practical Implications

The findings suggest that phishing detection systems may benefit from evaluating multiple structural indicators collectively rather than relying on isolated heuristic rules. The observed interaction between URL complexity, symbolic manipulation, and domain instability indicates that phishing behavior is better represented through behavioral feature combinations.

The results additionally support the development of lightweight and interpretable phishing detection mechanisms suitable for browser-side protection, security monitoring systems, and explainable cybersecurity applications. Because URL-based indicators can be extracted without intensive webpage rendering or third-party infrastructure dependence, the proposed analytical perspective may support scalable phishing screening environments.

The reduced discriminative capability of several traditional indicators further highlights the need for adaptive phishing analysis frameworks capable of responding to evolving phishing strategies and increasingly legitimate-looking phishing infrastructures.

5.5. Limitations and Future Work

Several limitations should be acknowledged. First, the study relies on a publicly available phishing dataset, which may not fully reflect the dynamic and rapidly evolving nature of real-world phishing campaigns. Temporal shifts in attacker behavior, regional differences in phishing strategies, and dataset-specific feature distributions may affect the generalizability of the identified structural patterns. Therefore, the findings should be interpreted as evidence derived from a large-scale benchmark dataset rather than as universally representative characteristics of all phishing environments.

Secondly, the analysis mainly considers lexical and domain-related features without considering the webpage rendering behavior, visual similarity analysis or network level indications. Lightweight structural analysis has certain computing advantages, however there can be multimodal aspects in phishing activity beyond URL construction alone.

Therefore, future study can employ the analysis of temporal phishing evolution, explainable artificial intelligence methods, and multimodal phishing indicators to build more adaptive and interpretable phishing intelligence frameworks. Additional analysis with real-time phishing streams and cross dataset validation may also improve generalizability and analytical robustness.

6. Conclusion

In this work, we show that combinations of structural abnormalities, symbolic manipulation and domain-related instability are better at capturing phishing activity than individual suspicious indications. The exploratory analysis showed that phishing websites have certain structural features, especially in terms of URL construction, symbolic obfuscation and domain lifecycle behavior. The results demonstrate that phishing URLs have a stronger right-skewed distribution and a higher structural variability than legal websites, showing that URL expansion is still

a prevalent obfuscation method. Correlation research also showed that phishing-related indicators behave as interlinked structural mechanisms, with longer URLs often linked to more symbolic complexity and syntactic manipulation.

The statistics also show that several old phishing indications still matter such as symbol manipulation in URLs and domain instability. However, several traditional heuristic indicators appear to be less discriminative in contemporary web environments, highlighting the need to evaluate multiple structural characteristics collectively rather than relying on isolated features.

Unlike the prediction-oriented phishing research mainly focusing on classification performance, this work is dedicated to the interpretable statistical investigation to improve the transparency of the analysis on the phishing behaviors. The proposed framework demonstrates how exploratory visualization, statistical comparison, and feature interaction analysis can improve the interpretability of phishing investigation. Methodologically, this research offers an interpretable exploratory approach that prioritises structural comprehension over predictive optimisation alone. The study illustrates that the investigation of phishing behavior may be performed by combining distribution analysis, non-parametric statistical testing and correlation-based exploration with transparent statistical interpretation instead of black-box predictive models.

Future research should validate the proposed framework using real-time phishing streams, cross-dataset evaluation, and multimodal phishing indicators. Integrating explainable AI techniques and temporal phishing evolution analysis may further enhance the adaptability and interpretability of phishing intelligence systems.

Data Availability Statement

The dataset used in this study was obtained from publicly available sources intended for cybersecurity research and educational purposes. Processed analytical data and supplementary materials supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: "*The author utilized AI-assisted tools to generate the initial conceptual layout for Figure 1, which was subsequently verified and refined for technical accuracy.*"

Appendix A. Feature Description and Analytical Configuration

Appendix A.1. Selected Feature Representation

Table A1 summarizes the principal phishing-related features analyzed in this study. The selected attributes include lexical URL indicators, symbolic manipulation features, and domain-related characteristics commonly associated with phishing behavior analysis.

Table A1. Selected phishing-related feature descriptions.

