Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Research Article

A Comparative Analysis of P-Value and Mutual Information Feature Selection Methods for Random Forest-Based Phishing Detection

Fahmi Bahtiar Adi Nugroho [a], *Wildanil Ghozi* [b,*], *Fauzi Adi Rafrastara* [c]

[a,b,c] *Universitas Dian Nuswantoro, Jl. Imam Bonjol No.207, Pendrikan Kidul, Kec. Semarang Tengah, Kota Semarang, Jawa Tengah 50131, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 11 November 2025
Revisi Akhir: 18 Desember 2025
Diterbitkan *Online*: 15 Januari 2026

KATA KUNCI

Phishing,
Feature Selection,
P-Value,
Mutual Information,
Random Forest

KORESPONDENSI

E-mail: wildanil.ghozi@dsn.dinus.ac.id*

A B S T R A C T

The application of ANOVA's P-value-based feature selection method, specifically the F-test, in phishing detection using the Random Forest algorithm reveals that a configuration of 25 features yields the fastest inference time, making it suitable for scenarios requiring high computational efficiency and responsiveness. However, if the user's primary priority is to achieve the highest level of detection accuracy, the 29-feature configuration is more feasible because it exhibits higher accuracy performance and better prediction stability. Consequently, there is no definitive trade-off between 25 and 29 features; instead, a selection of solutions can be tailored to the application's requirements. This methodology enables users to achieve an optimal equilibrium between superior performance and minimal inference time in a phishing detection system, contingent upon the implementation context and operational priorities. This study demonstrates that a simple statistical approach, such as the P-value, is not only competitive but also provides superior results compared to more complex methods, offering a practical and efficient solution for real-world implementation.

1. INTRODUCTION

Phishing represents a persistent cyber threat that exploits human vulnerabilities through social engineering methods to obtain personal and financial data. This threat continues to grow rapidly. The latest report from the Anti-Phishing Working Group (APWG) reveals that in the first quarter of 2025, more than one million phishing attacks were identified, marking one of the highest periods of phishing activity ever recorded. [1]. The sheer number of attacks raises the urgent need for more efficient, accurate, and responsive detection methods.

Conventional detection in the context of cybersecurity refers to an approach that predates the adoption of machine learning techniques. This approach emphasizes extracting static attributes from objects such as URLs, page structures, and web content, and then using simple rules, heuristics, or analysis to distinguish between phishing and legit sites [2]. These basic features are carefully selected to be compact, informative, and operational, allowing for quick detection in a production environment. In practical terms, conventional detection is often used as a baseline to assess performance improvements as more advanced ML techniques begin to be implemented. Feature selection, although simpler than fully supervised learning techniques, remains crucial because it enhances accuracy and efficiency without introducing unnecessary complexity.

A Machine Learning-based technique has been developed to combat the ever-evolving threat of phishing, yielding encouraging outcomes. The foundation of effective ML model development is features, which are individual attributes or characteristics extracted from each data sample. In this context, features can be understood as properties of URLs, page structures, and web content that are analyzed in detail across multiple attributes [3]. The quality and relevance of these features directly

https://doi.org/10.25077/TEKNOSI.v11i3.2025.377-386

impact the model's performance, so selecting informative attributes is crucial for improving computational accuracy and efficiency without introducing complexity. However, the use of all attributes can lead to overfitting and increase the computational load. A systematic method known as feature selection is required to determine a subset of the complete array of attributes.

This study specifically investigated the effectiveness of two statistically based feature selection methods. The first is the P-value, a technique that quantifies the statistical significance of the association between a feature and a target variable. This strategy posits that a low P-value signifies a statistically significant association between the feature and the target class, hence deeming it pertinent for inclusion in the model. [4]. The second is Mutual Information (MI), which measures the amount of information a feature provides to a target range. The main advantage of MI is its ability to detect non-linear relationships between variables. In these, not just linear relationships, a high MI value signifies a strong dependency between features and targets. [5]. Although both techniques have been widely used in other domains, their application in phishing detection is still limited.

In comparison, various methods of supervised feature selection have also been successfully implemented [6] and [7]. Principal Component Analysis (PCA) is a dimensionality reduction method that converts original data into a new, uncorrelated set of principal components Other prevalent methods encompass Information Gain, which assesses the significance of a feature by its capacity to diminish entropy OneR, a straightforward algorithm that identifies the most predictive feature and ReliefF, which appraises the quality of a feature based on its proficiency in differentiating between neighboring samples. While these methods have proven effective in improving classification performance, their primary reliance on labeled data is a limitation, given that such data is not always readily available or inexpensive to obtain. Until now, comprehensive evaluations that directly compare the effectiveness of statistical methods, such as P-value and Mutual Information, with various techniques of selecting monitored features are still scarce in the context of phishing site detection.

