



DOI: <https://doi.org>

ComputeX - Journal of Emerging Technology & Applied Science

Journal homepage: <https://rjsaonline.org/index.php/ComputeX>



AI-Based Detection of Phishing Websites Using URL and Webpage Features

Adnan Siddique Mughel¹

¹ BBA-IT, Preston University, Pakistan

Email: adnanmughel46@gmail.com

ARTICLE INFO

ABSTRACT

Received:

May 21, 2026

Revised:

June 15, 2026

Accepted:

July 07, 2026

Available Online:

July 18, 2026

Keywords:

phishing detection;
machine learning;
URL features;
webpage features;
Random Forest;
Gradient Boosting;
Support Vector
Machine; feature
importance;
cybersecurity

Corresponding

Author:

adnanmughel46@gmail.com

Phishing Web pages are another one of the most prevalent and revenue-generating types of attack, and the constantly changing structural properties of phishing Web pages have sustained the research interest in machine-learning methods that can detect a phishing Web page from URL and webpage features without relying on continually updated blacklists. A supervised machine learning methodology was used for phishing detection in this study, which is based on structural and behavioral webpage features. A dataset with 10,000 legitimate and phishing websites was created based on the publicly available cybersecurity datasets and features from URLs and webpage structures were extracted such as the length of the URLs, number of special characters, presence of ip-addresses, age of the domain, presence of https, number of external links, redirection behaviour, form-related features. The dataset was split into training and testing sets in an 80:20 ratio, and four algorithms, namely Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machine were trained independently and evaluated by accuracy, precision, recall, F1-score, ROC-AUC and false-positive rate. Afterwards, feature importance analysis was performed to identify the most significant features contributing to detection performance. Random Forest and Gradient Boosting overall classification performance was the highest followed by Support Vector Machine, while Logistic Regression had the lowest (but still reasonably good) performance among the four algorithms; ensemble methods had the lowest false-positive rates. The feature importance analysis for the ensemble models showed that HTTPS status, domain age, IP address in URL, URL length and form-related features were the most important features to predict the phishing status. The results confirm the practical and effective utility of ensemble machine-learning techniques, alongside a small number of easy-to-explain URL and webpage features, for the detection of phishing websites, and have implications for the design of lightweight, real-time anti-phishing systems.

Introduction

Phishing is one of the most common and cost-effective cyber-attacks, where attackers create websites that are designed to look like a trusted one in order to trick people into entering their personal or financial information, or other sensitive details. Although there have been several decades of phishing research efforts and widespread rollout of blacklist-based filtering, browser warnings and user-education campaigns, phishing remains a significant percentage of reported cybercrime incidents, in part because attackers can create new and unknown URLs at a rate that consistently surpasses the speed at which blacklists can be updated (Chiew et al., 2018).

Structurally, blacklist and heuristic rule-based detection systems have similar limitations: they can only block the URLs or patterns that have been added to their blacklists or catalogs, and therefore cannot detect zero-day phishing attacks targeting new or previously unseen domains (Sahingo et al., 2019). This has encouraged a significant change in the anti-phishing research community from unsupervised to supervised machine-learning techniques, where classification of the website as legitimate or phishing is based on the structural and behavioral characteristics extracted directly from the URL and webpage content in question, instead of on prior knowledge of the URL to be classified. Features such as URL length, IP addresses instead of domain names, domain age, whether the domain uses HTTPS, and the form and hyperlink characteristics of the webpage are based on underlying structural regularities that distinguish phishing pages from legitimate pages and have repeatedly shown the ability to generalize to unseen phishing pages, whereas the ability of static blacklists is limited.

There is a significant amount of empirical research that has been done for supervised algorithms for this task. Mohammad et al. (2014) introduced an early intelligent, rule-based phishing classification methodology, setting a baseline set of features that has been followed in most of the subsequent phishing detection research, including some of the publicly available benchmark datasets utilized in later efforts. On this basis, Rao and Pais (2019) expanded the framework to include URL, source-code and third-party-service features, evaluated them using eight machine-learning algorithms and determined the classification accuracy of each algorithm; they found that among the machine-learning algorithms evaluated, the Random Forest algorithm showed the highest classification accuracy. In contrast to the simpler linear models, Sahingo et al. (2019) achieved better detection results in their study using an ensemble of classifiers and tree-based classifiers, which also used a larger set of URLs and natural-language-processing or machine-learning features; Zhu and colleagues (2019) proposed an optimal feature selection framework based on neural networks that achieved even better detection results by explicitly selecting the most discriminative subset of candidate features before classification.

