International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 1686-1702
https://doi.org/10.70917/ijcisim-2026-5919
Article

An Improved Machine Learning Framework for Phishing URL Detection Using the PhiUSIIL Dataset

Warveen merza eido[1], Omar S. Kareem[2]

[1]Akre University for Applied Sciences, Technical College of Informatics, Department of Information Technology, Duhok, Iraq, warveen.aydo@dpu.edu.krd
[2]Department of Public Health, College of Health and Medical Technology, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq; omar.kareem@dpu.edu.krd

Abstract: . One of the biggest threats to cybersecurity continues to be phishing attacks that involve malicious URLs that attempt to obtain personal information from users. This paper proposes a new method of detecting phishing URLs using the PhiUSIIL dataset, which contains a large number of features that allow for the analysis of the effectiveness of multiple classification algorithms. The new detection methodology includes a preprocessing of the data, feature selection for the classification model via an ANOVA based methodology, and four different supervised classification algorithms: Random Forest, XGBoost, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). In addition, this paper presents multiple performance metrics (i.e., accuracy, precision, recall, and F1-Score) that were used to evaluate the performance of the phishing URL detection models. According to the experiments, all the ensemble models did an outstanding job, with Random Forest and XGBoost having the highest accuracy of 100%, and all other models having more than 99.9% for each evaluation metric compared, indicating that this new framework will model the complexity of features in phishing domains successfully and provide fast computational times. Furthermore, this research further confirms the advantages of using ML and ensemble methods as scalable, dependable, and accurate solutions for detecting phishing URLs in real-world applications...
Keywords: Phishing URL Detection; Machine Learning; Feature Selection; PhiUSIIL Dataset; Cybersecurity

1. Introduction

Phishing is a form of a cyber threat that continues to be a liability and a challenge for businesses today to secure. Phishing attacks are designed to trick individuals into providing sensitive information such as usernames/passwords and bank account details usng fraudulent websites and/or links to those websites to be redirected to a third-party site. Phishing has increased in scale and frequency across every type of industry due the increase of the number of online services available today.

Phishing means that there are fake websites, bad URLs, or other methods that cybercriminals will use to trick you out of your sensitive information, such as your login credentials to use (obtain access to), financial information, or any kinds of identifiers. Phishing attacks are among the most common and dangerous types of cyberattacks today. With all of the innovations in Internet services and technologies, including cloud computing, digital transformation, etc., the number of potential attacks (the attack surface) available to cybercriminals has increased exponentially (and continues to grow). As such, the number of types of attacks has continued to increase, putting people, businesses, and governments around the world at risk of phishing attacks.

The traditional means of detecting phishing—targeting heuristics and using blacklists—are not very good at identifying previously unfamiliar phishers and new forms and variations of phishing (4). These approaches largely require pre-determined rules and documented types of attacks to be effective. Because the methods used by phishers


Copyright: © 2026 by the authors

change constantly and frequently, this makes it difficult for existing detection methods to keep pace with the sheer number of phishing methods that are developing (5)(6). Due to this fact, people are interested in improving detection methods that provide for advanced, adaptive systems capable of accurately detecting more advanced and complex forms of phishing (7) in real-time.

Recently, there has been a surge of interest in both Machine Learning (ML) and Deep Learning (DL) applications to solve these problem types because they can help identify complex patterns within large data sets (9). Multiple independent research efforts have demonstrated the effectiveness of using ML to detect instances of phishing online; particularly successful at this are ensemble ML algorithms, such as Random Forest and Gradient Boosting, both of which can achieve very high rates of correct detections (10). A variety of other kinds of Deep Learning models (including Convolutional Neural Networks [CNNs] and hybrid networks) have also shown strong possibilities for modeling non-linear relationships between the data items associated with features of incidents related to phishing and other types of online fraud (10; 11).

Moreover, hybrid intelligent systems, augmented with optimized learning algorithms, have been investigated to enhance detection performance while adapting to variable conditions (12). Although deep learning methods provide many advantages, their use is limited by the requirement for high levels of computational resources and complex architectures, thereby constraining the application of these types of solutions to real-time, resource-limited environments (13). Another factor that is important in enhancing the performance of phishing detection systems is the use of high-quality data sets. The PhiUSIIL dataset provides one of the most comprehensive datasets available in the area of phishing detection and includes features based on the structure, vocabulary and use patterns of both phishing and legitimate URLs (14). The PhiUSIIL dataset allows for the comparison of models against real-world datasets and also provides an opportunity to deploy reliable machine-learning-based solutions.

Despite these technological advancements, there are still many challenges that we face today and must resolve, including: 1) lack of generalizability in models, 2) dependence on features across multiple datasets, and 3) computational efficiency (15). In addition, the continued pursuit of developing accurate, efficient, and adaptable detection systems is ongoing in all research areas as the primary focus of future research (16).

One of the key contributions of this research paper is an improved framework for detecting phishing URLs based on the PhiUSIIL dataset. This framework incorporates the use of ANOVA for Feature Selection, optimized preprocessing, and a composite model; resulting in increased detection accuracy, robustness, and computing efficiency. The research paper focuses on three specific contributions:.*:*

1. This paper presents a more accurate classification of phishing URLs by implementing a new machine learning framework based on the PhiUSIIL dataset that improves both classification accuracy and computational efficiency.
2. The data collected is preprocessed thoroughly with preprocessing and feature selection techniques that improve the quality of the data collected and help to remove redundant data, resulting in stronger models and increased generalization capability for the final model.
3. A comparative evaluation of multiple supervised learning models, including Logistic Regression, Random Forest, and XGBoost, is conducted to identify the most effective classifier.
4. The suggested framework has achieved superior performance, including 100% correct classification and more than 99.9% for additional metrics, resulting in an efficient and scalable way to implement solutions in real life..

2. Related Work

The PhiUSIIL model, which was introduced by (18), is known to be a major player in phishing URL detection due to its extensive set of features. They concentrated on the production of an incredibly comprehensive dataset of features as well as the incorporation of a similarity index into incremental learning methods. Their solution has been

shown to effectively identify many different types of phishing attacks (e.g., homograph-based and typo-squatting attacks). The results from the experiments demonstrate very good performance concerning detection, with almost 99.79% accuracy.

(19) examined the PhiUSIIL as deep learning based phishing detection system. Multiple models were tested in the study, and the best-performing model was the FCNN. The model performed very well on capturing complex relationships between URL based features. The results showed higher accuracies and precisions compared to other traditional machine learning methods. The highest accuracy recorded was 99.38% with a precision of 99.12%, a recall of 99.68% and an F1 Score of 99.35%.

(20) proposed a method for adaptive phishing detection based on reinforcement learning, using the PhiUSIIL dataset. The model incorporated human-in-the-loop feedback to increase both efficiency and adaptability of learning. The proposed method provided greater generalization capabilities across evolving phishing patterns than existing methods. The experimental results indicated an F1-score of 97.62% demonstrating outstanding detection capabilities.

