GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY

E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Analysis of Feature Selection Effects on Phishing URL Detection Performance

M. Ivandanizar[1*], Taufiq[2]
[1,2]Department of Informatics Engineering, Faculty of Engineering, Malikussaleh University, Aceh, Indonesia
[1]ivandanizar.240170174@mhs.unimal.ac.id, [2]taufiq.te@unimal.ac.id

open access

***Corresponding Author**

Article History:
Submitted: 24 Oktober 2026
Accepted: 27 Oktober 2026
Published: 28 OKTOBER 2026

Keywords:
Feature Selection; Machine Learning; Mutual Information; Phishing URL Detection; URL Classification.

GASET: Global Advances in Science, Engineering & Technology is licensed under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0).

ABSTRACT

Phishing remains a significant cybersecurity threat, highlighting the need for effective methods to identify malicious URLs. Machine learning has been widely applied to phishing detection, but its performance depends partly on the features used for classification. Feature selection offers a potential solution by reducing the number of input features while retaining relevant information. This study aims to analyze the effect of Mutual Information (MI)-based feature selection on the classification performance and training time of Random Forest, Extreme Gradient Boosting (XGBoost), and Linear Support Vector Machine (SVM) for phishing URL detection. The experiments used the PhiUSIIL Phishing URL (Website) Dataset, consisting of 235,795 records and 50 numerical and binary features after preprocessing. MI was used to rank features and construct five feature subsets containing the top 10, 20, 30, 40, and 50 features. Model performance was evaluated using accuracy, precision, recall, F1-score, and training time. Random and domain-based data splitting were also examined to assess performance under different evaluation conditions. The results show that feature selection can reduce input dimensionality while maintaining high classification performance. In particular, XGBoost achieved an F1-score of approximately 99.995% using the top 10 features in the combined URL and webpage feature group. However, the effect of feature selection varied across algorithms and feature groups, indicating that additional features did not consistently improve performance. These findings suggest that MI-based feature selection can support more compact phishing detection models. Further evaluation using independent datasets is needed to assess the generalizability of the results.

INTRODUCTION

Phishing remains a significant cybersecurity threat because it exploits users' trust to obtain sensitive information, including login credentials, financial details, and personal data. One common technique involves distributing deceptive URLs that direct users to websites designed to imitate legitimate services. As phishing techniques continue to evolve, distinguishing malicious URLs from legitimate ones has become an important research challenge. Machine learning has attracted considerable attention in this field because it can identify patterns in URL and website characteristics and use these patterns to classify potentially malicious samples. Safi and Singh (2023) reviewed a range of phishing website detection techniques, while Zieni, Massari, and Calzarossa (2023) examined existing approaches and the challenges associated with identifying phishing websites. Their reviews illustrate the diversity of available detection methods and the continuing need to evaluate their effectiveness under different experimental conditions.

The performance of a machine learning model depends not only on the classification algorithm but also on the features used to represent the data. Phishing detection datasets may contain numerous attributes describing URL structure, domain characteristics, and webpage properties. Although these attributes can provide useful classification information, their contributions are not necessarily equal. Some features may be strongly associated with the target class, whereas others may provide limited additional information. Including all available features can also increase computational requirements without guaranteeing better predictive performance. Feature selection is therefore an important consideration when developing phishing detection models. Bu & Kim (2022) investigated an optimized URL feature-selection approach, while Rani, Foozy, and Mustafa (2023) examined feature-selection techniques for phishing website detection using URL-based characteristics.

Mutual Information (MI) provides a statistical approach to ranking features according to their dependence on a target variable. In supervised classification, MI estimates how much information a feature provides about the class label. The resulting scores can be used to rank candidate features and construct subsets containing different numbers of highly ranked variables. This makes MI useful for investigating whether a smaller feature representation can retain sufficient information for classification while reducing the number of inputs used during training. Vajrobol, Gupta, and Gaurav (2024) investigated MI-based feature selection for phishing URL detection using logistic regression. However, the effects of feature selection may vary across classification algorithms, meaning that the most suitable feature subset size for one model cannot automatically be assumed to be optimal for another.

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY

E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Dataset characteristics are also important when evaluating phishing detection methods. Prasad and Chandra (2024) introduced the PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository , which contains 235,795 records, including 100,945 phishing URLs and 134,850 legitimate URLs. The dataset combines URL-related characteristics with information derived from webpages, providing a structured representation for supervised classification. Nevertheless, high performance on a particular dataset does not necessarily demonstrate that a model will perform equally well on URLs collected from different domains or at different times. Rashid, Doyle, Han, and Seneviratne (2024) investigated phishing URL detection generalization across domains, highlighting the importance of considering differences between training and evaluation data when interpreting classification results.

Although previous studies have investigated phishing detection, feature selection, and machine learning classifiers, the relationship between feature subset size, predictive performance, and computational cost remains important to examine. Studies may use different datasets, preprocessing procedures, feature-selection strategies, and evaluation protocols, making direct comparisons difficult. Moreover, high classification accuracy alone does not indicate whether a model offers a favorable balance between predictive performance and training efficiency. A controlled evaluation that considers multiple classification metrics and training time can therefore provide a more comprehensive understanding of the effects of feature selection.