Feature	Description
URL Length	Total number of characters contained in the URL
Hostname Length	Character length of the hostname component
Number of Dots	Total number of “.” symbols in the URL

Feature	Description
Number of Hyphens	Total number of "-" symbols in the URL
Number of "@" Symbols	Frequency of "@" characters within the URL
Number of Slash Symbols	Total number of "/" symbols
Number of Digits	Numerical character frequency within the URL
HTTPS Token	Presence of HTTPS-related lexical indicators
Subdomain Count	Number of subdomain levels
URL Redirection	Presence of redirection-related behavior
External Redirection	Redirection toward external domains
Character Repetition	Consecutive repeated character patterns
Average Word Length	Mean token length extracted from URL segments
Longest Word Length	Maximum token length within URL segments
Phishing Hints	Presence of phishing-associated lexical keywords
Domain Age	Estimated operational age of the domain
Domain Registration Length	Remaining registration duration of the domain
DNS Record Availability	Presence of valid DNS registration
WHOIS Registration	Availability of WHOIS registration information
Statistical Report Indicator	Suspicious domain/IP statistical blacklist indicator

The feature extraction framework integrates URL parsing, lexical decomposition, HTML structure inspection, WHOIS analysis, DNS inspection, and external domain-related metadata retrieval procedures. Several handcrafted phishing-oriented indicators were adapted from previous phishing detection frameworks proposed in prior studies.

Appendix A.2. Data Preprocessing Procedure

The preprocessing workflow included duplicate removal, missing value handling, feature filtering, and statistical consistency inspection prior to exploratory analysis. Numerical variables exhibiting asymmetric distributions were evaluated using skewness inspection and non-parametric statistical procedures.

Several highly sparse and low-variance attributes were excluded during preprocessing to reduce redundancy and improve analytical interpretability. Feature selection additionally prioritized phishing-related indicators demonstrating meaningful structural relevance and statistical variability.

The preprocessing workflow further included lexical normalization, URL decomposition, symbolic extraction, and domain-level parsing procedures to ensure analytical consistency across phishing and legitimate website instances.

Appendix A.3. Statistical Assumptions and Analytical Procedures

Because several phishing-related variables exhibited non-normal and asymmetric distributions, non-parametric statistical procedures were employed throughout the analysis. The Mann–Whitney U test was used to evaluate statistical differences

between phishing and legitimate website classes without assuming Gaussian distribution behavior.

Spearman rank correlation was additionally employed to evaluate monotonic relationships among phishing-related structural indicators. Effect size estimation using rank-biserial correlation was incorporated to evaluate the magnitude of observed differences beyond statistical significance alone.

All statistical analyses were conducted using a significance threshold of $p < 0.05$.

Appendix A.4. Supplementary Visualization Description

Additional supplementary visualizations were generated to support distributional inspection and feature interaction analysis. These supplementary plots included extended skewness distributions, feature density visualizations, outlier inspection plots, and additional correlation-based exploratory figures.

All figures were generated using Python-based scientific visualization libraries and exported in publication-quality high-resolution format to ensure graphical consistency and readability.

References

Abdelhamid, N., Ayesh, A., & Thabtah, F. (2014). Phishing detection based Associative Classification data mining. *Expert Systems with Applications*, *41*(13), 5948–5959. https://doi.org/https://doi.org/10.1016/j.eswa.2014.03.019

Ali, W., & Ahmed, A. A. (2019). Hybrid intelligent phishing website prediction using deep neural networks with genetic algorithm-based feature selection and weighting. *IET Information Security*, *13*(6), 659–669. https://doi.org/https://doi.org/10.1049/iet-ifs.2019.0006

Alsariera, Y. A., Adeyemo, V. E., Balogun, A. O., & Alazzawi, A. K. (2020). AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites. *IEEE Access*, *8*, 142532–142542. https://doi.org/10.1109/ACCESS.2020.3013699

Alsharaiah, M., Abu-Shareha, A., Abualhaj, M., Baniata, L., Adwan, O., Al-saaidah, A., & Oraiqat, M. (2023). A New Phishing-Website Detection Framework Using Ensemble Classification and Clustering. *International Journal of Data and Network Science*, *7*(2), 857–864. https://doi.org/10.5267/j.ijdns.2023.1.003

Asiri, S., Xiao, Y., Alzahrani, S., & Li, T. (2024). PhishingRTDS: A real-time detection system for phishing attacks using a Deep Learning model. *Computers & Security*, *141*, 103843. https://doi.org/https://doi.org/10.1016/j.cose.2024.103843

Barik, K., Misra, S., & Mohan, R. (2025). Web-based phishing URL detection model using deep learning optimization techniques. *International Journal of Data Science and Analytics*, *20*(5), 4449–4471. https://doi.org/10.1007/s41060-025-00728-9