2. RELATED WORKS

Recent advancements in phishing detection research, utilizing machine learning, have made significant progress, particularly in feature selection to enhance the effectiveness and accuracy of classification models. One of the studies referenced in this study is the [6] that utilizes various feature selection methods, such as PCA, to identify the essential attributes of phishing websites, the research indicates that employing feature selection techniques can significantly enhance classification efficacy, particularly when integrated with algorithms such as Random Forest and Naïve Bayes.

Recent studies on phishing detection using machine learning highlight the importance of aligning feature selection with adaptive model designs. The results of previous studies by [7] Conducted a systematic comparison of six feature selection algorithms in the hyperlink-based dataset category and showed that Information Gain, in combination with Random Forest, was able to achieve superior accuracy in the top 20 features. As in previous studies, this conclusion also emphasizes that not all features contribute significantly to predictive performance, so it is only necessary to consider the most supportive features.

Another comparative study on phishing detection using a few machine learning algorithms also proves that the selection of the correct classification algorithm has a significant impact on the results. [8] Conducted a comparison of five major classification algorithms, such as Logistic Regression, Decision Trees, Random Forest, Adaptive Boosting, and Extreme Gradient Boosting, using the Phishing Websites Kaggle dataset, which contains thousands of URL samples. The results of this study demonstrate that ensemble learning systems, such as Random Forest with voting mechanisms and prediction aggregation from multiple decision trees, provide better results than other algorithms or hybrid approaches.

A separate study on phishing classification, which employed a comparable classification methodology, underscored the importance of meticulous feature selection and multi-algorithmic comparison. In its systematic evaluation, [9] will use the four classification algorithms: Decision Tree, K-Nearest Neighbor, Random Forest, and Support Vector Machine, and focus their analysis on one very complete metric, including True Positive, True Negative, False Positive, and False Negative, to avoid too high FP and FN values that will always be lower than the real world. The research reveals the ranking of multiple evaluation metrics, indicating that uniform performance in classification for the imbalance category is not always possible, as accuracy alone loses meaning.

[10] Construct a URL phishing detection model that is resilient to novel (zero-day) attacks by utilizing a synthesis of attributes derived from contemporary phishing behavior studies. They compared the performance of 11 classification algorithms, including Random Forest, LightGBM, and Gradient Boosting, and found that models that underwent regular retraining performed best. However, the feature selection approach applied emphasizes manual selection based on domain knowledge rather than feature selection methods such as P-value and Mutual Information.

Meanwhile, [11] explicitly highlights the importance of feature selection in URL-based phishing detection. They applied Information Gain (IG) and TreeSHAP techniques to rank and evaluate features using Naïve Bayes, Random Forest, and XGBoost algorithms. The best results are obtained with XGBoost on the top 15 features. The study shows that selective feature trimming can maintain and even improve model accuracy while reducing computational burden.

Lastly, [12] Develop a deep learning system for detecting actual phishing attempts via browser extensions. While their focus is on the use of the RNN-GRU architecture, they also emphasize the importance of prediction efficiency, as well as the use of a minimal number of features for real-time deployment. This research shows that feature reduction through selection or reduction techniques, such as PCA, is crucial for real-world applications.

https://doi.org/10.25077/TEKNOSI.v11i3.2025.377-386

3. METHOD

This research underwent several main stages, as illustrated in Figure 1.

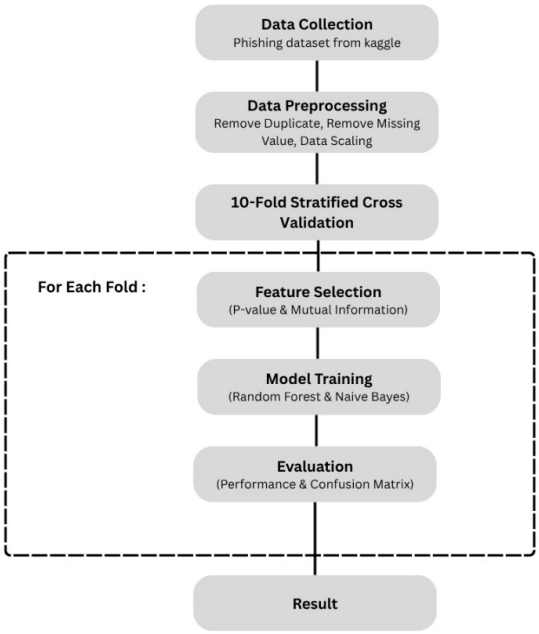

Figure 1. Main stages of research

3.1. Experimental Setup

Machine learning model training is a complex process, particularly when working with large datasets. The roles of hardware and software are crucial in handling the large datasets presented in Table 1, where the ideal combination can improve efficiency, accelerate training, and maximize the performance of the resulting model.