To address this issue, this study investigates the effect of Mutual Information-based feature selection on phishing URL detection using three machine learning classifiers: Random Forest, XGBoost, and Linear Support Vector Machine (Linear SVM). The experiments use the PhiUSIIL Phishing URL Dataset and evaluate five feature subset sizes consisting of the top 10, 20, 30, 40, and 50 features ranked by MI. Feature ranking is performed using the training data only, and the dataset is divided into training and testing sets through a stratified 80:20 split. Classification performance is assessed using accuracy, precision, recall, and F1-score, while training time is measured to compare the computational efficiency of the evaluated configurations.

The contribution of this study is a controlled comparison of three classification algorithms across five MI-based feature subset sizes using a common dataset and evaluation procedure. Rather than assuming that increasing or decreasing the number of features will always improve classification performance, the study examines how predictive metrics and training time vary across the tested configurations. The findings are intended to provide empirical evidence for selecting feature subset sizes that balance predictive performance and computational cost. Since the experiments are conducted on a single dataset, the conclusions are interpreted within the evaluated experimental setting and do not, by themselves, establish generalization to all real-world phishing URLs.

LITERATURE REVIEW

Machine Learning for Phishing URL Detection

Machine learning has been widely applied to phishing URL detection because it enables classification models to learn patterns from labeled data and use them to distinguish malicious URLs from legitimate ones. Different algorithms approach this task through different learning mechanisms, and their effectiveness depends on the characteristics of the input features, the composition of the training data, and the evaluation procedure.

Ahammad et al. (2022) investigated several machine learning algorithms for phishing URL detection, including Random Forest, Decision Tree, LightGBM, Logistic Regression, and Support Vector Machine. Their work examined URL-related characteristics and demonstrated the relevance of comparing multiple classifiers for identifying malicious URLs. Alani and Tawfik (2022) developed PhishNot, a cloud-based machine learning system for phishing URL detection. Their research illustrates how machine learning can be incorporated into a practical detection framework rather than being evaluated solely as an isolated classification algorithm.

Random Forest combines predictions from multiple decision trees, enabling it to represent complex relationships between input variables. XGBoost employs a boosting strategy in which successive trees are trained to improve the predictions of the existing ensemble. Linear SVM, in contrast, identifies a linear decision boundary that separates classes in the feature space. These differences provide a useful basis for investigating whether changes in feature subset size affect different classifiers in similar or different ways.

Mushtaq, Javed, Mohd, and Ud (2024) investigated an ensemble learning approach for URL phishing detection using feature selection and voting-based classification. Their experiments considered datasets containing different numbers of features and evaluated the effect of combining classifiers. The study illustrates that feature selection can be integrated with ensemble learning to develop phishing detection systems. However, ensemble models combine predictions from multiple learners, whereas the present study evaluates Random Forest, XGBoost, and Linear SVM as individual classifiers. This distinction allows the effects of feature subset size to be examined without introducing an additional ensemble layer.

Mousavi and Bahaghighat (2025) examined machine learning and deep learning approaches for phishing website detection under different feature-representation scenarios. Their experiments considered feature-selection strategies and compared several classification approaches, including Random Forest and SVM. Their findings illustrate that feature representation and model configuration can influence classification performance. Nevertheless, results obtained using one

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

dataset and experimental configuration cannot automatically determine which classifier will perform best under a different feature-selection procedure.

Taken together, these studies indicate that phishing detection performance depends on the interaction between classifier choice, feature representation, and experimental conditions. A consistent comparison of several algorithms using the same feature-ranking procedure is therefore useful for investigating how classification performance responds to changes in feature subset size.

Feature Selection and Mutual Information

Feature selection aims to identify a subset of input variables that retains useful information for a prediction task. In phishing URL detection, available features may describe URL length, domain characteristics, suspicious character patterns, and other properties derived from URLs or webpages. Although these variables can provide relevant classification signals, not every feature necessarily contributes equally. Removing less informative variables may reduce the dimensionality of the input data, although excessive reduction can also exclude information needed to distinguish phishing URLs from legitimate ones.

Bu and Kim (2022) investigated an optimized feature-selection approach that combines a genetic algorithm with deep learning for phishing website detection. Their work demonstrates how feature selection can be incorporated into a classification framework to identify a more suitable input representation. Rani et al. (2023) also examined feature-selection techniques for phishing website detection using URL-based characteristics and machine learning. These studies support the importance of evaluating feature relevance instead of automatically treating all available variables as equally informative.

Mutual Information is a filter-based feature-ranking method that measures the statistical dependence between a feature and a target variable. In supervised classification, MI estimates how informative an individual feature is about the class label. Features can subsequently be ranked according to their MI scores, allowing a predefined number of highly ranked features to be selected for model training. Because MI is not restricted to measuring linear relationships, it can be useful for ranking features in datasets containing diverse URL and webpage characteristics.

Vajrobol et al. (2024) investigated MI-based feature selection for phishing URL detection using logistic regression. Their study provides a relevant foundation for examining the role of MI in this domain. However, the usefulness of a selected feature subset may differ across classification algorithms because different models exploit the information contained in the input variables in different ways. A subset that performs well with one classifier does not necessarily produce the same results with a tree-based model or a linear support vector classifier.

