Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Optimizing URL-Based Phishing Detection Using XGBoost and ReliefF Feature Selection

Wahyu Suryaning Tyas[1)*], Fauzi Adi Rafrastara[2)], Wildanil Ghozi[3)]
[1)2)3)] Informatics Engineering, Facullty of Computer Science, Universitas Dian Nuswantoro
[1)]111202214731@mhs.dinus.ac.id, [2)]fauziadi@dsn.dinus.ac.id, [3)] wildanil.ghozi@dsn.dinus.ac.id

Submitted :Dec 1, 2025 | **Accepted** : Dec 29, 2025 | **Published** : Jan 04, 2026

Abstract: Phishing is a significant cybersecurity threat in which attackers exploit manipulated URLs to deceive users and obtain confidential information. As phishing attacks continue to grow in complexity, automated machine learning based detection methods have become essential to strengthen digital security. This study proposes a URL based phishing detection model using boosting algorithms while analyzing the role of feature selection in improving classification performance and computational efficiency. The experiments were conducted on a dataset consisting of 10000 instances with 50 features and balanced class labels. After data preparation, 48 features were retained as input variables, and min max normalization was applied to ensure uniform feature scaling. Three boosting algorithms namely Gradient Boosting, XGBoost, and AdaBoost were evaluated using accuracy, precision, recall, and F1 score. Among these methods, XGBoost achieved the highest accuracy of 98.8 percent, demonstrating its effectiveness in learning complex URL patterns. Subsequently, three feature selection techniques namely Information Gain, Chi Square, and ReliefF were applied and evaluated using 10 fold cross validation. The results indicate that ReliefF provides the most effective feature reduction by selecting 37 features while maintaining the same classification accuracy. Unlike previous studies that mainly focus on classifier comparison, this study demonstrates that integrating XGBoost with ReliefF enables significant feature dimensionality reduction without compromising predictive accuracy. This finding highlights an efficient trade off between detection performance and computational complexity. Overall, the proposed framework offers a robust, efficient, and scalable solution for fast and adaptive phishing detection in modern cybersecurity environments.

Keywords: Feature Selection; Machine Learning; Phishing Detection; ReliefF; XGBoost

INTRODUCTION

Phishing is a form of cyberattack that has a significant impact on user information security. This attack is carried out by sending fake messages or creating websites that mimic trusted sources to obtain sensitive data such as passwords and account credentials. The growing use of web-based services has further increased the opportunities for these attacks to succeed (Budiono et al., 2025). Therefore, users must be more cautious when providing personal information on digital platforms.

The rising number of reports collected by the Indonesia Anti-Phishing Data Exchange (IDADX) highlights the growing urgency of phishing threats. IDADX recorded 6,106 reports in the fourth quarter of 2022, and this number surged to 26,675 in the first quarter of 2023 (Fatiha et al., 2024). Globally, the average loss caused by data breaches in 2023 reached USD 4.45 million, and the FBI reported total losses exceeding USD 10 billion in 2021. These figures demonstrate that phishing has significant economic consequences that extend far beyond a simple privacy threat.

At the national level, (BSSN, 2023) recorded 26,771,610 phishing incidents in 2023. These activities have become a significant threat because they exploit various digital platforms to obtain sensitive data. In addition, 56,128,160 data exposures were found on the darknet, potentially increasing the risk of phishing attacks. Of the 1,814 reported cyber incidents, some involved data theft caused by phishing. Experts predict that this threat will increase in 2025, requiring a more adaptive security response.

Conventional phishing detection methods, such as blocklists and allowlists, have limitations when dealing with new attack variants that can be easily modified. These weaknesses encourage the development of more adaptive approaches that can automatically recognize attack patterns (Zieni et al., 2023). Machine learning emerges as a

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

practical solution because it can analyze URL characteristics and website attributes through learning from data (Sirisha, 2025). This approach enables systems to identify patterns that conventional methods cannot detect. Thus, machine learning offers greater flexibility in addressing evolving attacks.

This study aims to enhance phishing detection accuracy by applying machine learning algorithms combined with feature selection techniques. It evaluates the performance of three boosting algorithms and analyzes the impact of different feature selection methods on model accuracy. The objective is to identify the most effective integration of algorithms and feature subsets to develop a phishing detection system that is both efficient and highly accurate.