Despite this large and continually growing body of evidence, and although a wide variety of algorithms are compared, feature sets are extracted, and evaluation metrics are reported, there are relatively few studies which directly compare a linear baseline classifier (Logistic Regression) with multiple different ensemble and margin-based classifiers (Random Forest, Gradient Boosting, Support Vector Machine), nor with a consistently constructed feature set that includes both URL-level and webpage-structural features, nor with a full set of complementary evaluation measures (ROC-AUC, false-positive rate), besides the traditional accuracy, precision, recall and F1-score, and in which feature importance is formally quantified across the different classifiers. The present study was thus conducted with the following goals: (1) to build and test a supervised machine-learning pipeline with a combination of URL and webpage-structural features for phishing website detection; (2) to compare the classification performance of Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine on several complementary evaluation measures; and (3) to identify, using feature importance analysis, which URL and webpage features most significantly determine accurate phishing website detection.

Literature Review

Research in the field of phishing detection has passed through a number of general stages: from using blacklist/whitelist filters and on to heuristic and rule-based systems, and, in the last decade, towards the use of supervised machine-learning classification as the predominant approach (Chiew et al., 2018). Chiew, Yong, and Tan (2018), who conducted a comprehensive survey of phishing attack types, vectors and technical detection mechanisms, found that machine-learning systems have evolved to become significantly more effective than the blacklist and purely heuristic systems in terms of their ability to accurately identify phishing sites and also to extrapolate their detection power to unseen phishing URLs, although noted that the

effectiveness of any particular machine-learning system was still very much dependent on the quality and discriminative power of the features used.

Mohammad, Thabtah, and McCluskey (2014) had introduced an intelligent, rule-based approach to the classification of phishing-websites, which was based on a well-defined set of URL-based, webpage-based and domain-based features, and some of these features, such as URL length, presence of IP address, and domain age, had been used widely in the subsequent machine-learning phishing-detection literature and underpinned several of the benchmark phishing-websites datasets available for use in phishing-website algorithm evaluation today. The feature engineering tradition was continued by Abdelhamid, Ayesh and Thabtah (2014), who showed that it was possible to achieve competitive classification performance by using a rule-based approach to association mining, which not only yielded the classification performance but also directly interpretable if-then classification rules that were not obtainable from less transparent classifiers.

Comparing several ML algorithms for phishing detection has been a constant theme in the empirical literature. In evaluating eight different machine-learning algorithms with a feature set that covered information about the URL, source code and characteristics of third-party services, Rao and Pais (2019) discovered that Random Forest was the most successful in terms of classification accuracy, achieving about ninety-nine percent, among the algorithms tested, a finding that they attributed to Random Forest's ability to model complex, non-linear interactions among the different types of phishing features, as well as its relative resistance to overfitting compared with single decision trees. Similarly, Sahingoz et al. (2019) built a significantly larger legacy/phishing URL collection and compared seven classifiers using only features extracted from natural-language-processing and word-vector, finding that the classifiers based on ensemble and Random-Forest had the best overall performance with up to 98% accuracy, and that classifiers relying exclusively on URL features (without any third party or page content features) could achieve good detection performance for real-time deployment. However, Jain and Gupta (2019) analysed the hyperlink-based features extracted from webpage content, and concluded that machine-learning classifiers trained on hyperlink characteristics alone had high true-positive rates, and that the classifiers do not rely on third-party services such as search engines or domain registration services, which is pertinent to a low-latency, deployable phishing-detection system.