Bari et al. (2025) proposed a filter-based feature-selection framework for phishing URL detection that combines several procedures, including the removal of constant, duplicate, quasi-constant, and correlated features, as well as Mutual Information and analysis of variance. Their framework was evaluated using multiple machine learning classifiers and ensemble configurations. The study illustrates how feature-selection procedures can be integrated into a broader phishing detection pipeline. However, when several selection procedures are combined, the contribution of an individual method may be difficult to isolate without additional controlled comparisons.

Alazaidah et al. (2026) investigated feature-selection techniques for identifying important features in phishing website detection. Their work further emphasizes the importance of evaluating feature relevance when developing classification models. In addition, Shafin (2025) proposed an explainable feature-selection framework that uses SHAP and LIME to identify informative features for phishing website detection. These approaches differ from MI in their feature-selection mechanisms, but they share the broader objective of identifying useful input variables while evaluating their relationship with model performance.

The number of selected features is therefore an important experimental variable. A smaller subset may reduce computational requirements but could exclude information that contributes to classification. Conversely, retaining more features may preserve additional information without producing a meaningful improvement in predictive performance. The relationship between feature subset size, classification performance, and training time should consequently be established through empirical evaluation.

Based on this consideration, the present study uses MI to rank 50 predictive features and evaluates five subset sizes: the top 10, 20, 30, 40, and 50 features. The same feature-ranking procedure is applied to Random Forest, XGBoost, and Linear SVM. Feature ranking is performed using the training data only, preventing test labels from influencing the feature-selection process and providing a more consistent basis for comparing the evaluated configurations.

Classification Performance, Computational Efficiency, and Generalization

The evaluation of phishing detection models requires consideration of several performance measures. Accuracy describes the proportion of correctly classified samples, while precision measures the proportion of predicted positive samples that are correctly classified. Recall indicates the proportion of actual positive samples successfully identified, and the F1-score combines precision and recall through their harmonic mean. These metrics provide complementary information about classification performance and help prevent conclusions from relying exclusively on overall accuracy.

Ahammad et al. (2022) compared multiple classification algorithms for phishing URL detection, illustrating the importance of evaluating different models under a common experimental setup. Mousavi and Bahaghighat (2025) also

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

examined classification performance across several machine learning and deep learning scenarios. Their work demonstrates that performance depends on the combination of algorithm, feature representation, and experimental configuration. Consequently, the relative performance of the classifiers in the present study must be determined from the experimental results rather than inferred from previous studies using different datasets.

Computational efficiency is another important consideration when evaluating feature selection. Reducing the number of input variables may lower the computational workload, but the actual effect on training time depends on the learning algorithm, implementation, and data characteristics. Bari et al. (2025) incorporated several feature-selection procedures into a phishing detection framework, illustrating how feature selection can be combined with classification to investigate detection performance. However, the computational benefit of reducing feature count should be measured directly for each model configuration rather than assumed from the number of retained features alone.

Generalization is equally important when interpreting phishing detection results. A model may perform well on a particular benchmark dataset but encounter different patterns when evaluated on URLs from previously unseen domains or different collection periods. Rashid et al. (2024) investigated cross-domain generalization in phishing URL detection, highlighting the challenges associated with differences between training and evaluation data. This issue is particularly relevant when URLs from the same domains may appear in both training and testing partitions.

In the present study, the PhiUSIIL dataset provides a common basis for comparing all model configurations. The data are divided using a stratified 80:20 split, and feature ranking is performed using the training data only. Accuracy, precision, recall, F1-score, and training time are then evaluated for each combination of classifier and feature subset size. These procedures support a controlled comparison within the selected dataset, although domain-separated evaluation or testing on an independent dataset would be needed to establish broader generalization.

Research Gap and Study Contribution

Previous studies have investigated phishing detection using different machine learning algorithms, feature-selection techniques, and ensemble architectures. Some have focused on optimizing feature selection within a particular classification framework, while others have combined multiple selection procedures with several classifiers. Although these approaches contribute to the development of phishing detection systems, their findings do not necessarily establish how the number of selected features affects different classifiers under the same experimental conditions.

An important consideration is that the relationship between feature subset size and classification performance may vary across learning algorithms. A smaller subset may retain sufficient information for one classifier but exclude useful variables for another. Similarly, reducing the number of input features does not necessarily produce a proportional reduction in training time. These relationships require direct evaluation rather than assumptions based solely on feature count.

To address this issue, the present study investigates MI-based feature selection using Random Forest, XGBoost, and Linear SVM on the PhiUSIIL Phishing URL Dataset. Five feature subset sizes—10, 20, 30, 40, and 50—are evaluated using a consistent experimental procedure. The comparison considers accuracy, precision, recall, F1-score, and training time.

The main contribution of this study is a controlled comparison of three individual classifiers across five MI-ranked feature subset sizes. By maintaining a common dataset, feature-ranking method, and evaluation procedure, the study examines how feature subset size relates to predictive performance and computational cost. Rather than assuming that a particular subset size is universally optimal, the study evaluates the observed differences across the tested configurations and interprets the findings within the limits of the experimental dataset.

METHOD

Figure 1 illustrates the research methodology, beginning with dataset preparation and preprocessing, followed by stratified data splitting, Mutual Information-based feature selection, model training, and performance evaluation.