Existing phishing detection research has demonstrated the effectiveness of various machine learning models, particularly ensemble-based approaches. For example, (Putri & Wijayanto, 2022) compared several classifiers, including Naïve Bayes, Random Forest, Decision Tree, and SVM, and reported that Random Forest achieved the highest accuracy of 90.77% on a small-scale dataset. Similarly, (Fauzan et al., 2025) showed that Gradient Boosting outperformed Random Forest in email phishing classification, achieving an accuracy of 97.4%. More recently, (Fahri, 2025) applied Random Forest with automated hyperparameter optimization on a large URL-based dataset and achieved an accuracy exceeding 98%. While these studies report strong classification performance, most focus primarily on algorithm comparison or hyperparameter optimization as the main means of improving performance.

However, despite the growing body of literature, the relationship between feature dimensionality and generalization stability in boosting-based phishing detection models remains underexplored. Many existing studies emphasize accuracy improvements without systematically examining how feature selection strategies influence model efficiency, robustness, and stability in high-dimensional URL-based data. As a result, it remains unclear whether reported high performance is achieved through genuinely informative features or merely through increased model complexity.

To address this gap, this study investigates the impact of feature dimensionality on the generalization performance of boosting-based phishing detection models by systematically combining three boosting algorithms, Gradient Boosting, XGBoost, and AdaBoost with multiple feature selection strategies. The scientific contributions of this study are threefold. First, this study provides an empirical analysis of how feature dimensionality affects generalization stability in boosting-based phishing detection models, thereby extending prior accuracy-centric evaluations. Second, it demonstrates that integrating XGBoost with ReliefF enables substantial feature dimensionality reduction without compromising classification accuracy, highlighting the effectiveness of distance-based feature selection in preserving non-linear discriminative patterns in URL-based phishing data. Third, this study advances ensemble learning research by emphasizing feature efficiency, rather than model complexity alone, as a critical theoretical factor in designing robust and scalable phishing detection systems.

METHOD

This study used Orange Data Mining software as the primary tool for data processing and analysis. In general, the research methodology comprises four main stages, namely data preparation, preprocessing, modeling, and model performance evaluation. This methodological structure was designed to systematically analyze the impact of feature selection and boosting-based models on phishing detection performance. The complete research workflow is presented in Figure 1 to illustrate the interrelationships among the analytical stages.

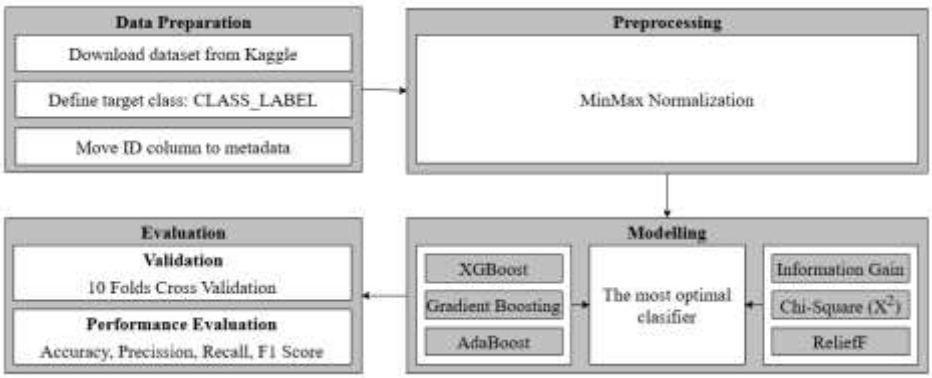

Fig 1. Research process flow

The Data Preparation stage begins by downloading the dataset from Kaggle in .csv format (https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning) and importing it into Orange Data Mining. The ID column is moved to the metadata section to avoid affecting classification, while the

*name of corresponding author

CLASS_LABEL attribute is set as the target. This structuring keeps the dataset consistent, organized, and free of irrelevant attributes. Proper data preparation is essential because it strongly influences model performance and analysis accuracy (Hermawan, 2024; Vidhya et al., 2023).

In the preprocessing stage, Min–Max normalization is applied to standardize each numerical feature to a range of 0–1. This scale adjustment prevents features with large values from dominating the model and improves algorithm stability and convergence efficiency. With uniform feature scales, the model can recognize data patterns more accurately and reduce bias caused by differences in value ranges (Ahmed et al., 2022; Alshdaifat et al., 2021). This stage serves as an essential foundation for the modeling process. The Min–Max normalization is formulated in Equation (1), where X represents the original feature value, $X_{\min}$ is the minimum value of the feature, and $X_{\max}$ is the maximum value. This step provides a crucial foundation for the subsequent modeling process.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

The modeling stage applies three boosting algorithms, namely Gradient Boosting, XGBoost, and AdaBoost, which were selected for their capability to model complex and non-linear patterns commonly observed in phishing-related data. Gradient Boosting employs a stage-wise learning strategy in which each successive model is trained to correct the prediction errors of the preceding model by iteratively fitting residuals. This mechanism enables a structured learning process that enhances sensitivity to complex data distributions, particularly in high-dimensional feature spaces (Rayadin et al., 2024). Moreover, the iterative nature of Gradient Boosting enables it to effectively capture non-linear relationships characteristic of phishing data, making it a suitable baseline for boosting-based analysis.