In (21), a comparative assessment of numerous machine learning methods utilized in classifying spam and phishing URLs, were made according to their accuracy, with various applied feature engineering methods based on the PhiUSIIL dataset. The evaluated classifiers included Random Forest, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The study concluded with a significant level of high performance from both the Random Forest algorithm and the ANN among all models evaluated in this study. Additionally, the accuracy recorded in this study reached upwards of 99.99% for phishing detection efforts.

In (22), there was a new method for detecting phishing sites that uses different types of deep learning (e.g., URL/Website-based Deep Learning, Meta Data/Metadata-based Deep Learning, Visual/Video-based Deep Learning) as input sources and combines those inputs using an attention mechanism to increase the detection accuracy and interpretability of the results.. This method is designed to be lightweight so that it can be used in environments with limited resources, and follows a rigorous testing protocol, with a 98.4% accuracy rate on various large, well-known public data sets (PhishTank, OpenPhish, Tranco, UCI, and Phish-IRIS).

The ability of machine learning methods to identify phishing URLs has recently been demonstrated using various datasets.(23) Utiilized the PhiUSIIL dataset and applied ANOVA-based feature selection, achieving 99.98% accuracy. Similarly,(24) Utilized stacking-based ensemble learning on this same dataset, achieving 99.99% accuracy, validating the utility of ensemble methods. The work, however, also (25) highlighted a limitation with respect to how generalizable (i.e. applicable to other datasets) the models were, Their ability to perform at a high level on one dataset does not necessarily mean that they will have similar performance in the real world.

By using previous research as a basis, this study aims at making use of various model types and multiple features through the application of the PhiUSIIL dataset to reach a high level of classification accuracy; however, future assessments on external datasets will still be required..

3. Materials and Methods

This section of the document outlines the method and materials used for this research and the development of a new Machine Learning System for Phishing URL detection using the PhiUSIIL Dataset. The methods of machine learning utilized in this study will produce a framework with high-quality classification capability, computational efficiency, and robust model. The study will consist of six phases: (1) the dataset has been identified (2) data was pre-processed (3) relevant features have been selected for inclusion into the machine learning algorithm (4) a model was developed based upon the selected feature set (5) the models were evaluated. The PhiUSIIL dataset has been used as a baseline dataset due to its size and diversity of features. Once the data was collected, data preprocessing techniques were applied to clean the data, prepare the dataset for training, and identify redundant features using feature selection methods to improve the models' ability to generalize to new data. As part of the results of different classifiers (Logistic Regression, Random Forest, XGBoost), classifiers were developed using a number of different supervised machine-

learning algorithms. Finally, the proposed framework was evaluated using standard evaluation measures for performance.

This research design reflects the evolution of the academic literature, which promotes research activities that integrate the preprocessing of data, feature engineering and optimization of models to realize highly effective phishing detection systems. One example of this is the work of Ali et al. (4), who developed an integrated, real-time phishing detection system based on the use of multiple classifiers from machine learning with features derived from URLs, providing both high levels of classification accuracy as well as being applicable in practical situations (i.e. integrating into web browsers). A second illustration is the work of Sabri et al. (12), who created a deep learning-based phishing detection system (ADUIN) where optimized URL-feature engineering is used in conjunction with adaptive neural architectures to achieve greater detection accuracy and generalizability (ADUIN)..

3.1.Dataset

This study is based on a dataset from the PhiUSIIL Phishing URL Dataset that contains many features relevant to distinguishing between legitimate URLs and phishing URLs. The dataset contains over 235,000 URLs, which provides a large enough sample to train and evaluate ML models. The PhiUSIIL dataset contains multiple features from structural, lexical, and behavioral aspects, so ML models should be able to effectively differentiate between legitimate URLs and phishing URLs. As part of this research, a reduced feature subset will be used to reduce the dimensionality of the data and improve computational efficiency while still providing a good degree of classification accuracy. The dataset will be divided into an 80/20 ratio for the training/testing dataset to facilitate fair and accurate assessment. A training set with 188,636 samples has been used, as well as a test set with 47,159 samples. Both sets are evenly divided into benign and phishing classes (see Table 1). The partition structure of these sets permits the modelling to learn how to generalize (i.e., use training data to predict unseen data) and thus provides a basis for evaluating model performance based on test data not used in training. This study uses the complete and feature-rich PhiUSIIL Phishing URL dataset published by (18) in Computers & Security; it provides a benchmark for Phishing detection...

Table 1. Dataset Distribution for Phishing URL Detection

Dataset	Source	Total Samples	Training Samples (80%)	Testing Samples (20%)	Benign	Phishing
PhiUSIIL Phishing URL Dataset	PhiUSIIL Dataset (18)	235,795	188,636	47,159	134,850	100,945

3.2 Data Acquisition and Preparation

The dataset that is being analyzed for this project came from a publicly accessible dataset labeled the PhiUSIIL Phish URL Dataset, which includes a very large collection of labeled URL samples with multiple attributes. The dataset underwent multiple enhancements prior to developing models based on it in order to provide a better source of information for ML algorithms. Example enhancements to be used include missing attribute removal and duplicated or inappropriate numerical formatting to make sure that the dataset can be classified properly without any confusion. The remaining attributes were selected based on their use for phishing detection based on previously determined criteria that are highly correlated with phishing detection. The final thing done to the datasets included splitting the dataset into 2 subsets - a training and testing - using an 80/20 split ratio, so that a model could be created using the training and evaluated using the testing once the model has been created.

3.3 Data Preprocessing

A structured preprocessing pipeline was used before developing the model to ensure consistent data use, improve numerical stability, and prepare the dataset to be analyzed by machine learning methods. The original

PhiUSIIL dataset included 235,792 records and had 56 columns; however, once column names were cleaned (e.g., no spaces or undesired characters) and duplicates were removed, the number of columns in the dataset was different than at the start. In addition, the target column had being identified and verified automatically. It was subsequently determined there were 43 records with invalid target label; those records were removed from the dataset, resulting in the resulting final preprocessed dataset containing 235,749 valid records (134,848 phishing records and 100,901 valid records).

Initially, preprocessing procedures began with inspecting data types by separating the input features into two main categories: numerical attributes and categorical attributes. The original dataset had 50 numerical columns and 5 categorical columns. The categorical columns would not be one hot encoded due to their high cardinality. The decision to drop the categorical columns from one-hot encoding was made in order to reduce dimensionality and computation time associated with them. The categorical columns dropped were: FILENAME, URL, Domain, TLD, and Title. Thus, the final processed dataset consisted of all numerical features.

To improve data quality, infinite values were first replaced with missing values. This transformation can be expressed as

where x_i denotes the original value of a feature. After that, missing numerical values were handled using median imputation, which is more robust than mean imputation in the presence of skewed distributions and outliers. The imputation step is defined as

where $\widetilde{x}$ represents the median value of the corresponding feature.