Table 1. Hardware & Software Specification Table

Category	Details
Processor	Intel® Core™ i7-7700HQ CPU @ 2.8 GHz
RAM	16 GB
Graphics Card	Nvidia® GeForce GTX 1050 Ti 4 GB
Storage	NVMe 1 TB
Operating System	TeaLinuxOS
Python	v3.10.10
Jupyter	v2024.2.0

3.2. Dataset Collection

To achieve an objective comparison, this study adopted the same dataset used by [6] namely the "Phishing Website Dataset" which can be found on the Kaggle platform. The main features of this dataset are seen in Table 2:

Table 2. Dataset Phishing Table

Criterion	Details
Number of Samples	11.055 websites
Number of Features	30 Features
Class Distribution	4.898 Phishing (-1 label)
	6.157 legitimate (1 label)
Source	[13] Kaggle

3.3. Dataset Preprocessing

3.3.1. Handling Missing Values

Addressing missing values is a crucial aspect of data preparation that seeks to maintain the quality and integrity of the dataset. This study conducted a thorough analysis of missing values in the Phishing Website Dataset, which comprised 11,055 website samples. The results of the analysis show that the dataset is in a clean condition with no missing values or corrupted data, so no further imputation techniques are required. According to [14]The appropriate management of missing values is essential, as it might influence the validity of statistical analysis outcomes and the efficacy of the constructed machine learning model.

3.3.2. Remove Duplicate

Identifying and removing duplicate data is a crucial step in maintaining data validity and preventing bias in the model learning process. As explained by [15], Duplicate data can introduce inaccuracies and redundancies that could potentially affect the study's conclusions. In this study, the duplicate removal process was implemented to prevent overfitting caused by the low variation in the dataset. The Phishing Website dataset used has been proven to be free of duplicate data after going through the validation process, ensuring that each sample contributes uniquely to the model training process.

3.3.3. *Data Scaling (Standard Scaler)*

The implementation of data scaling using the Standard Scaler is performed before the 10-fold cross-validation process. This standardization technique modifies the numerical data distribution to attain a mean of zero and a standard deviation of one, which is optimal for machine learning algorithms such as Random Forest and Naïve Bayes. According to [15]Data scaling is one of the most essential preprocessing techniques because it can improve model convergence and stability. This approach was chosen to ensure feature scale consistency across cross-validation iterations with the formula:

$$x' = \frac{(x - \mu)}{\sigma} \quad (1)$$

Standardization mathematically, μ and σ It is calculated from the training data for each feature to ensure that each feature possesses a mean of zero and a standard deviation of one after the transformation. On cross-validation, μ and σ calculated separately for each fold, ensuring there is no leakage of information from the test data to the scaling parameters.

3.4. *Feature Selection*

3.4.1. *P-Value*

The feature selection technique utilizing the p-value from the ANOVA F-test is a univariate statistical strategy that assesses the relevance of each feature in isolation from the target variable. The ANOVA F-test calculates the ratio between intergroup variants and in-group variants to generate the F-statistic, which is then converted into a p-value to determine the level of statistical significance. [16] which uses the formula:

$$\mathrm{F} = \frac{Variants\ Between\ Groups}{Variants\ in\ Group} \quad (2)$$

Variance between groups measures how much a feature differs on average between different classes (phishing and legitimate sites). In contrast, variance within a group measures the diversity of feature values within the same class. A high F-value, which indicates that the difference between classes is greater than the variation within the class, is then converted to a P-value. A low P-value signifies a statistically significant association between the feature and the target class; hence, it is deemed pertinent for inclusion in the model. The primary advantage of this method is its computational simplicity, which enables the effective identification of relevant features without the need for complex algorithms.

The main advantage of this method lies in its simplicity of computation, which does not require complex algorithms such as ensembles or dimension reduction, yet remains effective in identifying relevant features. This method has demonstrated efficacy across various study domains, including email spam detection, by employing one-way ANOVA F-test statistics to assess the similarity of pertinent variables [18].