Building upon the Gradient Boosting framework, XGBoost was developed to achieve higher computational efficiency through parallel processing and optimized memory management. In addition, XGBoost incorporates explicit regularization mechanisms that control model complexity and mitigate overfitting, thereby producing more stable performance on unseen data (Lin et al., 2023; Zhang et al., 2022). Support for multiple objective functions and pruning strategies further enhances its flexibility in handling diverse feature structures and data variations. In this study, XGBoost was implemented using the default parameter configuration provided by Orange Data Mining to ensure methodological consistency and to emphasize the influence of feature selection rather than extensive hyperparameter optimization. Consequently, XGBoost represents an appropriate choice for achieving high predictive performance while maintaining computational efficiency.

In contrast to the previous algorithms, AdaBoost adopts an adaptive learning approach by increasing the weights of misclassified samples in subsequent iterations. This strategy enables the model to focus on more difficult instances, thereby progressively improving classification accuracy (Sukmawati et al., 2025). By combining multiple weak classifiers into a single ensemble, AdaBoost produces a model that is capable of handling challenging decision boundaries. Its adaptive nature makes AdaBoost a competitive and lightweight alternative for comparative evaluation in phishing detection tasks.

After identifying the most effective boosting algorithm, the study proceeds by applying three feature selection methods, namely Information Gain, Chi-Square (χ^2), and ReliefF, to identify features that contribute most significantly to the classification process. This step aims to eliminate irrelevant, redundant, or weakly informative attributes, thereby improving both model efficiency and predictive accuracy. Information Gain evaluates each feature based on its contribution to reducing entropy, with higher values indicating greater relevance for distinguishing between phishing and legitimate websites (Savyanavar et al., 2024). Features with high Information Gain scores are retained because they are considered informative and directly contribute to improving the model's predictive capability. The Information Gain calculation is presented in Equation (2), where $IG(Y, X)$represents the information gain between feature $X$and the class label Y, $H(Y)$denotes the initial entropy, and $H(Y \mid X)$corresponds to the conditional entropy after accounting for feature X.

$$IG(Y, X) = H(Y) - H(Y \mid X) \quad (2)$$

The Chi-Square (χ^2) method evaluates how strongly a feature is related to the target class by comparing the difference between observed frequencies and those expected under independence. Features with high Chi-Square scores are interpreted as having a meaningful association with the phishing category, making them suitable for prioritization during modeling to enhance classification performance (Awasthi & Goel, 2024). The level of dependency between each feature and the class label is determined using Equation (3), where O_i represents the observed frequency in the dataset, and E_i represent the expected frequency assuming independence between the feature and the class label.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$x^2 = \sum_{i=1} \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

In contrast to statistical dependency-based methods such as Chi-Square, ReliefF employs a distance-based evaluation strategy that examines disparities in feature values between nearest-hit instances, which belong to the same class, and nearest-miss instances, which belong to different classes. Features that consistently differentiate between these instances receive higher weights and are therefore considered relevant for classification (Iwanowski et al., 2025). By accounting for local context and inter-feature relationships, ReliefF is particularly effective in handling complex data distributions and non-linear feature interactions. This formulation calculates the feature weights that indicate these contributions using Equation (4), where W[A] represents the weight of feature A, R_i denotes a randomly selected reference sample, H_i is the nearest hit, and M_i is the nearest miss.

$$W[A] = W[A] - \frac{1}{m}\sum_{i=1}^{m} diff(A, R_i, H_i) + \frac{1}{m}\sum_{i=1}^{m} diff(A, R_i, M_i) \quad (4)$$

The selected feature subsets are subsequently integrated with the most effective classification algorithm to construct a phishing detection model that is accurate, efficient, and adaptable to unseen data. Model performance is evaluated using a 10-fold cross-validation strategy, which balances bias and variance and provides a reliable estimate of generalization performance (White & Power, 2023). Within this framework, each fold is used once as the test set, while the remaining folds serve as training data. The final performance metrics are obtained by averaging the results across all iterations. Model evaluation is based on accuracy, precision, recall, and F1-score derived from the confusion matrix, enabling a comprehensive assessment of classification effectiveness (Iwanowski et al., 2025; Rainio et al., 2024).