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2021). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access*, *10*, 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. *Knowledge and Information Systems*, *64*(6), 1457–1500. https://doi.org/10.1007/s10115-022-01672-x

Chiew, K. L., Tan, C. L., Wong, K., Yong, K. S. C., & Tiong, W. K. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Information Sciences*, *484*, 153–166. https://doi.org/https://doi.org/10.1016/j.ins.2019.01.064

da Silva, C. M. R., Feitosa, E. L., & Garcia, V. C. (2020). Heuristic-based strategy for Phishing prediction: A survey of URL-based approach. *Computers & Security*, *88*, 101613. https://doi.org/https://doi.org/10.1016/j.cose.2019.101613

Do, N. Q., Selamat, A., Krejcar, O., Yokoi, T., & Fujita, H. (2021). Phishing Webpage Classification via Deep Learning-Based Algorithms: An Empirical Study. *Applied Sciences*, *11*(19). https://doi.org/10.3390/app11199210

ENISA. (2022). *ENISA Threat Landscape 2022*. ENISA. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

ENISA. (2023). *ENISA Threat Landscape 2023*. ENISA. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023

Gualberto, E. S., De Sousa, R. T., De B. Vieira, T. P., Da Costa, J. P. C. L., & Duque, C. G. (2020). From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection. *IEEE Access*, *8*, 76368–76385. https://doi.org/10.1109/ACCESS.2020.2989126

Hannousse, A., & Yahiouche, S. (2021). Web page phishing detection. *Mendeley Data*, *3*. https://doi.org/10.17632/c2gw7fy2j4.3

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J., & Park, P. S. (2024). Devising and Detecting Phishing Emails Using Large Language Models. *IEEE Access*, *12*, 42131–42146. https://doi.org/10.1109/ACCESS.2024.3375882

Jain, A. K., & Gupta, B. B. (2018). Towards detection of phishing websites on client-side using machine learning based approach. *Telecommunication Systems*, *68*(4), 687–700. https://doi.org/10.1007/s11235-017-0414-0

Jain, A. K., & Gupta, B. B. (2019). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, *10*(5), 2015–2028. https://doi.org/10.1007/s12652-018-0798-z

Khan, A. I., & Unhelkar, B. (2024). An Enhanced Anti-Phishing Technique for Social Media Users: A Multilayer Q-Learning Approach. *International Journal of Advanced*

Computer Science and Applications, *15*(1). https://doi.org/10.14569/IJACSA.2024.0150103

Kytidou, E., Tsikriki, T., Drosatos, G., & Rantos, K. (2025). Machine learning techniques for phishing detection: A review of methods, challenges, and future directions. *Intelligent Decision Technologies*, *19*(6), 4356–4379. https://doi.org/10.1177/18724981251366763

Li, Y., Yang, Z., Chen, X., Yuan, H., & Liu, W. (2019). A stacking model using URL and HTML features for phishing webpage detection. *Future Generation Computer Systems*, *94*, 27–39. https://doi.org/https://doi.org/10.1016/j.future.2018.11.004

Linh, D. M., & Hung, T. C. (2025). A feature-engineered dataset of benign and phishing URLs for machine learning and large language models evaluation. *Data in Brief*, *63*, 112162. https://doi.org/https://doi.org/10.1016/j.dib.2025.112162

Mahdavifar, S., & Ghorbani, A. A. (2019). Application of deep learning to cybersecurity: A survey. *Neurocomputing*, *347*, 149–176. https://doi.org/https://doi.org/10.1016/j.neucom.2019.02.056

Marchal, S., François, J., State, R., & Engel, T. (2014). PhishStorm: Detecting Phishing With Streaming Analytics. *IEEE Transactions on Network and Service Management*, *11*(4), 458–471. https://doi.org/10.1109/TNSM.2014.2377295

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Predicting phishing websites based on self-structuring neural network. *Neural Computing and Applications*, *25*(2), 443–458. https://doi.org/10.1007/s00521-013-1490-z

Mousavi, S., & Bahaghighat, M. (2025). Phishing Website Detection: An In-Depth Investigation of Feature Selection and Deep Learning. *Expert Systems*, *42*(3), e13824. https://doi.org/https://doi.org/10.1111/exsy.13824

Nguyen, V., Wu, T., Yuan, X., Grobler, M., Nepal, S., & Rudolph, C. (2024). *An Innovative Information Theory-based Approach to Tackle and Enhance The Transparency in Phishing Detection*. https://arxiv.org/abs/2402.17092