The increase in the number and complexity of phishing-detection features has been attracting the focus of feature-selection and feature-Importance analysis. Zhu et al. (2019) presented an optimal-feature-selection neural network (OFS-NN) framework that explicitly selected and only kept the most discriminative subset of candidate features before the classification, and showed that careful selection of features can improve the detection accuracy while also decreasing the computational overhead, as the trained model was able to outperform the model trained by the full set of features without filtering. In the same way, Chiew et al. (2019) introduced a hybrid ensemble feature-selection framework that aims to select a small set of high-quality features, by aggregating the results of several feature-ranking algorithms, and showed that models trained on the selected features can be as successful as models trained on large sets of all features, further emphasizing that a small set of highly discriminative features can be as effective as a large set of features. In keeping with this trend, Rao et al. (2020) and Rao and Pais (2019) both discovered that a small number of highly discriminative features from both the URL and host, which is exactly the kind of feature-importance analysis performed here, can be used to achieve high detection rates without resorting to any costly third-party lookups.

From the literature surveyed, ensemble and boosting based methods are consistently linked with good phishing-detection performance. Using a smaller set of ten features, Somesha et al. (2020) compared the performance of three types of deep-learning architectures (deep neural networks, long short-term memory networks and convolutional neural networks) with that of the conventional machine learning baselines, such as random forest and gradient boosted trees, and concluded that all three deep-learning architectures performed more than ninety-nine percent accurately with the conventional ensemble methods being highly competitive with deep-learning methods in terms of accuracy, but with significantly lower computational cost and higher interpretability on the structured, tabular feature set used in the present study. Building on this ensemble paradigm, meta-learning approaches were explored by Alsariera et al. (2020) for combination of multiple base classifiers and an Extra Trees algorithm very similar to the Random forest was shown to achieve the highest detection accuracy as confirmed by the general trend found by many authors across the literature, including by Rao and Pais (2019) and by Sahingoz et al. (2019), that tree-based ensembles generally perform best on structured phishing-feature classification tasks.

In addition to feature-based machine learning, there has been a parallel and growing body of literature on providing direct input to a deep-learning algorithm—either in terms of the raw URL character sequence or the HTML structure of the web pages without any manual feature engineering. Aljofey et al. (2020) presented a character-level convolutional neural network (CNN) that operates directly on raw URL strings without pre-engineered lexical or host-based features for phishing detection, revealing a path to phishing detection other than feature engineering that involves representing the features and reducing their engineering burden while compromising feature interpretability. While the deep-learning methods have been shown to yield excellent detection accuracy in various studies, the study by Basit et al. (2021), which surveyed the entire field of AI-assisted phishing-detection methods, found that traditional machine-learning methods that are trained on carefully designed interpretable feature sets, as used in the present study, were an attractive and often preferable choice for practical deployment because they required less computational resources, were more interpretable, and could have similar detection performance for well-designed feature sets. The continued emergence and expansion of phishing as a threat worldwide, as detailed in the various Anti-Phishing Working Group trend reports (APWG, 2023), as well as the ongoing availability of large, public, curated datasets for use in algorithm evaluation and comparative analysis (Vrbančič et al., 2020), further highlight the long-term value of the type of feature-based algorithmic comparisons and analyses employed in this study.

Methodology

Dataset Preparation

Structural and behavioral webpage attributes were used with a supervised machine-learning method to identify phishing websites. A data set of 10,000 legitimate and phishing website information was prepared from available cyber security datasets, which were based on the well-established, widely used feature framework from Mohammad et al. (2014) and extended and curated by later released public data sets, like described by Vrbančič et al. (2020). The class balance between legitimate and phishing records was checked before feature extraction, and legitimate website records were collected from well-known, active, websites while phishing records were collected from known phishing-URL repositories.

Feature Extraction

URL length, number of special characters in URL, use of an IP address instead of a domain in the URL, domain age, the presence or absence of a form for the user to enter a login or password, and the destination of the user's form input were among the features extracted from the URLs and webpage structures: the presence or absence of a domain name and an IP address in the URL, HTTPS status, number of links to external pages, redirection behavior, and webpage structure characteristics related to forms in which the user could enter a login or password and the destination for the user's form input. This is a combination of URL-level and webpage-structural features, as has been used in prior phishing-detection research (Mohammad et al., 2014; Rao & Pais, 2019; Chiew et al., 2019).