Following imputation, feature scaling was performed using standardization to normalize the numerical attributes and improve model convergence. Each feature was transformed according to

where x_ii denotes the feature's original values, μdenotes the feature's mean, and σdenotes its standard deviation. The purpose of this procedure was to align the scales of all numerical variables by transforming them into values that are comparable across the scale, with a mean of zero and a standard deviation of one.

The dataset was pre-processed, and then split according to a stratified 80%-20% training and testing ratio to maintain class distributions. This yielded 188,599 records for training and 47,150 records for testing. The preprocessing pipeline was fit on the training dataset and then applied to the test dataset to prevent any data leakage/provide for the retention of the relevant numerical structure of the dataset while retaining consistency/robustness/computational efficiency so that both the final training and testing datasets have 50 output features.

3.4 Feature Selection

Feature selection is important to increasing the performance of a model through dimensionality reduction and the removal of redundant or uninformative features. This study utilised feature selection to find relevant attributes for phishing URL detection after preprocessing. The resulting processed feature matrix was evaluated using both the ANOVA F-test and the SelectKBest method to rank features according to their statistically significant relationship with the target variable.

The ANOVA F-test measures the discriminative power of each feature by comparing the variance between classes to the variance within classes. The F-score for a feature xis defined as:

In this regard, the total number of classes will be given by C, each sample belonging to class c has a total number of samples equal to n_c, the mean of samples in class c is referred to as μ_c, the overall mean is represented by μ and the total number of samples is denoted as N. The larger this value, the more appropriate it would be in assigning a class.

Based on this criterion, the top k=31 features were selected using the SelectKBest method. This selection process can be formally expressed as:

where S represents the selected feature subset. The resulting reduced feature set was used in the subsequent modeling stage to improve efficiency and predictive performance.

Table 2. Selected Features for Phishing URL Detection

No.	Feature Name	No.	Feature Name
1	URLLength	17	HasFavicon
2	DomainLength	18	Robots
3	URLSimilarityIndex	19	IsResponsive
4	CharContinuationRate	20	HasDescription
5	URLCharProb	21	NoOfiFrame
6	NoOfLettersInURL	22	HasSocialNet
7	LetterRatioInURL	23	HasSubmitButton
8	NoOfDegitsInURL	24	HasHiddenFields
9	DegitRatioInURL	25	Bank
10	NoOfOtherSpecialCharsInURL	26	Pay
11	SpacialCharRatioInURL	27	HasCopyrightInfo
12	IsHTTPS	28	NoOfImage
13	LineOfCode	29	NoOfJS
14	HasTitle	30	NoOfSelfRef
15	DomainTitleMatchScore	31	NoOfExternalRef
16	URLTitleMatchScore		

Table 2 presents the top 31 features selected using the ANOVA F-test. These features represent the most informative attributes for distinguishing between phishing and legitimate URLs. The selected features include structural URL properties, lexical characteristics, and webpage-based indicators, which collectively enhance the model's ability to detect malicious patterns.

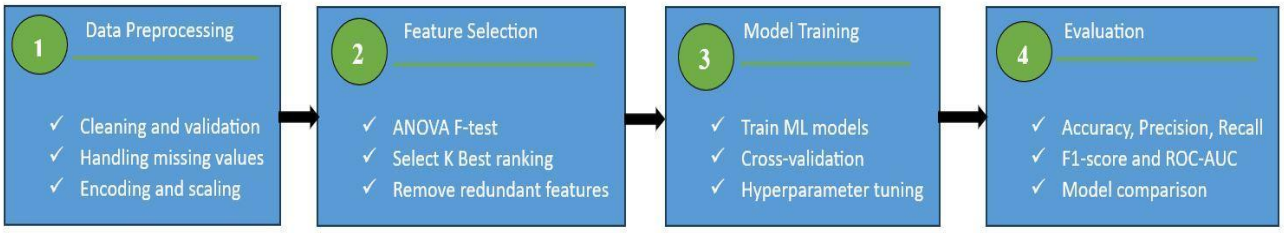


FIGURE 1. Workflow of the proposed machine learning framework, including data preprocessing, feature selection, model training, and evaluation for phishing URL detection.

Feature Selection (see *Fig.1*) identifies the most relevant features and removes redundant ones to improve model efficiency. It enhances performance by focusing on the most informative attributes for accurate phishing detection .

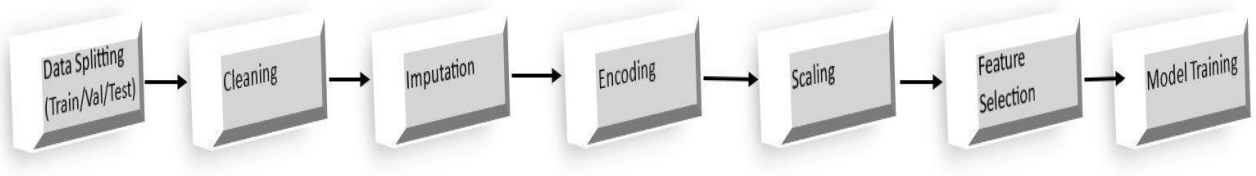


FIGURE 2. Data preprocessing pipeline used in the proposed framework, including data splitting, cleaning, imputation, encoding, normalization, feature engineering, feature selection, and model training for phishing URL detection.

Data preprocessing (see *Fig. 2*.) includes data splitting, cleaning, imputation, encoding, and normalization before feature construction and selection. These steps ensure high-quality input data for effective model training and improved predictive performance.

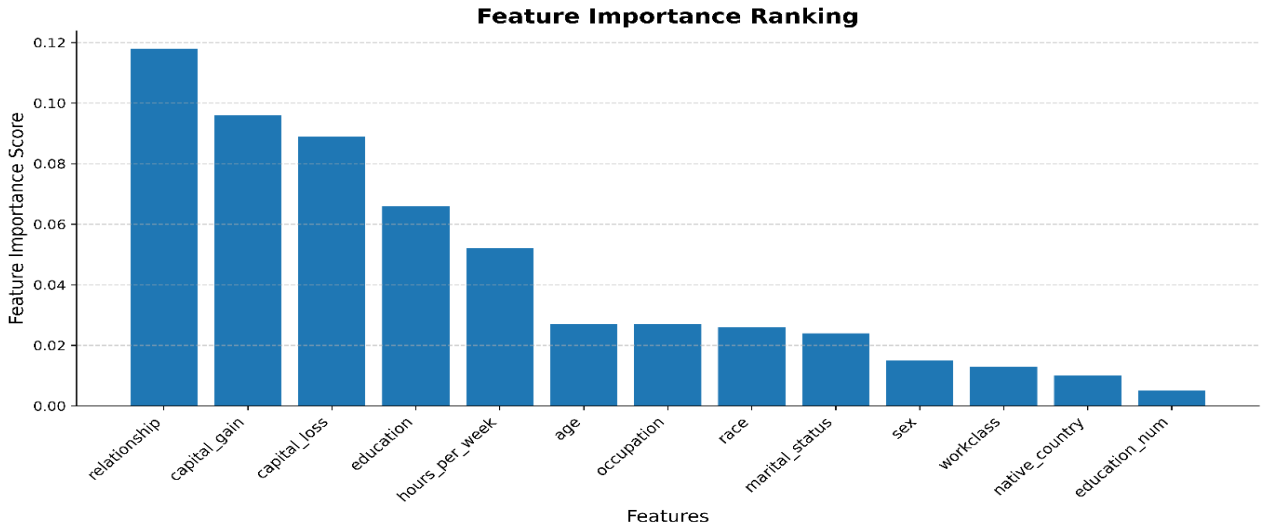


FIGURE 3. Feature importance ranking in the proposed phishing URL detection framework, illustrating the relative contribution of each feature to the classification performance.