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

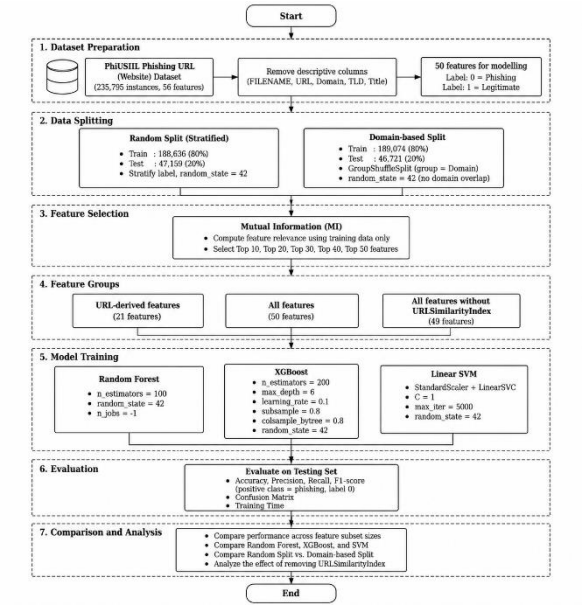


Figure 1. Research Methodology Flowchart for Phishing URL Detection Using Feature Selection and Machine Learning Algorithms

Dataset

The dataset used in this study was the PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository Dataset, obtained from the UCI Machine Learning Repository. It comprises 235,795 records and 56 columns, including 100,945 phishing samples labeled 0 and 134,850 legitimate samples labeled 1. Before model development, five descriptive columns—FILENAME, URL, Domain, TLD, and Title—were excluded from the model inputs, leaving 50 candidate features for classification.

Data Preprocessing and Splitting

Two data-splitting strategies were applied using an 80:20 training-to-testing ratio. The first used stratified random splitting with `random_state=42` to maintain the class distribution in both subsets. The second employed `GroupShuffleSplit`, using Domain as the grouping variable to prevent the same domains from appearing in both training and testing sets. This approach enabled an additional assessment of model performance on previously unseen domains.

Mutual Information-Based Feature Selection

Mutual Information (MI) was used to measure the statistical dependence between individual features and the target label. The MI value is expressed as follows (1):

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right) \qquad (1)$$

Features were ranked using the training data to prevent information from the testing set from influencing the selection process. Five feature subsets were constructed using the top 10, 20, 30, 40, and 50 features. These subsets were evaluated to determine how the number of selected features affected classification performance.

Feature Group Experiments

Additional experiments were conducted to examine the contribution of different feature configurations. Three configurations were compared: 21 URL-derived features, the complete set of 50 features, and 49 features after excluding `URLSimilarityIndex`. These experiments aimed to assess the contribution of URL-derived information and determine how removing `URLSimilarityIndex` influenced model performance.

Classification Models

Three machine learning algorithms were evaluated: Random Forest, XGBoost, and Linear Support Vector Machine (SVM). Random Forest was configured with 100 estimators, while XGBoost used 200 estimators, a maximum tree depth of 6, and a learning rate of 0.1. Linear SVM employed `StandardScaler` for feature standardization, with the regularization parameter set to C=1.

Each algorithm was trained using the selected feature subsets to compare its predictive performance and computational efficiency.

Performance Evaluation

Model performance was assessed using accuracy, precision, recall, and F1-score. The formulas for these metrics are presented below.

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Accuracy measures the proportion of correctly classified samples (2):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision measures the proportion of correctly identified phishing URLs among all URLs predicted as phishing (3):

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Recall measures the proportion of phishing URLs correctly identified by the model (4):

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F1-score represents the harmonic mean of precision and recall (5):

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision \times Recall} \quad (5)$$

Here, TPTP denotes correctly classified phishing URLs, TNTN denotes correctly classified legitimate URLs, FPFP denotes legitimate URLs incorrectly classified as phishing, and FNFN denotes phishing URLs incorrectly classified as legitimate. Since phishing corresponds to label 0 in the dataset, this class was treated as the positive class when calculating precision, recall, and F1-score.

Confusion matrices were also used to examine classification errors, while training time was recorded to assess computational efficiency.

Comparative Analysis

The experimental results were compared across feature subset sizes, classification algorithms, and data-splitting strategies. The analysis also examined the effect of excluding `URLSimilarityIndex` on classification performance. These comparisons were used to evaluate the influence of feature selection and identify model configurations that offered an appropriate balance between detection performance and training efficiency.

RESULT

Dataset Characteristics and Data Splitting

The PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository Dataset contains 235,795 records, consisting of 100,945 phishing samples and 134,850 legitimate samples. After excluding descriptive columns and separating the target label, 50 candidate features were retained for the classification experiments. Phishing samples were assigned label 0, whereas legitimate samples were assigned label 1.