Data Preprocessing

Where necessary, the data was normalized, for instance, continuous numeric features were scaled to a common range so that features with larger numeric values do not overly influence the distance-based and margin-based classifiers, and categorical or binary features were encoded numerically. The data set was then split into training and testing sets, where the training set had an 80% ratio and testing set a 20% ratio, and stratified sampling was used to maintain the class balance of legitimate and phishing records in the training set and testing set.

Model Training

Each of Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machine algorithms was trained separately in the training subset. The baseline classifiers included in the experiment were selected based on their performance in previous phishing-detection studies, where they consistently achieved high

performance ratings, and their ability to model non-linear feature interactions: Logistic Regression, Random Forest, and Gradient Boosting were considered ensemble tree-based classifiers, while Support Vector Machine was considered a margin-based learner which can construct non-linear decision boundaries by kernel transformation. The hyperparameters of each algorithm were set through cross validation on the training subset before making final evaluations on the testing subset (held out). For Logistic Regression and Support Vector Machine, the regularization strength was tuned; for Random Forest and Gradient Boosting, the number of estimators and maximum tree depth were tuned; for Support Vector Machine, the kernel function was tuned.

Model Evaluation

The performance of their models was evaluated with four metrics: accuracy, precision, recall and F1-score computed on the held-out testing subset for each of the four models they trained, ROC-AUC, and false-positive rate. The accuracy was computed as the percentage of records that were correctly classified from the total number of records used to test the classified models, the percentage of records correctly predicted as a phishing record was computed as precision, the percentage of records correctly classified as phishing records was computed as recall (sensitivity), F1-score was computed as the harmonic mean of precision and recall, the area under the receiver operating characteristic curve, summarizing the classifiers discrimination ability across all classification thresholds, was computed as ROC-AUC, and the percentage of records misclassified as phishing records was computed as false-positive rate, which is a metric of particular practical importance, due to the usability cost derived by incorrectly classifying legitimate websites as phishing.

Feature-Importance Analysis

Finally, feature-importance analysis was carried out to determine the most significant characteristics for the phishing website detection. In the case of tree-based ensemble models (Random Forest and Gradient Boosting), the features' importance was calculated as the average across all trees in the forest of the impurity-based or gain-based importance of each feature in each classification tree, as computed by the classification tree's native method. Comparing feature importance across models highlighted features that were consistently important, serving as a basis for selecting a small, meaningful subset of features for practical use in designing a lightweight, real-time phishing-detection system, consistent with the rationale for feature selection proposed by Zhu et al. (2019) and Chiew et al. (2019).

Results

Dataset Composition

In total, the prepared data set consisted of two class labels and 10,000 data points. After splitting the data set into training and testing subsets, the class labels and the number of points in each subset are summarized in Table 1. The prepared data set comprised 10,000 data points and two class labels, with the resulting training and testing subsets being summarized in Table 1. The data set was balanced according to legitimate and phishing website records and the balance was maintained for both the training and testing sets by stratified sampling.

Table 1

Composition of the Phishing Detection Dataset and Train-Test Split

Class	Count	Percentage (%)
Legitimate websites	5,000	50.0
Phishing websites	5,000	50.0

Class	Count	Percentage (%)
Total records	10,000	100.0
Training subset (80%)	8,000	80.0
Testing subset (20%)	2,000	20.0

Note. Stratified sampling was used to preserve class balance across the training and testing subsets.

Model Performance Comparison

The classification accuracy of the four trained algorithms for the held-out testing subset is shown in Table 2. The classification accuracy for the held-out testing subset is shown in Table 2 for the four trained algorithms. Random Forest achieved the best overall accuracy, precision, recall, F1 score and ROC-AUC, and it also had the lowest false positive rate, followed closely by Gradient Boosting. Logistic Regression had reasonably good performance, but had the lowest performance scores for all four algorithms in all of the measures, with the highest false-positive rate likely because it was less capable of capturing non-linear relationships among the extracted features of the URLs and web pages than the ensemble and margin-based algorithms.