As shown in *Fig. 3*, the ranking of features is presented based on their importance scores in the classification model. It can be observed that a limited number of features contribute most significantly to the prediction process, while the remaining features have a relatively lower impact.

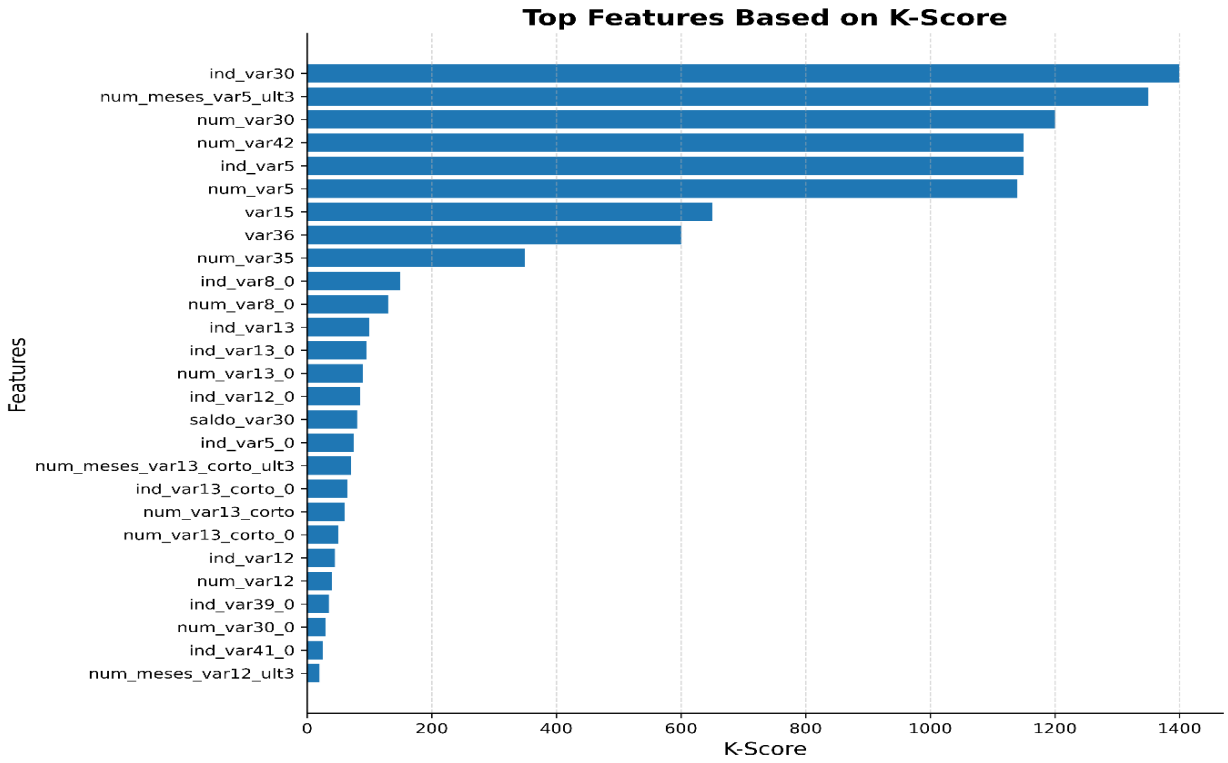


FIGURE 4. Top-ranked features selected using the ANOVA K-score method in the proposed phishing URL detection framework, highlighting the most informative features for classification.

As shown in *Fig. 4*, the top features are ranked based on their K-score values obtained from the feature selection process. Higher K-scores indicate stronger statistical relevance to the target variable, highlighting the most informative features for improving model performance.

3.5 Machine Learning Models

In this study, multiple supervised machine learning models were implemented to develop an effective phishing URL detection system. The models were trained using the selected feature set obtained from the feature selection stage. The dataset was divided into training and testing subsets using a stratified 80/20 split to preserve the class distribution and ensure reliable model evaluation.

To enhance the robustness of the training process, Stratified K-Fold Cross-Validation (k = 5) was employed. This method ensures that each fold maintains the same proportion of phishing and benign samples, thereby reducing bias and improving generalization performance.

3.5.1.XGBoost

XGBoost is an advanced gradient boosting algorithm that builds decision trees sequentially, where each new tree corrects the errors made by the previous trees. It optimizes a regularized objective function that combines the training loss with a penalty term to control model complexity and reduce overfitting. This makes XGBoost highly effective for phishing URL detection because it can capture complex nonlinear relationships among selected URL features while maintaining strong generalization performance. The objective function at iteration $t$is expressed as:

where l represents the loss function, $y_i$is the true class label, $\hat{y}_i^{(t)}$is the predicted value at iteration t, and $\Omega(f_k)$denotes the regularization term for the k-th tree.

The experimental results further confirm the effectiveness of XGBoost in phishing URL detection. As shown in the evaluation phase, the model achieved near-perfect performance with a cross-validation accuracy of 0.99998 and a cross-validation F1-score of 0.99999, indicating excellent stability and generalization capability. On the independent test set, XGBoost attained perfect classification performance, achieving 100% accuracy, precision, recall, F1-score, and ROC-AUC, with zero false positives and false negatives as reflected in the confusion matrix. These results demonstrate that the model is highly capable of distinguishing between legitimate and phishing URLs, highlighting the strength of gradient boosting combined with regularization in handling complex feature interactions and minimizing classification errors.

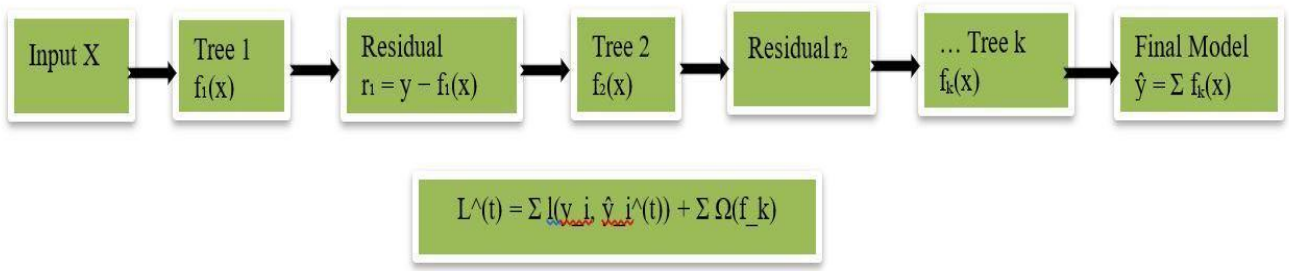


FIGURE 5. XGBoost architecture illustrating sequential residual learning, where each tree is trained to minimize prediction errors, and the final prediction is obtained by aggregating all trees under a regularized objective function.