Although Orange Data Mining offers a visual and user-friendly environment that facilitates rapid experimentation and reproducibility, it has limitations in supporting advanced hyperparameter tuning and fine-grained model control compared to script-based frameworks. Therefore, this study prioritizes methodological clarity and experimental consistency over extensive parameter optimization to ensure fair and reliable model comparison.

In the confusion matrix, TP (True Positive) represents the number of phishing URLs correctly identified as phishing, TN (True Negative) refers to the number of legitimate URLs that are successfully recognized by the model as non-phishing, FP (False Positive) represents legitimate URLs incorrectly classified as phishing, and FN (False Negative) refers to phishing URLs that the model erroneously classifies as legitimate (Narayana et al., 2023). Understanding these components is essential because they form the basis for calculating accuracy, precision, recall, and F1-score, providing insight into both the correctness and reliability of the model's predictions.

Accuracy quantifies the ratio of correctly classified instances to the total number of samples, serving as a widely accepted metric for assessing overall model performance. This metric offers a preliminary understanding of how well the model separates phishing URLs from legitimate ones (Agustina et al., 2024). Because the dataset in this study is balanced, accuracy provides a more reliable reflection of the model's actual effectiveness. The formula for calculating accuracy is shown in Equation (5).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

Precision is used to assess the model's accuracy in identifying phishing data without producing incorrect optimistic predictions (Mahmud Sujon et al., 2024). This metric is highly relevant in the security context because mislabeling legitimate sites as phishing can compromise system reliability and user experience. A high precision value indicates the model's effectiveness in minimizing false positives. The precision formula is shown in Equation (6).

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

Recall measures how effectively the model identifies all samples that are truly phishing, prioritizing minimizing false negatives. This metric is critical because failing to detect even a single attack can pose a significant security risk. A high recall value indicates the model's capability to capture a wide range of phishing threats. The recall formula is presented in Equation (7).

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

F1-score is a combined metric that harmonizes precision and recall, providing a more stable evaluation when the two metrics are imbalanced (Sitarz, 2023). In this study, the F1-score is used to ensure that the model not only produces accurate predictions but also consistently identifies all potential attacks. A high F1-score indicates that the model has balanced and reliable performance. The F1-score formula is presented in Equation (8).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

Overall, the use of accuracy, precision, recall, and F1-score provides a comprehensive assessment of the model's predictive capability. These metrics enable a reliable evaluation of the model's ability to consistently distinguish between phishing and legitimate URLs under diverse data conditions. Consequently, this evaluation framework supports the development of phishing detection systems that are both accurate and operationally reliable in real-world cybersecurity environments.

The experimental design of this study is grounded in feature efficiency and ensemble generalization theory. The use of multiple boosting algorithms establishes a comparative baseline, while the application of three distinct feature selection paradigms—entropy-based, dependency-based, and distance-based—facilitates methodological triangulation. In addition, 10-fold cross-validation is employed to enhance statistical robustness and reduce sampling bias, thereby improving the reliability of generalization analysis.

This research design is specifically structured to address the relationship between feature dimensionality and generalization stability in boosting-based phishing detection models. By combining variations in boosting algorithms with multiple feature selection strategies, the design enables a controlled separation between the effects of model complexity and feature dimensionality. Therefore, this approach provides a methodologically valid and empirically grounded framework for analyzing the role of feature efficiency in improving generalization stability.

RESULT

The dataset used in this study was obtained from Kaggle, titled Phishing_Legitimate_full.csv, and contains 10,000 instances of legitimate and phishing websites with a balanced class distribution. The dataset includes 50 features extracted from URLs and website characteristics. One feature, CLASS_LABEL, is used as the target, with 1 indicating a phishing website and 0 indicating a legitimate website, while the ID column is moved to metadata to prevent it from affecting the classification process. This dataset, therefore, leaves 48 features as input variables for model training. General characteristics of the dataset are presented in Table 1.

Table 1 General characteristics of the dataset

Characteristics	Description
Number of instances	10.000
Number of features	50 total features (48 input features, 1 target feature, 1 metadata/ID feature)
Feature customization	The ID feature is moved to metadata because it has no predictive value
Variable targets	CLASS_LABEL (0=phishing, 1=legitimate)
Class balance	The dataset is balanced between phishing and legitimate classes
Data source	Kaggle (Phishing_Legitimate_full.csv)

The data preprocessing stage involved applying Min–Max normalization to standardize the scale of all numerical features to 0–1. This normalization ensures that each feature contributes equally during model training, preventing specific attributes from dominating due to scale differences. After normalization, all feature values are proportionally distributed and have a uniform range. The preprocessing results in a cleaner, more consistent dataset that is ready for use in the phishing detection model training stage.