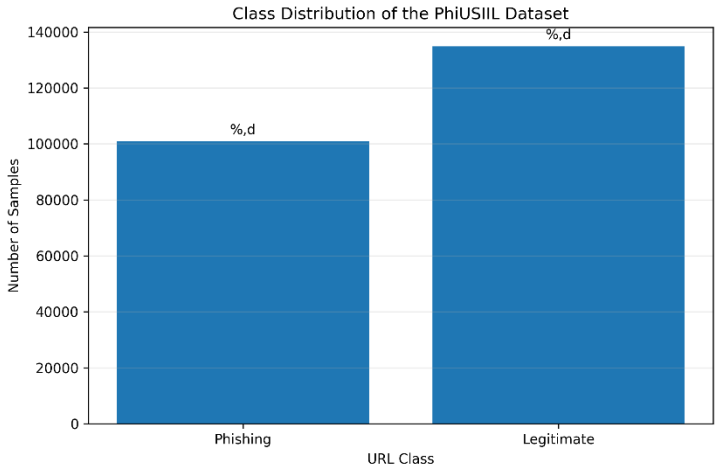


Figure 2. Distribution of Phishing and Legitimate Samples in the PhiUSIIL Dataset

The dataset was evaluated using two data-splitting strategies: stratified random splitting and domain-based splitting. The random split allocated 188,636 records for training and 47,159 records for testing. The domain-based split allocated 189,074 records for training and 46,721 records for testing, with no overlapping domains between the training and testing sets.

Table 1. Dataset Characteristics and Data-Splitting Results

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Parameter	Random Split	Domain-Based Split
Total records	235,795	235,795
Training samples	188,636	189,074
Testing samples	47,159	46,721
Phishing samples in test set	20,189	19,798
Legitimate samples in test set	26,970	26,923
Overlapping domains	Present	0

Note: The domain-based split uses Domain as the grouping variable to prevent domain overlap between the training and testing sets

Mutual Information Feature Ranking

Mutual Information (MI) was used to rank the candidate features according to their statistical dependence on the target label. The feature-ranking process was performed using the training data.

The ten features with the highest MI scores are presented in Table 2. The corresponding scores are also visualized in Figure 3.

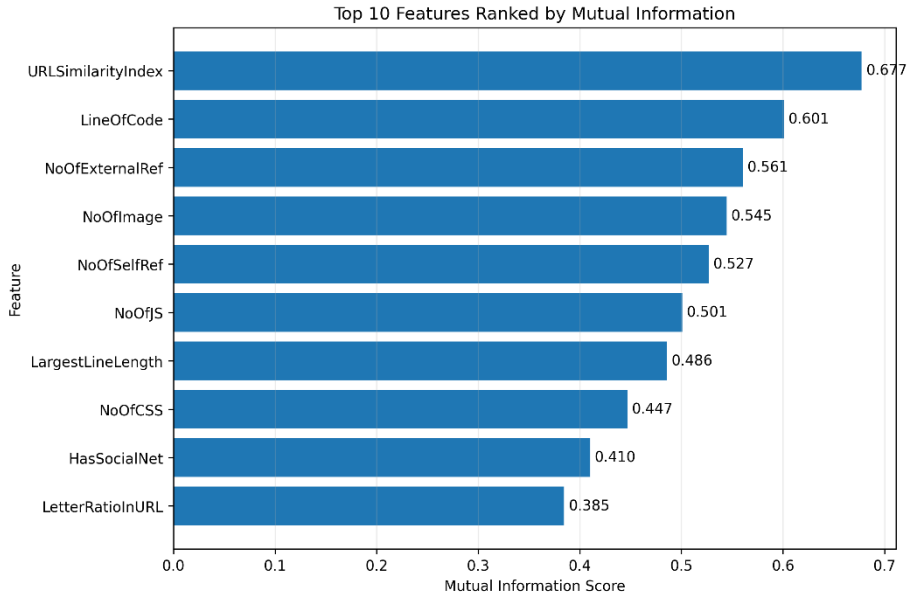


Figure 3. Top 10 Features Ranked by Mutual Information

Table 2. Top 10 Features Ranked by Mutual Information

Rank	Feature	MI Score
1	URLSimilarityIndex	0.677496
2	LineOfCode	0.600989
3	NoOfExternalRef	0.560673
4	NoOfImage	0.544708
5	NoOfSelfRef	0.527011
6	NoOfJS	0.500922
7	LargestLineLength	0.486051
8	NoOfCSS	0.447054
9	HasSocialNet	0.410158
10	LetterRatioInURL	0.384643

Classification Performance Across Feature Groups

The classification experiments compared three feature configurations: 21 operational URL-derived features, all 50 candidate features, and 49 features after excluding `URLSimilarityIndex`. Each configuration was evaluated using all available features or selected subsets based on Mutual Information.

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Table 3 presents the XGBoost performance results using the domain-based split

Table 3. XGBoost Performance Across Feature Groups Using the Domain-Based Split

Feature Group	Configuration	No. of Features	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
URL-derived	All	21	99.7881	99.9341	99.5656	99.7495
URL-derived	Top 10 MI	10	99.7111	99.8681	99.4494	99.6583
URL-derived	Top 20 MI	20	99.7838	99.9290	99.5606	99.7445
URL + webpage	All	50	99.9957	100.0000	99.9899	99.9949
URL + webpage	Top 10 MI	10	99.9957	99.9949	99.9949	99.9949
URL + webpage	Top 20 MI	20	99.9957	100.0000	99.9899	99.9949
Excluding URLSimilarityIndex	All	49	99.9957	100.0000	99.9899	99.9949
Excluding URLSimilarityIndex	Top 10 MI	10	99.7560	99.8531	99.5707	99.7117
Excluding URLSimilarityIndex	Top 20 MI	20	99.9872	100.0000	99.9697	99.9848

Performance Across Feature Subset Sizes

The experiments evaluated feature subsets containing the top 10, 20, 30, 40, and 50 features ranked by Mutual Information. For the sensitivity analysis, feature rankings were recalculated after `URLSimilarityIndex` was excluded.