Nisreen Innab Ahmed Abdelgader Fadol Osman, M. A. M. A. M. A.-Z. B. M. E. F. H. Z. M. F. A. (2024). Phishing Attacks Detection Using Ensemble Machine Learning Algorithms. *Computers, Materials & Continua*, *80*(1), 1325–1345. https://doi.org/10.32604/cmc.2024.051778

Omari, K. (2023). Comparative Study of Machine Learning Algorithms for Phishing Website Detection. *International Journal of Advanced Computer Science and Applications*, *14*(9). https://doi.org/10.14569/IJACSA.2023.0140945

Orunsolu, A. A., Sodiya, A. S., & Akinwale, A. T. (2022). A predictive model for phishing detection. *Journal of King Saud University - Computer and Information Sciences*, *34*(2), 232–247. https://doi.org/https://doi.org/10.1016/j.jksuci.2019.12.005

Petrosyan, A. (2024). *Targets of External Attacks Global 2023*. Statista. https://www.statista.com/statistics/1451097/targets-of-external-attacks-worldwide/

Popescul, D., & Radu, L. D. (2025). AI in phishing detection: a bibliometric review. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1496580

Rahebi, M. G., & Rahebi, J. (2025). An explainable hybrid deep learning-optimization framework for robust phishing attack detection using GAN and transformer-based feature learning. *Ain Shams Engineering Journal*, *16*(12), 103745. https://doi.org/10.1016/j.asej.2025.103745

Rao, R. S., & Pais, A. R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, *31*(8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0

Saeed M. Alshahrani Nayyar Ahmed Khan, J. A. W. A. S. (2022). URL Phishing Detection Using Particle Swarm Optimization and Data Mining. *Computers, Materials & Continua*, *73*(3), 5625–5640. https://doi.org/10.32604/cmc.2022.030982

Safi, A., & Singh, S. (2023). A systematic literature review on phishing website detection techniques. *Journal of King Saud University - Computer and Information Sciences*, *35*(2), 590–611. https://doi.org/https://doi.org/10.1016/j.jksuci.2023.01.004

Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345–357. https://doi.org/https://doi.org/10.1016/j.eswa.2018.09.029

Somesha, M., Pais, A. R., Rao, R. S., & Rathour, V. S. (2020). Efficient deep learning techniques for the detection of phishing websites. *Sādhanā*, *45*(1), 165. https://doi.org/10.1007/s12046-020-01392-4

Uddin, K. M. M., Biswas, N., Rikta, S. T., Nur -A-Alam, Md., & Mostafiz, R. (2025). Explainable Machine Learning for Phishing Site Detection: A High-Efficiency Approach Using Boosting Models and SHAP. *The Journal of Engineering*, *2025*(1), e70110. https://doi.org/https://doi.org/10.1049/tje2.70110

Vrbančič, G., Fister, I., & Podgorelec, V. (2020). Datasets for phishing websites detection. *Data in Brief*, *33*, 106438. https://doi.org/https://doi.org/10.1016/j.dib.2020.106438

Wang, W., Zhang, F., Luo, X., & Zhang, S. (2019). PDRCNN: Precise Phishing Detection with Recurrent Convolutional Neural Networks. *Security and Communication Networks*, *2019*(1), 2595794. https://doi.org/https://doi.org/10.1155/2019/2595794

Wilk-Jakubowski, J. L., Pawlik, L., Wilk-Jakubowski, G., & Sikora, A. (2025). Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024). *Electronics*, *14*(18). https://doi.org/10.3390/electronics14183744

Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Trans. Inf. Syst. Secur.*, *14*(2). https://doi.org/10.1145/2019599.2019606

Yang, P., Zhao, G., & Zeng, P. (2019). Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. *IEEE Access*, *7*, 15196–15209. https://doi.org/10.1109/ACCESS.2019.2892066

Yi, P., Guan, Y., Zou, F., Yao, Y., Wang, W., & Zhu, T. (2018). Web Phishing Detection Using a Deep Learning Framework. *Wireless Communications and Mobile Computing*, *2018*(1), 4678746. https://doi.org/https://doi.org/10.1155/2018/4678746

Zhang, Y., Hong, J. I., & Cranor, L. F. (2007). CANTINA: A Content-Based Approach to Detecting Phishing Web Sites. *Proceedings of the 16th International Conference on World Wide Web, WWW '07*, 639–648. https://doi.org/10.1145/1242572.1242659