The modeling stage was conducted by applying three classification algorithms, namely Gradient Boosting, XGBoost, and AdaBoost, to develop a phishing detection system based on machine learning techniques. These algorithms were chosen for their ability to handle complex, non-linear data patterns and for their proven ability to deliver stable predictive performance across diverse data conditions. The training process involved optimal model parameter tuning and performance evaluation using accuracy, precision, recall, and F1-score metrics to differentiate between phishing and legitimate websites. Evaluation results showed that all three algorithms achieved high accuracy, though performance varied. A comprehensive summary of the results is shown in Table 2 to support selecting the model with the best performance.

Table 2 Comparison of classification model performance

Algorithm	Accuracy	Precision	Recall	F1-Score
XGBoost	98,8%	98,8%	98,8%	98,8%
Gradient Boosting	97,8%	97,8%	97,8%	97,8%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

AdaBoost	97%	97%	97%	97%

Based on the experimental results, XGBoost achieved the highest accuracy (98.8%), outperforming Gradient Boosting (97.8%) and AdaBoost (97.0%). This superiority can be attributed to the combination of regularization mechanisms and more adaptive learning strategies in XGBoost, which enable more effective control of model complexity in high-dimensional data. With built-in regularization, XGBoost can minimize overfitting while maintaining strong generalization performance as the number of features and non-linear interactions increase, a characteristic of URL-based phishing data.

In contrast, conventional Gradient Boosting relies on a stage-wise learning process without explicit complexity control, making it more susceptible to error accumulation when handling complex feature interactions. AdaBoost, which focuses on increasing the weights of misclassified samples, tends to amplify the influence of noise and outliers, leading to reduced performance stability. Therefore, these differences in learning mechanisms and complexity management explain why XGBoost demonstrates more stable and superior performance compared to other boosting methods in the context of URL-based phishing detection.

After identifying the best algorithm, XGBoost, the next stage involved applying three feature selection methods, namely Information Gain, Chi-Square (χ^2), and ReliefF. Feature selection was used to identify the attributes most relevant to the target variable, thereby improving the efficiency and accuracy of model training. This step also aims to reduce noise and eliminate uninformative features that could potentially decrease classification performance. Performance evaluation was conducted by assessing accuracy, precision, recall, and F1-score for the XGBoost model after applying each feature selection method. The test results showed that each method produced different numbers of selected features and model performance, as presented in the following tables.

Table 3 Model performance based on information gain

Number of ranks	Accuracy	Precision	Recall	F1-Score
40	98,7%	98,7%	98,7%	98,7%
41	98,7%	98,7%	98,7%	98,7%
42	98,8%	98,8%	98,8%	98,8%
43	98,8%	98,8%	98,8%	98,8%
44	98,8%	98,8%	98,8%	98,8%

Table 4 Model performance based on Chi-Square (X^2)

Number of ranks	Accuracy	Precision	Recall	F1-Score
42	98,7%	98,7%	98,7%	98,7%
43	98,7%	98,7%	98,7%	98,7%
44	98,8%	98,8%	98,8%	98,8%
45	98,8%	98,8%	98,8%	98,8%
46	98,8%	98,8%	98,8%	98,8%

Table 5 Model performance based on ReliefF

Number of ranks	Accuracy	Precision	Recall	F1-Score
35	98,7%	98,7%	98,7%	98,7%
36	98,6%	98,6%	98,6%	98,6%
37	98,8%	98,8%	98,8%	98,8%
38	98,8%	98,8%	98,8%	98,8%
39	98,8%	98,8%	98,8%	98,8%

Based on Tables 3–5, Information Gain and Chi-Square demonstrate stable classification performance across different numbers of selected features, indicating that both methods effectively identify attributes that are statistically relevant to the target class. Information Gain achieves optimal performance with 42 selected features, while Chi-Square reaches its best performance with 44 features, suggesting that additional features beyond these points contribute limited performance improvement. This behavior reflects the univariate evaluation nature of both methods, which assess features independently and therefore have limited capability to capture inter-feature interactions.