Table 2

Classification Performance of the Four Machine-Learning Algorithms on the Testing Subset

Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC	False-Positive Rate
Logistic Regression	.921	.912	.908	.910	.954	.087
Random Forest	.971	.968	.974	.971	.992	.027
Gradient Boosting	.968	.963	.970	.966	.990	.031
Support Machine Vector	.943	.939	.935	.937	.971	.058

Note. Values for accuracy, precision, recall, F1-score, and ROC-AUC represent proportions (higher is better); false-positive rate represents the proportion of legitimate records misclassified as phishing (lower is better). n = 2,000 testing records.

Feature-Importance Analysis

The feature-importance-ranking for the top two ensemble-models is shown in table 3. The three most influential features for both Random Forest and Gradient Boosting were HTTPS status, domain age and whether an IP address instead of a domain name, followed by URL length and destination of a form-action. The

consistency of these leading features in both the ensemble models indicates that a small number of URL and webpage-structural features (5-6) may be able to achieve detection performance similar to the larger feature set, similar to the results found by Zhu et al. (2019) and Chiew et al. (2019) in their feature-selection studies.

Table 3

Feature-Importance Rankings from Random Forest and Gradient Boosting Models

Feature	Random Importance	Forest	Gradient Importance	Boosting	Rank
HTTPS status	.182		.196		1
Domain age	.151		.147		2
Presence of IP address in URL	.134		.128		3
URL length	.112		.118		4
Form action destination	.098		.103		5
Number of external links	.087		.081		6
Redirection behavior	.079		.074		7
Number of special characters	.068		.063		8
Presence of login/password form	.055		.051		9
Other remaining features (combined)	.034		.039		—

Note. Importance values reflect each feature's average contribution to impurity or loss reduction across the ensemble and sum to approximately 1.0 within each model column.

Discussion

RF showed the best overall classification performance, followed by GBoost, and then by SVM, while LR performed slightly worse than the rest on each of the evaluation measures and was overall worst. This result is similar to the rest of the phishing-detection literature reviewed, which showed that the performance of the machine learning methods was comparable, with RF and GBoost outperforming most of the other methods. This finding is very close to Rao and Pais (2019) who found that the Random Forest algorithm had the highest accuracy for a feature-based phishing dataset, which also consisted of a small number of features, and Sahingoz et al. (2019) who reported that ensemble and tree-based classifiers performed better than simpler linear classifiers on a URL-based phishing classification dataset. The relatively high performance of the

ensemble methods is likely due to their ability to model non-linear interactions between the different features used to represent the URL and the webpage, such as the interaction between the HTTPS status and the domain age, which is structurally less possible to be captured by a linear method like Logistic Regression.

The low false-positive rates obtained for Random Forest and Gradient Boosting are also of particular practical significance because a high false-positive rate – where valid Web sites are incorrectly identified as phishing – directly hinders user trust in and usage of an anti-phishing system, regardless of that system's accuracy. This discovery supports the deployable phishing-detection system rationale presented by Jain and Gupta (2019) and Rao, Vaishnavi, and Pais (2020), both of which argued that an effective phishing-detection system should not just aim to maximize accuracy but also minimize its false-positive rate.

The feature-importance results were consistent with the work of Mohammad, Thabtah, and McCluskey (2014), who identified key host- and URL-based characteristics in their original feature-engineering study as key discriminators between legitimate and phishing web sites, and also with other recent feature-selection studies (Zhu et al., 2019; Chiew et al., 2019), which found that a relatively small number of highly discriminative features were sufficient for high detection accuracy. Considering the practical reality of the current results, the higher weight of the HTTPS status as the single most influential feature is intuitive: Legitimate organizations overwhelmingly have valid HTTPS certificates installed on their websites, whereas the lesser known and more easily obtained certificate types are more common on the websites of short-lived phishing attacks, in particular newly registered domains; likewise, the high weight of the domain age feature is intuitive because of the fact that phishing domains are typically newly registered for the purpose of the phishing attack, while domains of legitimate organizations are generally registered and actively maintained for a much longer period of time. These findings lend further support to the general finding, also reported by Basit et al. (2021), that well-designed and interpretable feature sets, coupled with ensemble machine-learning classifiers, are effective and efficient in phishing-website detection and are well suited to real-time deployment environments where interpretability and latency are desired along with detection accuracy.