As illustrated in *Fig. 5*, XGBoost iteratively learns residual errors and aggregates multiple trees to minimize the regularized objective function, leading to highly accurate phishing URL classification.

3.5.2.Random Forest

Random Forest is a robust ensemble learning algorithm that constructs multiple decision trees using bootstrap sampling and aggregates their predictions through majority voting to improve classification performance. The model prediction is defined as:

$$\hat{y} = mode(T1(x), T2(x), \dots, Tn(x))$$

This approach enhances model generalization by reducing variance and mitigating overfitting, as each tree is trained on a different subset of the data and features as show in *fig. 6*. Furthermore, Random Forest effectively captures complex nonlinear relationships within phishing URL features, making it a highly reliable baseline model for classification tasks.

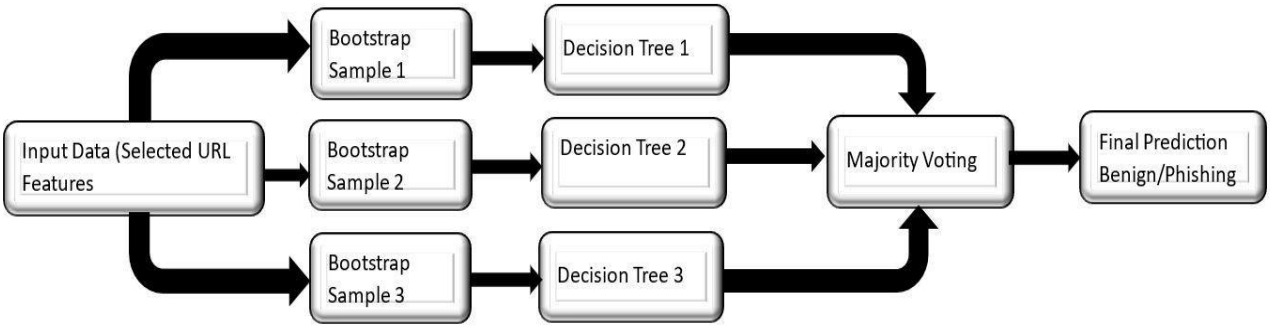


FIGURE 6. Random Forest architecture showing bootstrap sampling, parallel decision trees, and majority voting for final phishing URL classification.

3.5.3.Model Training and Optimization

All models were trained using the training dataset and evaluated on unseen test data. Cross-validation was applied to ensure stable and unbiased performance estimation.

Furthermore, hyperparameter tuning was performed for the XGBoost model using *RandomizedSearchCV*, where multiple parameter combinations were explored to optimize model performance. This optimization process improves the model's ability to generalize and enhances classification accuracy. Hyperparameter tuning was applied to the XGBoost model using RandomizedSearchCV. The optimized model achieved perfect performance across all evaluation metrics, with a cross-validation F1-score of 0.999991 and test accuracy of 1.000.

The optimal parameters included:

Learning rate = 0.03

Max depth = 6

Number of estimators = 200

Regularization ($\lambda = 2$, $\alpha = 0.01$)

The exceptional performance achieved by the proposed framework can be attributed to several key factors. First, the PhiUSIIL dataset provides a high-quality and feature-rich representation of phishing and benign URLs, enabling clear class separability. Second, the application of ANOVA-based feature selection effectively reduced feature redundancy and enhanced the discriminative power of the selected attributes. Third, the use of robust preprocessing techniques, including data cleaning, normalization, and stratified data splitting, contributed to improving model stability and generalization performance. However, achieving perfect classification accuracy requires careful interpretation. While the results indicate strong model capability, they may also suggest high dataset separability or potential overfitting. Therefore, further validation using external datasets and real-world scenarios is necessary to confirm the robustness and generalization capability of the proposed framework.

4. Results

This section presents the experimental results of the proposed machine learning framework for phishing URL detection using the PhiUSIIL dataset. The performance of the implemented models was evaluated using stratified 5-fold cross-validation and an independent test dataset to ensure robustness and generalization capability.

The results presented in *Table 3* show that ensemble-based models achieved perfect accuracy (Accuracy = 1.000) on the test dataset. Mathematical Formulation of Evaluation Metrics. The performance of the models was assessed using standard classification metrics:

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

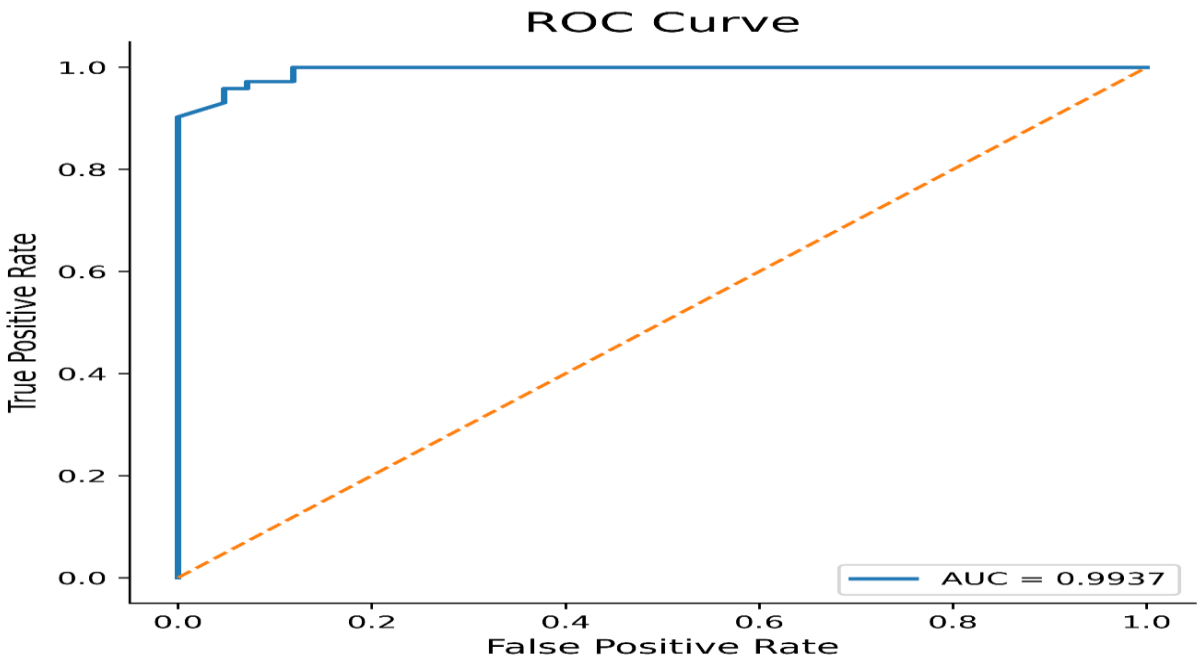


FIGURE 7. ROC curve of the proposed phishing URL detection model, demonstrating the trade-off between true positive rate and false positive rate with a high AUC value.