In contrast, ReliefF achieves comparable classification performance using a considerably smaller number of selected features. Optimal accuracy is obtained with only 37 features, highlighting ReliefF's ability to reduce feature dimensionality without degrading model performance. This advantage stems from its distance-based evaluation mechanism, which incorporates local neighborhood information and inter-feature relationships, making

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15651
e-ISSN : 2541-2019
p-ISSN : 2541-044X

it more effective in capturing non-linear patterns commonly present in URL-based phishing data. Consequently, integrating ReliefF with XGBoost results in a more efficient and stable model configuration.

Model validation using a 10-fold cross-validation strategy further confirms the robustness of this configuration. Consistent performance across different data partitions indicates that the proposed model exhibits reliable generalization capability. Moreover, this validation approach helps reduce the risk of overfitting by ensuring that the observed performance is not dependent on a specific data split.

DISCUSSIONS

This study evaluates the superiority of the proposed approach by comparing its performance with recent state-of-the-art (SOTA) methods for URL-based phishing detection using machine learning. The comparison focuses on the algorithms employed, the number of selected features, and the primary evaluation metrics, namely accuracy, precision, recall, and F1-score, as summarized in Table 6. The evaluation results indicate that the XGBoost with the ReliefF model consistently achieves the highest performance. This model achieves 98.8% accuracy and demonstrates strong stability across all evaluation metrics.

Table 6 Comparison of research model performance with recent studies

Study	Model	Number of features	Accuracy	Precision	Recall	F1-Score
Proposed	XGBoost + ReliefF	37	98,8%	98,8%	98,8%	98,8%
(Fahri, 2025)	Random forest	48	98,20%	98,22%	98,22%	98,22%

Although the 0.6% improvement in numerical accuracy appears relatively modest, this difference has significant implications for large-scale phishing detection systems. In real-world scenarios where millions of URLs are processed daily, even a small increase in accuracy can result in thousands of additional phishing URLs being correctly identified. Consequently, this performance improvement directly reduces security risks and enhances the treliability of phishing detection systems.

The superior performance of the proposed model is primarily attributed to the synergistic integration of ReliefF-based feature selection and the XGBoost learning algorithm. (Agustina et al., 2024) argue that distance-based feature selection methods such as ReliefF effectively capture inter-feature interactions and local neighborhood information that are often overlooked by univariate filter-based approaches, including Chi-Square and Information Gain. This property is particularly relevant for URL-based phishing detection, where malicious patterns frequently arise from combinations of multiple URL attributes rather than isolated features. Consequently, ReliefF enables the construction of a more informative and compact feature subset without degrading classification performance.

From an algorithmic perspective, XGBoost further strengthens this efficient feature representation through its built-in regularization mechanisms, optimized tree construction, and effective model complexity control. Previous studies, including (Iwanowski et al., 2025), have shown that boosting algorithms without explicit regularization, such as conventional Gradient Boosting, are more prone to overfitting, whereas adaptive methods such as AdaBoost tend to amplify noise by increasing the influence of misclassified instances. In contrast, XGBoost incorporates regularization directly into the boosting framework, thereby improving generalization stability. As a result, XGBoost maintains consistent prediction performance even when the feature set is reduced.

The effectiveness of the proposed approach is further supported by validation using 10-fold cross-validation, which demonstrates robust generalization across different data partitions. These findings empirically confirm the generalization behavior described by (Oyelakin et al., 2023), specifically regarding XGBoost's ability to maintain stable performance when learning from complex, high-dimensional cybersecurity features. The results indicate that performance gains are not solely dependent on algorithm selection but are also strongly influenced by the applied feature selection strategy.

A comparative analysis of feature selection methods shows that Chi-Square and Information Gain maintain high accuracy, however both tend to select a larger number of features. This condition increases computational cost and the potential for feature redundancy. In contrast, ReliefF offers a better balance between performance and complexity by producing a more compact feature subset, thereby reducing the risk of overfitting and improving overall system efficiency.

From a practical perspective, the combination of high accuracy and feature efficiency results in a phishing detection system that is more reliable and responsive. A smaller number of features enables faster feature extraction and prediction, making the approach well-suited for real-time or near-real-time phishing detection systems. Furthermore, reducing reliance on redundant or noise-prone features enhances the system’s robustness against variations in emerging attack patterns.

Overall, integrating ReliefF-based feature selection with the XGBoost algorithm demonstrates that improvements in phishing detection performance can be achieved not only through algorithmic optimization but

*name of corresponding author

also through efficient feature management. This approach empirically demonstrates that feature reduction improves generalization stability in boosting-based models. By simultaneously emphasizing accuracy and efficiency, the proposed method offers a scalable and practical solution for deployment in modern cybersecurity environments.