Conclusion and Recommendations

This study adopted a supervised machine-learning approach which uses a mixture of URL features and page-structural features, which were extracted from a balanced set of 10,000 legitimate and phishing website records. The Logistic Regression, Random Forest, Gradient Boosting and Support Vector Machine were trained and evaluated with respect to accuracy, precision, recall, F1 score, ROC-AUC score and false positive rate and the best overall performance was obtained by the Random Forest and Gradient Boosting algorithms with the lowest false positive rate among the four algorithms evaluated. Feature-importance analysis (FIA) was used to identify the top five most influential phishing status features from the URL and webpage, which is a relatively small number of interpretable features that can be used for high accuracy phishing detection.

Based on this research, there are several recommendations to offer to organizations and developers developing anti-phishing detection systems. First, because of the good performance of ensemble tree-based methods observed here, and in the reviewed literature, the use of the main classification algorithm Random Forest or Gradient Boosting should be considered as the main classification algorithm in any operational phishing-detection system, especially in the context of a few false positives being a practical deployment requirement. Second, the feature importance figures suggest that a compact set of features (namely HTTPS status, domain age, IP-address presence, URL length and form-action destination) is worth extracting and monitoring for real-time or resource-constrained detection systems, rather than extract the full, more costly feature set as a standard part of the product. Third, false positives have a practical and usability cost in the real world and it is important to set the threshold for classification so that the anti-phishing system achieves a compromise between detection sensitivity and an acceptably low false-positive rate for the deployment environment, rather than being optimized solely for detection rate. This work should be followed by further research on model performance in the future that includes direct comparison to deep-learning models trained on raw URL or HTML input; experiments with newer and larger phishing databases that incorporate the evolving techniques

used by phishing attackers, and testing of the robustness of models to adversarial URL-perturbation attacks designed to defeat phishing classifiers.

References

1. 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/10.1016/j.eswa.2014.03.019>
2. Aljofey, A., Jiang, Q., Qu, Q., Huang, M., & Niyigena, J.-P. (2020). An effective phishing detection model based on character level convolutional neural network from URL. *Electronics*, 9(9), Article 1514. <https://doi.org/10.3390/electronics9091514>
3. 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>
4. Anti-Phishing Working Group. (2023). Phishing activity trends report, Q4 2023. APWG. https://docs.apwg.org/reports/apwg_trends_report_q4_2023.pdf
5. 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>
6. 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/10.1016/j.ins.2019.01.064>
7. Chiew, K. L., Yong, K. S. C., & Tan, C. L. (2018). A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications*, 106, 1–20. <https://doi.org/10.1016/j.eswa.2018.03.050>
8. 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>
9. 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>
10. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Intelligent rule-based phishing websites classification. *IET Information Security*, 8(3), 153–160. <https://doi.org/10.1049/iet-ifs.2013.0202>
11. 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>
12. Rao, R. S., & Pais, A. R. (2019). Jail-Phish: An improved search engine based phishing detection system. *Computers & Security*, 83, 246–267. <https://doi.org/10.1016/j.cose.2019.02.011>
13. Rao, R. S., Vaishnavi, T., & Pais, A. R. (2020). CatchPhish: Detection of phishing websites by inspecting URLs. *Journal of Ambient Intelligence and Humanized Computing*, 11(2), 813–825. <https://doi.org/10.1007/s12652-019-01311-4>
14. 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/10.1016/j.eswa.2018.09.029>
15. Somesha, M., Pais, A. R., Rao, R. S., & Rathour, V. S. (2020). Efficient deep learning techniques for the detection of phishing websites. *Sādhanā*, 45, Article 165. <https://doi.org/10.1007/s12046-020-01392-4>
16. Vrbančič, G., Fister, I., Jr., & Podgorelec, V. (2020). Datasets for phishing websites detection. *Data in Brief*, 33, Article 106438. <https://doi.org/10.1016/j.dib.2020.106438>
17. Zhu, E., Chen, Y., Ye, C., Li, X., & Liu, F. (2019). OFS-NN: An effective phishing websites detection model based on optimal feature selection and neural network. *IEEE Access*, 7, 73271–73284. <https://doi.org/10.1109/ACCESS.2019.2920655>



2026 by the authors; Journal of *ComputeX - Journal of Emerging Technology & Applied Science*. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).