The XGBoost results for the combined URL-and-webpage feature group and the reduced feature group are summarized in Table 3. The comparison includes the number of retained features and the resulting accuracy, precision, recall, and F1-score.

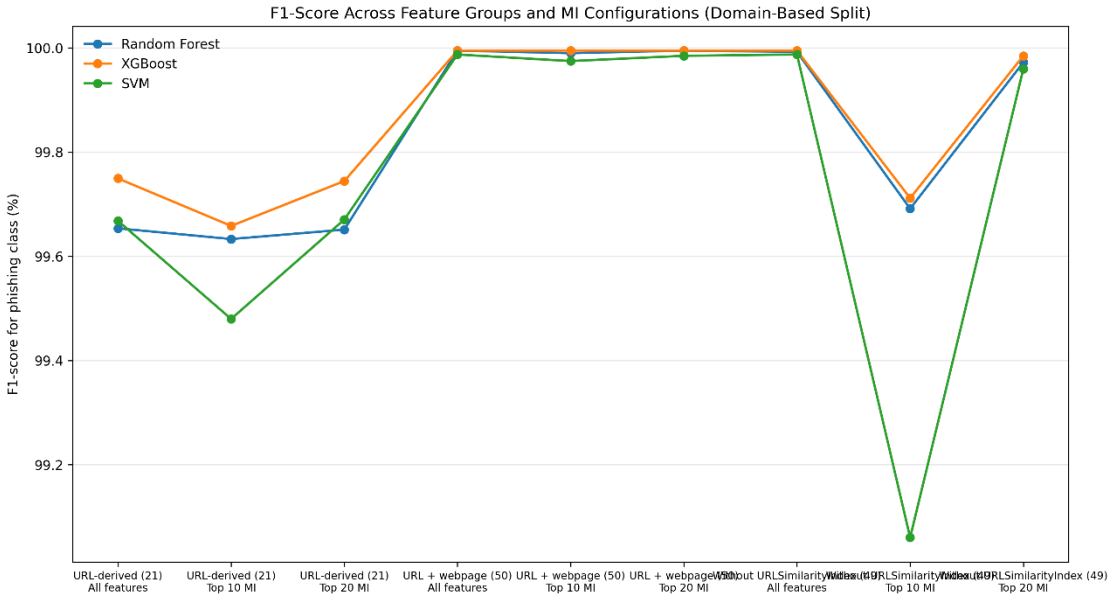


Figure 4. Comparison of F1-Scores Across Feature Subset Sizes and Classification Algorithms

Performance Comparison Between Random and Domain-Based Splits

The classification results were also compared between the random and domain-based splitting strategies. Table 4 presents the currently reported XGBoost F1-scores for the full feature configuration and the configuration excluding `URLSimilarityIndex`.

Table 4. XGBoost F1-Score Under Random and Domain-Based Splitting

Feature Configuration	Random Split F1-Score (%)	Domain-Based Split F1-Score (%)
All 50 features	100.0000	99.9949
49 features excluding URLSimilarityIndex	99.9950	99.9949

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY

E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Confusion Matrix Results

Confusion matrices were generated to record the classification outcomes for phishing and legitimate samples. The matrices use label 0 for phishing and label 1 for legitimate URLs.