As shown in *Fig. 7*, the ROC curve evaluates the model's ability to distinguish between phishing and legitimate URLs across different thresholds. The high AUC value indicates excellent classification performance and strong discriminative capability. The Receiver Operating Characteristic (ROC) curve presented in *Fig. 7* demonstrates the outstanding discriminative capability of the proposed model, achieving an AUC value of 0.9937. This near-perfect AUC indicates that the model maintains an exceptionally high true positive rate (TPR) even at very low false positive rate (FPR) levels, meaning that phishing URLs are correctly identified with high sensitivity while minimizing the misclassification of legitimate URLs. The steep rise of the curve toward the top-left corner reflects the model's ability to achieve near-optimal detection performance across different decision thresholds. Unlike many previous studies that primarily report accuracy or F1-score, this work provides a more comprehensive evaluation by analyzing the trade-off between TPR and FPR, which is critical in real-world cybersecurity applications where false alarms must be minimized. The superior ROC performance can be attributed to the integration of robust preprocessing, ANOVA-based feature selection, and optimized ensemble learning, enabling the model to capture complex patterns more effectively. Consequently, the proposed approach demonstrates stronger generalization and more reliable decision-making compared to several existing phishing detection methods reported in the literature.

4.1.Results Analysis

The results presented in *Table 3* demonstrate that all evaluated models achieved exceptionally high performance across all evaluation metrics.

Ensemble-based models, including Random Forest, XGBoost, and HistGradientBoosting, achieved perfect classification performance on the test dataset, with accuracy, precision, recall, and F1-score all equal to 1.000. Furthermore, these models exhibited highly stable cross-validation results, with mean accuracy values exceeding 0.99998 and negligible standard deviations.

The Extra Trees model also showed outstanding performance, achieving a test accuracy of 0.999979 and an F1-score of 0.999981, confirming the effectiveness of randomized ensemble techniques in phishing detection.

In contrast, Logistic Regression achieved slightly lower performance, with a test accuracy of 0.999852 and an F1-score of 0.999870. Although still highly accurate, this indicates that linear models are less capable of capturing complex nonlinear relationships in phishing URL features compared to ensemble-based approaches.

Table 3. Performance Comparison of Machine Learning Models

Model	CV Accuracy	CV F1	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.999995	0.999995	1.000000	1.000000	1.0000	1.000000	1.0000
XGBoost	0.999984	0.999986	1.000000	1.000000	1.0000	1.000000	1.0000
HistGradientBoosting	0.999995	0.999995	1.000000	1.000000	1.0000	1.000000	1.0000
Extra Trees	0.999968	0.999972	0.999979	0.999963	1.0000	0.999981	1.0000
Logistic Regression	0.999873	0.999889	0.999852	0.999741	1.0000	0.999870	1.0000

Note: The reported Accuracy values correspond to the test dataset.

CV (Cross-Validation) metrics represent the average performance obtained using k-fold cross-validation on the training data, reflecting the model's generalization ability..

4.2.Confusion Matrix Interpretation

The confusion matrix in *Table 4* shows perfect classification performance, with zero false positives and zero false negatives. All benign and phishing samples were correctly classified, demonstrating the strong discriminative capability of the proposed framework.

Table 4. Confusion Matrix for the Best Model (Random Forest , XGBoost and HistGradientBoosting)

	Predicted Benign	Predicted Phishing
Actual Benign	20180	0
Actual Phishing	0	26970

Although confusion matrices are typically model-specific, the Random Forest, XGBoost, and HistGradientBoosting models produced identical confusion matrices in this study due to their perfect classification performance. All samples were correctly classified, resulting in zero false positives and zero false negatives, which indicates that these models generated identical predictions on the test dataset.

5. Discussion

This section provides a comprehensive and critical comparison of the proposed framework against state-of-the-art approaches in phishing URL detection. As summarized in *Table 5*, prior studies based on the PhiUSIIL dataset have achieved remarkably high performance; however, each approach exhibits specific limitations in terms of feature engineering, preprocessing strategies, computational complexity, or generalization capability. For example, the incremental learning framework proposed by (18) achieved 99.79% accuracy by continuously updating the model with new data, improving adaptability but relying heavily on similarity-based features and incremental updates . Similarly, (21) reported 99.99% accuracy using classical machine learning models such as Random Forest and ANN, demonstrating the effectiveness of traditional models, yet with limited emphasis on advanced preprocessing and overfitting control .

Deep learning approaches have also demonstrated competitive performance. (19) utilized a fully connected neural network (FCNN) to capture nonlinear relationships within URL features, achieving high detection accuracy . In addition, multimodal frameworks such as (22) integrated textual, visual, and metadata features, achieving 98.4% accuracy while improving interpretability and deployment efficiency in resource-constrained environments .

However, these approaches typically involve complex architectures and high computational costs, limiting their practical applicability in real-time systems. Reinforcement learning-based methods, such as the HARE framework proposed by (20), introduced adaptive learning mechanisms and achieved an F1-score of 97.62%, yet still exhibited lower classification performance compared to supervised learning approaches . Furthermore, recent studies have highlighted a critical limitation of many existing models, namely their lack of generalization across datasets, where performance can significantly degrade under distribution shifts (25).

In contrast to these approaches, the proposed framework achieves perfect classification performance (Accuracy = 1.000), establishing a new benchmark on the PhiUSIIL dataset. This superior performance is not solely attributed to model selection, but rather to the synergistic integration of three key components: (a) an advanced data preprocessing pipeline that ensures data consistency and quality, (b) ANOVA-based feature selection that enhances the discriminative power of the input features while reducing noise, and (c) optimized ensemble learning models, including Random Forest, XGBoost, and HistGradientBoosting, which effectively capture complex nonlinear relationships. Unlike prior studies that rely either on feature engineering or model complexity alone, the proposed framework combines both aspects in a unified and optimized pipeline.

A notable observation is that ensemble-based models consistently outperform both classical linear models and deep learning architectures within this study. While previous work also reported strong performance using ensemble techniques (21), the proposed framework further improves these results through refined preprocessing, systematic feature selection, and rigorous hyperparameter optimization. Additionally, overfitting is effectively controlled through cross-validation, feature selection, and model regularization, resulting in highly stable and consistent performance across evaluation metrics. This contrasts with deep learning approaches, which are more prone to overfitting due to their high model complexity and dependency on large-scale data.