These findings support and extend ensemble learning theory by showing that the stability of generalization in boosting-based models is influenced not only by classifier strength but also by the interaction between model regularization and feature selection mechanisms. The results reinforce the theoretical assumption that interaction-aware, distance-based feature selection improves the effectiveness of regularized boosting by preserving non-linear discriminative patterns while reducing overfitting. Consequently, this study strengthens existing ensemble learning frameworks by emphasizing feature efficiency as a critical factor in achieving stable generalization rather than relying solely on increased model complexity.

CONCLUSION

This study demonstrates that integrating the XGBoost algorithm with the ReliefF feature selection method achieves superior performance for URL-based phishing detection, attaining 98.8% accuracy while using a substantially reduced number of features compared to prior approaches. The proposed model outperforms conventional Random Forest–based solutions, which achieve approximately 98.2% accuracy with a larger feature set, indicating that high classification performance does not necessarily depend on high feature dimensionality. These results confirm that ReliefF effectively identifies the most discriminative features, while XGBoost leverages internal regularization to learn complex patterns with strong generalization capability. As a result, the proposed approach improves computational efficiency and reduces model complexity without sacrificing predictive performance.

More importantly, this study explicitly addresses a critical gap in phishing detection research, where prior work has largely focused on classifier comparison or hyperparameter optimization without systematically examining the relationship between feature dimensionality and generalization stability. Through empirical analysis of multiple feature selection strategies within boosting-based models, this work demonstrates that distance-based feature selection plays a decisive role in preserving non-linear discriminative patterns in high-dimensional URL data. These findings move beyond accuracy-centric evaluation and provide evidence that feature efficiency is a key determinant of stable generalization in phishing detection models.

From a theoretical perspective, the results reinforce and extend ensemble learning principles by showing that regularized boosting models combined with interaction-aware feature selection are more suitable for complex cybersecurity data than approaches relying solely on statistically independent feature evaluation. This insight contributes to a deeper conceptual understanding of how feature selection mechanisms interact with ensemble models and highlights feature efficiency as an essential design consideration in boosting architectures. Consequently, this study provides a clearer theoretical foundation for developing robust and scalable phishing detection systems.

Despite these contributions, this study has limitations, particularly the reliance on a static URL-based dataset that may not fully capture the evolving characteristics of contemporary phishing attacks. Future research may extend this work by incorporating dynamic or real-time data, exploring richer feature representations, or integrating adaptive and hybrid ensemble strategies to further enhance robustness against emerging attack patterns. Nevertheless, the findings presented in this study provide a solid foundation for future research on feature-efficient, generalization-stable phishing detection models in large-scale cybersecurity environments..

REFERENCES

Agustina, T., Masrizal, & Irmayanti. (2024). Performance Analysis of Random Forest Algorithm for Network Anomaly Detection using Feature Selection. *Jurnal Dan Penelitian Teknik Informatika*, *8*(2), 1116–1123. https://doi.org/10.33395/sinkron.v8i2.13625

Ahmed, H. A., Ali, P. J. M., Faeq, A. K., & Abdullah, S. M. (2022). An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method. *ARO-The Scientific Journal of Koya University*, *10*(2), 29–37. https://doi.org/10.14500/aro.10970

Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., & El-Salhi, S. M. F. S. (2021). The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. *Data*, *6*(2), 1. https://doi.org/10.3390/data

Awasthi, A., & Goel, N. (2024). An Approach for Efficient and Accurate Phishing Website Prediction Using Improved ML Classifier Performance for Feature Selection. *International Journal of Experimental Research and Review*, *40*(Special Issue), 73–89. https://doi.org/10.52756/ijerr.2024.v40spl.006

BSSN. (2023). *LANSKAP KEAMANAN SIBER INDONESIA 2023*.

Budiono, B., Fadillah, F. R., & Arinudin, N. (2025). The Dangers of Phishing to Personal Data Security. *Formosa Journal of Applied Sciences*, *4*(3), 831–844. https://doi.org/10.55927/fjas.v4i3.61

*name of corresponding author

Fahri, M. (2025). Penerapan Algoritma Random Forest untuk Deteksi Phishing pada Website. *Jurnal Ilmiah Teknologi Sistem Informasi*, *6*(2), 186–194. https://doi.org/10.62527/jitsi.6.2.472

Fatiha, M. R., Setiawan, I., Ikhsan, A. N., & Yunita, I. R. (2024). Optimisasi Sitem Deteksi Phishing Berbasis Web Menggunakan Algoritma Decision Tree. *Jurnal Ilmiah IT CIDA : Diseminasi Teknologi Informasi*, *10*(2), 97–108. https://doi.org/10.55635/jic.v10i2.212