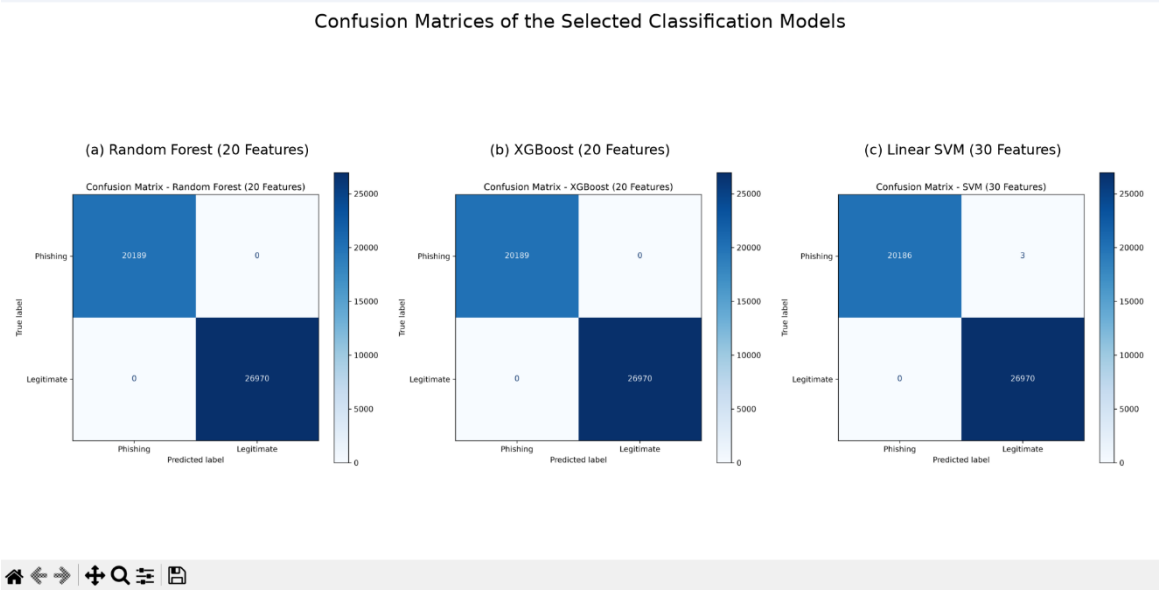


Figure 5. Confusion Matrices of the Evaluated Classification Models.

DISCUSSION

Effect of Feature Selection on Classification Performance

The experimental findings show that mutual information (MI)-based feature selection can reduce the number of input features while preserving high classification performance. However, the extent to which feature selection affects performance depends on the feature group and the classification algorithm being used.

For the URL-derived feature group, XGBoost achieved an F1-score of 99.75% when all 21 features were included. Using the top 10 features resulted in an F1-score of 99.66%, while the top 20 features achieved 99.74%. In comparison, the combined URL and webpage feature group achieved an F1-score of approximately 99.995% using only the top 10 features, which was comparable to the performance obtained with all 50 features.

These results indicate that a relatively small subset of features can retain much of the predictive information needed for phishing URL classification. Nevertheless, reducing the number of features does not always lead to improved performance. The outcome depends on how effectively the selected features capture relevant patterns in the dataset. Therefore, MI-based feature selection offers a potential means of reducing feature dimensionality while maintaining strong classification performance.

Comparison of Classification Algorithms and Feature Groups

The evaluation of Random Forest, XGBoost, and Support Vector Machine (SVM) revealed that all three algorithms achieved high classification performance. Within the URL-derived feature group, XGBoost recorded an F1-score of 99.75%, exceeding the scores obtained by Random Forest and SVM under the corresponding configuration.

The combined URL and webpage feature group produced higher classification performance than the URL-derived feature group. Using all 50 features, XGBoost achieved an F1-score of approximately 99.995%. This result suggests that information obtained from webpage characteristics may provide additional indicators for differentiating phishing URLs from legitimate ones.

However, this performance difference cannot be explained solely by the number of features, since the two groups also contain different types of information. From a practical perspective, URL-derived features may be suitable for preliminary detection, whereas webpage-related features can provide additional information but may require further data collection and processing.

The training-time results also differed across the evaluated algorithms. SVM required less training time in the tested top 10 feature configuration. Nevertheless, selecting an appropriate classifier requires consideration of several factors, including predictive performance, training time, feature extraction costs, and prediction latency.

Generalization Performance and the Role of URLSimilarityIndex

This study employed random and domain-based data splitting to examine model performance under different evaluation conditions. In contrast to random splitting, domain-based splitting prevents domains from appearing in both

GASET
GLOBAL ADVANCES IN SCIENCE, ENGINEERING & TECHNOLOGY
E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

the training and testing sets. This approach provides a more demanding assessment of the model's ability to classify samples associated with previously unseen domains.

In the domain-based experiment, XGBoost achieved an F1-score of approximately 99.995% using 49 features after excluding URLSimilarityIndex. A similarly high F1-score was obtained in the corresponding random-split experiment. These findings demonstrate that the model maintained strong performance under both splitting strategies. However, they do not necessarily indicate that comparable performance would be achieved on independent datasets or against emerging phishing campaigns.

The URLSimilarityIndex feature warrants further examination because it obtained the highest MI score, approximately 0.6775, indicating a strong statistical relationship with the target variable. Its observed distribution also showed a strong association with the class labels. Therefore, understanding how this feature is calculated and whether its information is available during actual prediction is important for evaluating its suitability.

The model continued to achieve a high F1-score after URLSimilarityIndex was excluded, suggesting that its performance did not rely exclusively on this feature. However, this result alone cannot eliminate the possibility of other dataset-specific patterns or potential data leakage. Further investigation into feature provenance and validation using independent datasets is necessary to establish the reliability of the observed performance.

Research Limitations and Future Work

Despite the high classification performance obtained, several limitations should be acknowledged. First, the experiments were conducted using the PhiUSIIL dataset. Consequently, the findings may not fully reflect the characteristics of phishing URLs encountered in other environments. Although domain-based splitting reduces overlap between training and testing domains, evaluation on independently collected datasets remains necessary to assess broader generalization.

Second, the experiments considered selected feature groups and subset sizes. Further studies could investigate additional feature selection techniques and a wider range of feature subset configurations to determine whether comparable performance can be maintained with fewer features.

Third, the reliability and availability of highly informative features, particularly URLSimilarityIndex, require further investigation. Establishing that the selected features rely on information available at prediction time is essential for ensuring the validity of the classification results.

Future research should evaluate the proposed approach using independent datasets and temporally separated samples. In addition, measuring feature extraction time, prediction latency, and computational requirements would provide a more comprehensive assessment of the models' suitability for practical phishing URL detection.