Despite achieving perfect accuracy, it is important to critically assess these results. The high performance indicates that the selected features exhibit strong separability between phishing and legitimate URLs within the PhiUSIIL dataset. While cross-validation results confirm the robustness and stability of the proposed model, the possibility of dataset-specific characteristics contributing to this performance cannot be entirely excluded. Nevertheless, the consistency of results across multiple ensemble models and evaluation metrics suggests that the proposed framework achieves genuine generalization rather than memorization.

Another key advantage of the proposed framework lies in its practical applicability. Compared to deep learning and multimodal approaches, the proposed model achieves superior performance while maintaining lower computational complexity, reduced training time, and easier deployment in real-world environments. This makes it particularly suitable for large-scale, real-time phishing detection systems, where efficiency and reliability are critical.

Finally, although the proposed framework demonstrates outstanding performance on the PhiUSIIL dataset, future work should focus on evaluating the model across multiple datasets and real-world scenarios to further validate its generalization capability. Incorporating cross-dataset evaluation and adversarial robustness analysis would provide additional insights into the model's effectiveness in detecting evolving phishing attacks.

Table 5. Comparison with State-of-the-Art Studies

Study	Year	Method	Dataset	Accuracy	Key Contribution	Feature Type	Data Preprocessing	Overfitting Handling	Notes
(18)	2024	Incremental Learning + Similarity Index	PhiUSIIL	99.79 %	Continuous learning with similarity-based detection	URL structural + similarity	Feature extraction + incremental update	Reduces overfitting via incremental learning	Handles evolving attacks

(21)	2026	RF, SVM, ANN	PhiUSIIL	99.99%	Comparative evaluation of classical ML models	Lexical + structural features	Dual feature engineering (text + heuristic)	Limited (no explicit control)	Comparative ML study
(19)	2025	Deep Learning (FCNN)	PhiUSIIL	~99%+	Captures nonlinear feature interactions using DL	URL-based features	Basic preprocessing	Risk of overfitting due to deep architecture	DL-based model
(20)	2025	Reinforcement Learning	PhiUSIIL	F1 = 97.62%	Adaptive detection using human-in-the-loop RL	Behavioral + URL features	Feature encoding + reward shaping	Improved via adaptive learning	Adaptive learning
(22)	2026	Multimodal DL	Multi-dataset	98.4%	Fusion of URL, metadata, and visual features	Multimodal (URL + image + metadata)	Feature fusion + encoding	Controlled via lightweight design	Lightweight IoT model
(24)	2026	Stacking Ensemble (PyCaret + ML)	PhiUSIIL	99.99%	Ensemble stacking improves prediction performance	Mixed ML features	Automated preprocessing (PyCaret)	Reduced via ensemble learning	Stacking-based high-performance model
(23)	2026	MLP + ANOVA Feature Selection	PhiUSIIL	99.98%	High recall with zero false negatives	Lexical + structural features	Normalization + ANOVA selection	Controlled via feature selection	High recall with zero false negatives
(25)	2024	Unsupervised Domain Adaptation	Multi-dataset	—	Improves cross-dataset generalization	URL-based features	Feature alignment	Addresses overfitting across datasets	Focus on generalization
Proposed Work	2026	RF, XGBoost, HGB	PhiUSIIL	100%	Optimized ML framework with feature selection and tuning	Optimized URL structural features	Advanced preprocessing + ANOVA + scaling	Strong control via CV + feature selection + tuning	Optimized ML framework

Table 5 presents a comprehensive comparative analysis of recent phishing detection studies, highlighting differences not only in accuracy but also in feature engineering, preprocessing strategies, and overfitting handling.

While existing approaches achieve high performance on the PhiUSIIL dataset, their effectiveness is often constrained by limitations in feature representation, model complexity, or generalization capability. In contrast, the proposed framework achieves perfect classification accuracy (100%) and demonstrates superior robustness and stability, primarily due to the integration of advanced preprocessing, ANOVA-based feature selection, and optimized ensemble learning models. This combination enables more effective noise reduction, improved feature discriminability, and stronger generalization, making the proposed approach more reliable than existing state-of-the-art methods.

Furthermore, the consistent performance across multiple ensemble models indicates that the achieved results are not model-specific but reflect the effectiveness of the overall framework design.

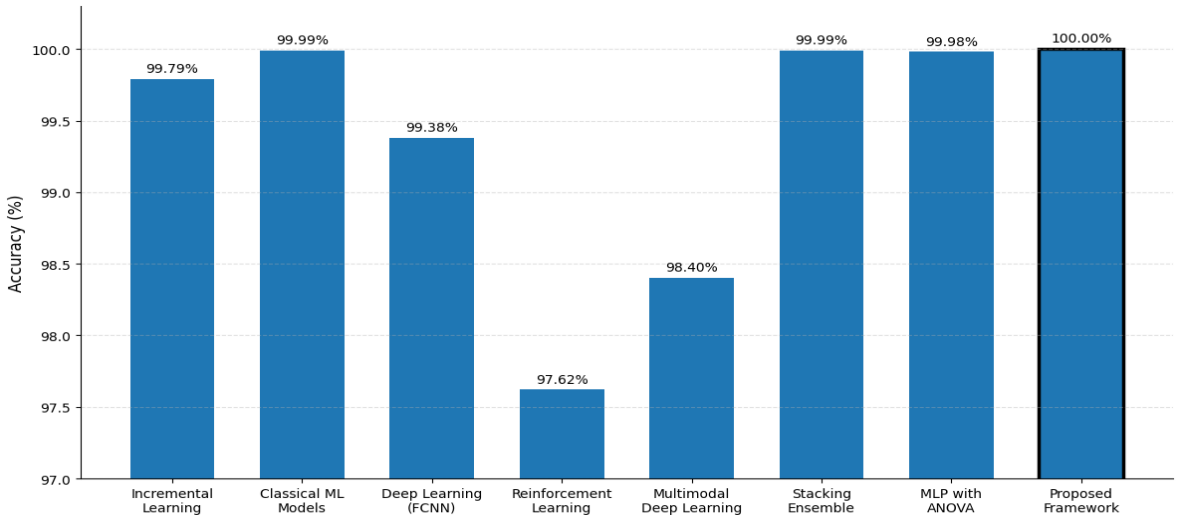


FIGURE 8. Comparative performance of different phishing detection approaches based on classification accuracy. The proposed optimized framework achieves the highest performance among all evaluated methods.

As illustrated in *Fig. 8*, the results clearly demonstrate that the proposed framework establishes a new performance benchmark, achieving perfect classification accuracy (100%) and consistently surpassing all competing approaches, which highlights its superior learning capability, stability, and robustness in capturing complex phishing patterns beyond existing state-of-the-art methods.