Fauzan, R., Vitianingsih, A. V., Cahyono, D., Maukar, A. L., & Suprio, Y. A. B. (2025). Application of Classification Algorithms in Machine Learning for Phishing Detection. *Indonesian Journal of Machine Learning and Computer Science*, *5*(2), 531–540. https://doi.org/10.57152/malcom.v5i2.1968

Hermawan, G. (2024). *Memahami Peran Dataset dalam Penelitian Kecerdasan Buatan: Kualitas, Aksesbilitas, dan Tantangan*. https://doi.org/10.13140/RG.2.2.34468.49288

Iwanowski, M., Olszewski, D., Graniszewski, W., Krupski, J., & Pelc, F. (2025). The Choice of Training Data and the Generalizability of Machine Learning Models for Network Intrusion Detection Systems. *Applied Science*, *15*(15), 1–22. https://doi.org/10.3390/app15158466

Lin, W., Shi, S., Huang, H., Wen, J., & Chen, G. (2023). Predicting risk of Obesity in Overweight Adults Using Interpretable Machine Learning Algorithms. *Frontiers in Endocrinology*, *14*, 01–10. https://doi.org/10.3389/fendo.2023.1292167

Mahmud Sujon, K., Binti Hassan, R., Tusnia Towshi, Z., Othman, M. A., Abdus Samad, M., & Choi, K. (2024). When to Use Standardization and Normalization: Empirical Evidence from Machine Learning Models and XAI. *IEEE Access*, *12*, 135300–135314. https://doi.org/10.1109/ACCESS.2024.3462434

Narayana, G., Manchala, U. D., Naresh, U., Kiran, S., Kiran, M. A., & Ch, R. K. (2023). Improving Phishing Website Detection with Machine Learning: Revealing Hidden Patterns for Better Accuracy. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*, 2321–8169. https://doi.org/10.17762/ijritcc.v11i8.8353

Oyelakin, A. M., Akanbi, M. B., Ogundele, T. S., Akanni, A. O., Gbolagade, M. D., Rilwan, M. D., & Jibrin, M. A. (2023). A Learning Approach for The Identification of Network Intrusions Based on Ensemble XGBoost Classifier. *Indonesian Journal of Data and Science*, *4*(3), 190–197. https://doi.org/10.56705/ijodas.v4i3.88

Putri, N. B., & Wijayanto, A. W. (2022). Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing. *Jurnal Sistem Komputer*, *11*(1), 59–66. https://doi.org/10.34010/komputika.v11i1.4350

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-56706-x

Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. *Jurnal Teknologi Informasi Dan Rekayasa Komputer*, *5*(2), 111–119. https://doi.org/10.37148/bios.v5i2.128

Savyanavar, A. S., Sankpal, P., & Mhala, N. C. (2024). Phishing Webpage Detection using Feature Selection Methods. *Journal of Electrical Systems*, *20*(5s), 447–452. https://doi.org/10.52783/jes.2070

Sirisha, M. L. (2025). Detection of Phishing Website Using Machine Learning. *International Journal of Computer Science and Mobile Computing*, *14*(4), 98–103. https://doi.org/10.47760/ijcsmc.2025.v14i04.008

Sitarz, M. (2023). Extending F1 Metric, Probabilistic Approach. *Advances in Artificial Intelligence and Machine Learning; Research*, *3*(2), 1025–1038. https://doi.org/10.54364/AAIML.2023.1161

Sukmawati, C. E., Pratama, A. R., Hikmayanti, H., & Juwita, A. R. (2025). Performance Optimization of Adaboost and XGBoost Algorithms on Obesity Disease Dataset with Smote Oversampling Technique. *Jurnal Pengembangan IT*, *10*(3), 771–780. https://doi.org/10.30591/jpit.v10i3.8536

Vidhya, N. G., Nirmala, D., & Manju, T. (2023). Quality Challenges in Deep Learning Data Collection in Perspective of Artificial Intelligence. *Journal of Information Technology and Computing*, *4*(1), 46–58. https://doi.org/10.48185/jitc.v4i1.725

White, J., & Power, S. D. (2023). k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation. *Sensors*, *23*(13), 1. https://doi.org/10.3390/s23136077

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and Application of XGBoost in Imbalanced Data. *International Journal of Distributed Sensor Networks*, *18*(6), 1–10. https://doi.org/10.1177/15501329221106935

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

*name of corresponding author