CONCLUSION

This study investigated the effect of mutual information (MI)-based feature selection on phishing URL detection performance using Random Forest, XGBoost, and Support Vector Machine (SVM). The experiments evaluated different feature subset sizes, compared URL-derived features with combined URL and webpage features, and examined model performance using random and domain-based data splitting.

The results demonstrate that MI-based feature selection can reduce the number of input features while maintaining high classification performance. In particular, XGBoost achieved an F1-score of approximately 99.995% using the top 10 features from the combined URL and webpage feature group, comparable to the performance obtained using all 50 features. The combined feature group also achieved higher performance than the URL-derived feature group in the evaluated experiments. These findings indicate that selecting informative features can reduce input dimensionality without necessarily compromising predictive performance.

Furthermore, XGBoost maintained a high F1-score in the domain-based evaluation, including after URLSimilarityIndex was excluded. However, these results are limited to the evaluated dataset and do not establish equivalent performance on independent datasets or in real-world deployment.

Future research should evaluate the proposed approach using independent and temporally separated datasets, investigate additional feature selection methods, and examine the reliability of highly informative features. Measuring feature extraction time and prediction latency would also help determine the practical applicability of the models for phishing URL detection.

REFERENCES

Ahammad, S. H., Kale, S. D., Upadhye, G. D., Pande, S. D., Babu, E. V., Dhumane, A. V., & Bahadur, M. D. K. J. (2022). Phishing URL detection using machine learning methods. *Advances in Engineering Software*, *173*, 103288. https://doi.org/10.1016/J.ADVENGSOFT.2022.103288

Alani, M. M., & Tawfik, H. (2022). PhishNot: A Cloud-Based Machine-Learning Approach to Phishing URL Detection. *Computer Networks*, *218*, 109407. https://doi.org/10.1016/J.COMNET.2022.109407

Alazaidah, R., BaniSalman, M., Alqawasmi, K. E., Abu Zaid, A., Hazaimeh, Y., Alshraiedeh, F. S., & Qumsiyeh, E. (2026). Identifying key features for phishing website detection through feature selection techniques. *Frontiers in*

E-ISSN: 3123-724X
Volume 2, **Number** 1, June 2026
https://doi.org/ 10.62671/gaset.v2i1.362

Computer Science, *7*, 1687867. https://doi.org/10.3389/FCOMP.2025.1687867/TEXT

Bari, N., Saleem, T., Shah, M., Algarni, A., Patel, A., & Ullah, I. (2025). A Filter-Based Feature Selection Framework to Detect Phishing URLs Using Stacking Ensemble Machine Learning. *Computer Modeling in Engineering & Sciences*, *145*(1), 1167–1187. https://doi.org/10.32604/CMES.2025.070311

Bu, S. J., & Kim, H. J. (2022). Optimized URL Feature Selection Based on Genetic-Algorithm-Embedded Deep Learning for Phishing Website Detection. *Electronics 2022, Vol. 11, Page 1090*, *11*(7), 1090. https://doi.org/10.3390/ELECTRONICS11071090

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/10.1111/EXSY.13824;PAGEGROUP:STRING:PUBLICATION

Mushtaq, S., Javed, T., Mohd, M., & Ud, S. ’. (2024). Ensemble Learning-Powered URL Phishing Detection: A Performance Driven Approach. *Journal of Informatics and Web Engineering*, *3*(2), 134–145. https://doi.org/10.33093/JIWE.2024.3.2.10

PhiUSIIL Phishing URL (Website) - UCI Machine Learning Repository. (n.d.). Retrieved October 14, 2026, from https://archive.ics.uci.edu/dataset/967/phiusiil%2Bphishing%2Burl%2Bdataset?hl=id-ID&utm_source=chatgpt.com

Prasad, A., & Chandra, S. (2024). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. *Computers & Security*, *136*, 103545. https://doi.org/10.1016/J.COSE.2023.103545

Rani, L. M., Foozy, C. F. M., & Mustafa, S. N. B. (2023). Feature Selection to Enhance Phishing Website Detection Based On URL Using Machine Learning Techniques. *Journal of Soft Computing and Data Mining*, *4*(1), 30–41. https://doi.org/10.30880/jscdm.2023.04.01.003

Rashid, F., Doyle, B., Han, S. C., & Seneviratne, S. (2024). Phishing URL detection generalisation using Unsupervised Domain Adaptation. *Computer Networks*, *245*, 110398. https://doi.org/10.1016/J.COMNET.2024.110398

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/10.1016/J.JKSUCI.2023.01.004

Shafin, S. S. (2025). An explainable feature selection framework for web phishing detection with machine learning. *Data Science and Management*, *8*(2), 127–136. https://doi.org/10.1016/J.DSM.2024.08.004

Vajrobol, V., Gupta, B. B., & Gaurav, A. (2024). Mutual information based logistic regression for phishing URL detection. *Cyber Security and Applications*, *2*, 100044. https://doi.org/10.1016/J.CSA.2024.100044

Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or Not Phishing? A Survey on the Detection of Phishing Websites. *IEEE Access*, *11*, 18499–18519. https://doi.org/10.1109/ACCESS.2023.3247135