3.4.2. *Mutual Information*

A statistical feature selection technique that measures the association between a feature and the target variable in a dataset, suitable for both classification and regression problems [18]. Unlike the ANOVA F-test, Mutual Information not only identifies linear relationships but can also reveal non-linear relationships between variables, allowing features with complex relationship patterns to be detected as relevant. A higher Mutual Information value for a feature signifies greater informativeness and the ability to enhance the accuracy of predictive models. Mutual Information is calculated using the combined and marginal probabilities of the (X) and targets (Y) with the formula:

$$\mathrm{I(X;Y)} = \left(\sum_{\mathrm{x \in X}} \sum_{\mathrm{y \in Y}} \mathrm{p(x,y)} \cdot \frac{\mathrm{p(x,y)}}{\mathrm{p(x)} \cdot \mathrm{p(y)}} \right) \quad (3)$$

Mutual Information measures the level of information dependence between a feature. (X) and target classes (Y). By comparing the combined probabilities $\mathrm{p}\,(x, y)$ with individual probabilities $\mathrm{p}(x)$ and $\mathrm{p}(y)$Mutual information can determine the informativeness of a feature. A high value of Mutual Information indicates that the X Provides a wealth of relevant information to reduce uncertainty when predicting Y. The main advantage of this method is its ability to detect non-linear relations, allowing even features with complex relationship patterns to be identified as essential attributes.

3.5. *Initial Model Training*

This research employs two primary categorization algorithms: Random Forest and Naïve Bayes. The assessment was conducted to compare the efficacy of the two algorithms to identify the optimal model for detecting phishing websites. The evaluation phase occurs after the pre-processing step and feature selection of the data. This is a concise overview of each algorithm used.

3.5.1. *Random Forest*

Ensemble learning algorithms consisting of several randomly constructed decision trees [19]. Each tree in the Random Forest is trained on a randomly selected data sample using the bootstrap method, and predictions are produced by aggregating the outputs of all trees through voting for classification or averaging for regression. The principal advantage of Random Forest is in its ability to reduce overfitting and improve forecast accuracy by using the strengths of numerous foundational models. [21].

Each tree in the Random Forest also implements random feature selection, where at each node (separation point), only a random subset of the available features is considered to perform the data selection. This differs from the typical decision tree, which considers all features. This random feature selection process ensures that each tree in the forest has a high level of diversity, so that when these trees are combined through voting or averaging, the prediction results become more stable and robust against overfitting. Properly, the final classification prediction of the Random Forest.

$$\hat{y} = mode\{h_1(x), h_2(x), \ldots, h_n(x)\} \quad (4)$$

In simple terms, this formula states that the final prediction $\hat{y}$ It is obtained by taking the mode or value that most often arises from the prediction results of all individual decision trees in the ensemble. Every $\mathrm{h_i}(x)$ Represent the output or prediction of the decision tree i to the input data sample x and by combining predictions from n decision tree that has been trained separately using a different subset of data (bagging), Random Forest uses a voting mechanism to determine the final class. This ensemble approach yields more robust and stable predictions compared to a single decision tree, as it reduces variance and the risk of

 https://doi.org/10.25077/TEKNOSI.v11i3.2025.377-386

overfitting by aggregating various independent models. This voting process enables Random Forest to achieve enhanced classification accuracy, particularly in the context of intricate or noisy datasets.

3.5.2. Naïve Bayes

It is a probabilistic classification technique, remarkably grounded on Bayes' Theorem. This algorithm functions by determining the likelihood of a data point belonging to a specific class based on its feature values [22]. A primary advantage of Naïve Bayes is its simplicity, rapid training speed, and efficacy with massive, complicated datasets, despite relying solely on a basic probabilistic framework. Bayes' Theorem is employed to revise the initial probability of a hypothesis upon acquiring new evidence or data. The subsequent formula articulates this theorem:

$$\mathrm{P}(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \quad (5)$$

This equation explains how to calculate posterior probabilities $\mathrm{P}(C|X)$ the probability of a data X included in the class C after considering the evidence or existing features. The components in the formula consist of P(X|C) is the likelihood (probability of data X given a class C), P(C) is a prior probability (the initial probability of class C before there is evidence, P(X) is evidence (marginal probability of the data X), and $\mathrm{P}(C|X)$ is the final result that represents the posterior probability (class probability) C After considering the data X).

3.6. Performance Validation

3.6.1. 10 Fold-Cross Validation

A fundamental evaluation technique in machine learning research to objectively assess model performance and prevent overfitting. This method involves randomly partitioning the dataset into 10 approximately equal folds, as illustrated in Figure 2. In each iteration, one fold is used as the test set, and the remaining nine folds are used for model training. This procedure is executed 10 times, allowing each data sample to serve as test data once while being utilized for training on 9 occasions. Subsequently, the evaluation findings from all iterations are averaged to yield a more consistent and dependable assessment of the model's performance. According to [22], in his research, which is a fundamental reference in machine learning evaluation, 10-fold cross-validation proved to be the best method for model selection compared to the more computationally expensive leave-one-out cross-validation, even when computing power allows for the use of more folds.