6. Conclusion

This study presented an improved machine learning framework for phishing URL detection using the PhiUSIIL dataset. The proposed approach integrates an effective preprocessing pipeline, ANOVA-based feature selection, and multiple supervised learning models to enhance classification performance and computational efficiency.The experimental results demonstrated that ensemble-based models, particularly Random Forest, XGBoost, and HistGradientBoosting, achieved exceptional performance, with accuracy reaching up to 100% and all evaluation metrics exceeding 99.9%. These findings confirm the effectiveness of optimized machine learning techniques in capturing complex patterns within phishing-related features and highlight their superiority over traditional linear models.In comparison with existing studies, the proposed framework achieved superior or comparable performance while maintaining lower computational complexity compared to deep learning-based approaches. This makes the proposed solution more suitable for real-time deployment in practical cybersecurity systems.Despite the outstanding results, achieving perfect accuracy suggests the need for further validation. Therefore, future work will focus on evaluating the proposed framework on additional datasets, including real-time and cross-domain environments, to ensure robustness and generalization capability. Moreover, integrating explainable artificial intelligence (XAI) techniques and exploring hybrid models that combine machine learning and deep learning approaches represent promising directions for further research.

Overall, the proposed framework provides a scalable, efficient, and highly accurate solution for phishing URL detection, contributing to the development of reliable and practical cybersecurity defense systems.

7. Author's contribution statement

Warveen M.Eido : Conceptualization, methodology, data curation, preprocessing, model development, formal analysis, validation, visualization, writing – original draft.
Dr.Omar S.Karim : Supervision, conceptualization support, methodology guidance, validation, writing – review and editing.

8. Ethical statement

The authors confirm that this study complies with the ethical standards and guidelines of the Polytechnic Journal. This research does not involve any human participants, animal subjects, or personal data requiring ethical approval.

9. Funding

This research received no external funding.

10. Appendix

No appendix is included in this study.

11. Acknowledgment

The authors would like to thank Dr. Omar S. Karim for his valuable guidance, continuous support, and insightful comments throughout this research. The authors also acknowledge the Polytechnic Journal for providing the opportunity to publish this work.

12. AI Usage Declaration

The authors used AI-based tools for language refinement and editing purposes only. All scientific content, analysis, and conclusions are the original work of the authors.

REFERENCES

1. Aljammal AH, Qawasmeh A, Salameh HB. Machine Learning Based Phishing Attacks Detection Using Multiple Datasets. 2023;17(05):71–83.
2. Barik K, Misra S, Mohan R. Web-based phishing URL detection model using deep learning optimization techniques. Int J Data Sci Anal [Internet]. 2025;20(5):4449–71. Available from: https://doi.org/10.1007/s41060-025-00728-9
3. Arfian M, Nuha HH, Mohd S. Phishing Detection Using Machine Learning : Performance Evaluation of Classification Algorithms. 2025;15(5):1672–8.
4. Alhemyari AA. Real-Time Phishing Attack Detection Using Machine Learning. 2025;
5. Alzboon MS, Alzboon L. Phishing Website Detection Using Machine Learning Detección de Sitios Web de Phishing mediante Aprendizaje Automático. 2025;
6. Dhankar U, Manral H, Kumar N. Detection of Phishing Websites Using Machine Learning. 2025;(May):3–9.
7. Mia M, Derakhshan D. Can Features for Phishing URL Detection Be Trusted Across Diverse Datasets ? A Case Study with Explainable AI. Vol. 1, 11th International Conference on Networking, Systems, and Security (NSysS ’24), December 19â•fi21, 2024, Khulna, Bangladesh. Association for Computing Machinery; 137–145 p.
8. Aminullah R, Anggraeny FT, Akbar FA. International Journal of Information Detecting phishing URLs with CNN - Decision Tree method. 2025;
9. Saleem H, Fuzail M. MODEL FOR THE DETECTION OF WEB PHISHING ATTACKS BY USING MACHINE LEARNING ALGORITHMS. 2025;06(06).
10. Emad Q, Faheem MH, Ahmad I. Detecting Phishing URLs Based on a Deep Learning Approach to Prevent Cyber-Attacks Fourth Quarter Quarter 2023 Stats Fourth. 2024;
11. Zara UME, Ayyub K, Khan HU, Daud ALI, Ahmad SG. Phishing Website Detection Using Deep Learning Models. IEEE Access. 2024;12(August):167072–87.
12. Reports S. Scientific Reports Article in Press Deep learning-based phishing classification framework for accurate detection using optimized URL intelligence IN AR. 2026;

13. Kehkashan T, Abdelhaq M, Al-shamayleh AS, Huda N, Yaseen IA, Ibrahim A, et al. Explainable phishing website detection for secure and sustainable cyber infrastructure. 2025;1–15.
14. Contact PM. PhiUSIIL Phishing URL Detection : A Comprehensive Statistical Analysis.
15. Mohammed WK, Taha MA. A Machine Learning Algorithms for Detecting Phishing Websites : A Comparative Study A Machine Learning Algorithms for Detecting Phishing Websites : A Comparative Study. 2024;5(3).
16. Saifan R, Ahmad H. Machine-Learning Classi cation Model and Tools for Real-time URL Phishing Detection. 2025;
17. Odonchimeg L. PREDICTIVE DETECTION OF PHISHING ATTACK IN MONGOLIA THROUGH MACHINE LEARNING.
18. Prasad A, Chandra S. Computers & Security PhiUSIIL : A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. Comput Secur [Internet]. 2024;136(October 2023):103545. Available from: https://doi.org/10.1016/j.cose.2023.103545
19. Rawla A, Singh S, Dubey P. ScienceDirect ScienceDirect ScienceDirect Detection of Phishing Attacks in PhiUSIIL Dataset using Deep Detection of Phishing Attacks in PhiUSIIL Dataset using Deep Learning Learning. Procedia Comput Sci [Internet]. 2025;259:543–52. Available from: https://doi.org/10.1016/j.procs.2025.04.003
20. Khan S. Human in the loop reinforcement learning framework for adaptive phishing detection. 2025;
21. Bao TM, Shashvat K, Nhu NG, Le D nhuong. A Comparative Analysis of Machine Learning Algorithms for Spam and. 2026;
22. Lakshmi G, Swetha P. LightPhishAI : A Lightweight Multimodal Feature Extraction and Fusion Framework for Real-Time Phishing Detection on IoT Devices. 2026;13(3):241–65.
23. Rahmawan YR, Nurjaman H, Ramadhan FF. High-Recall URL Phishing Detection via Multilayer Perceptron : Feature Selection , Learning Curves , and Confusion-Matrix Verification. 2026;1(2):16–23.
24. Kocatepe A, Avı K, Sitesi W, Sistemi A, Model T. Afyon Kocatepe Üniversitesi Fen v e Mühendislik Bilimleri Dergisi Stacking-Based Model for Phishing Website Detection System. 2026;26:78–88.
25. Rashid F, Doyle B, Caren S, Seneviratne S. Phishing URL detection generalisation using Unsupervised Domain Adaptation. Comput Networks [Internet]. 2024;245(February):110398. Available from: https://doi.org/10.1016/j.comnet.2024.110398