The primary advantage of 10-fold cross-validation lies in its ability to provide more accurate performance estimates than the standard train-test split approach, as it utilizes all the data for evaluation and reduces the bias that may arise from using unrepresentative data. In the context of phishing detection research, the implementation of 10-fold cross-validation ensures that each preprocessing step, such as data standardization using Standard Scaler and P-value-based feature selection, is carried out separately on each fold to avoid data leakage, so that the results of the evaluation truly reflect the model's overview capabilities on data that have never been seen before. Based on a comprehensive study conducted by [7] This method employs a comparable 10-fold cross-validation technique, providing an optimal balance between bias and variance in assessing the model's performance. The findings indicate substantial stability and dependability in evaluating machine learning algorithms for phishing detection.

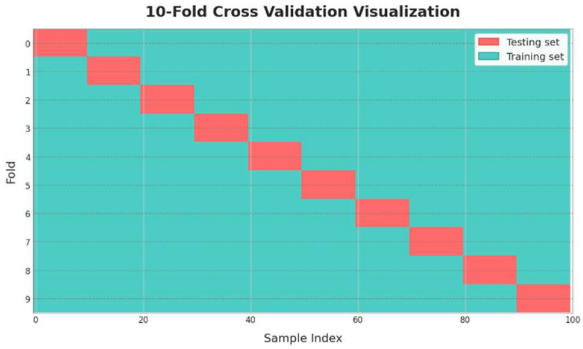

Figure 2. Example of 10-fold cross-validation

3.6.2. Confusion Matrix

One of the essential evaluation metrics in machine learning is the confusion matrix, which provides a comprehensive representation of a classification model's performance by comparing the model's predictions with the actual values in a two-dimensional table. According to [23] The confusion matrix, also referred to as the contingency table, underpins the computation of various prevalent performance metrics. This matrix comprises four primary components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), facilitating a comprehensive evaluation of the classification model's efficacy. As explained by [24] The Confusion Matrix not only provides information about the overall accuracy of the model but also identifies specific error patterns made by the algorithm, thus facilitating more targeted optimization in model development.

In the context of performance validation, the confusion matrix is very effective for identifying weaknesses of classification models because it explicitly visualizes where the model is confused in distinguishing between classes, as highlighted by [23] This matrix enables researchers to understand the trade-off between benefits (true positives) and costs (false positives). The implementation of a confusion matrix in machine learning research has become a standard evaluation metric recommended by various computational libraries, such as scikit-learn. The matrix structure facilitates the calculation of derivative metrics, including precision, recall, F1-score, and specificity, which cannot be obtained through conventional accuracy measurements. The evaluation approach using a confusion matrix also enables objective comparisons between different algorithms in the same dataset, as demonstrated in various studies on phishing detection and medical classification. These metrics offer comprehensive insights into the performance characteristics of the model within each class. Our study leveraged four key performance parameters:

a. Accuracy: This measurement shows a comparison between the correct predictions.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

b. Precision: This measures the comparative accuracy of optimistic predictions.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (7)$$

c. Recall: Evaluation metrics that measure the model's ability to detect positive data correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

d. F1 Score: An evaluation metric that combines Precision and Recall into a single value, by taking the harmonic mean of both.

$$\text{F1-}SSre = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (9)$$

The four standard performance measures of binary ML-based classifiers are:

a. True Positive (TP): The amount of data that is actually positive and predicted positive by the model.
b. False Positive (FP): The amount of data that is actually negative but incorrectly predicted as positive by the model.
c. True Negative (TN) : The amount of data is actually negative and predicted negatively by the model.
d. False Negative (FN): The amount of data that is actually positive but is predicted negatively by the model.

4. RESULT AND DISCUSSION

This study analyzed the effectiveness of ANOVA F-test-based feature selection on 11,055 URLs in the Phishing Website Dataset with a 10-fold validation scheme and two classification algorithms, Random Forest and Naïve Bayes. Feature ranking, based on Figures 3 and 4, consistently places SSLfinal_State and URL_of_Anchor as the two most informative attributes in both the P-Value and Mutual Information graphs. Both achieve the highest score, far surpassing the following features such as Prefix_Suffix, web_traffic, and having_Sub_Domain.

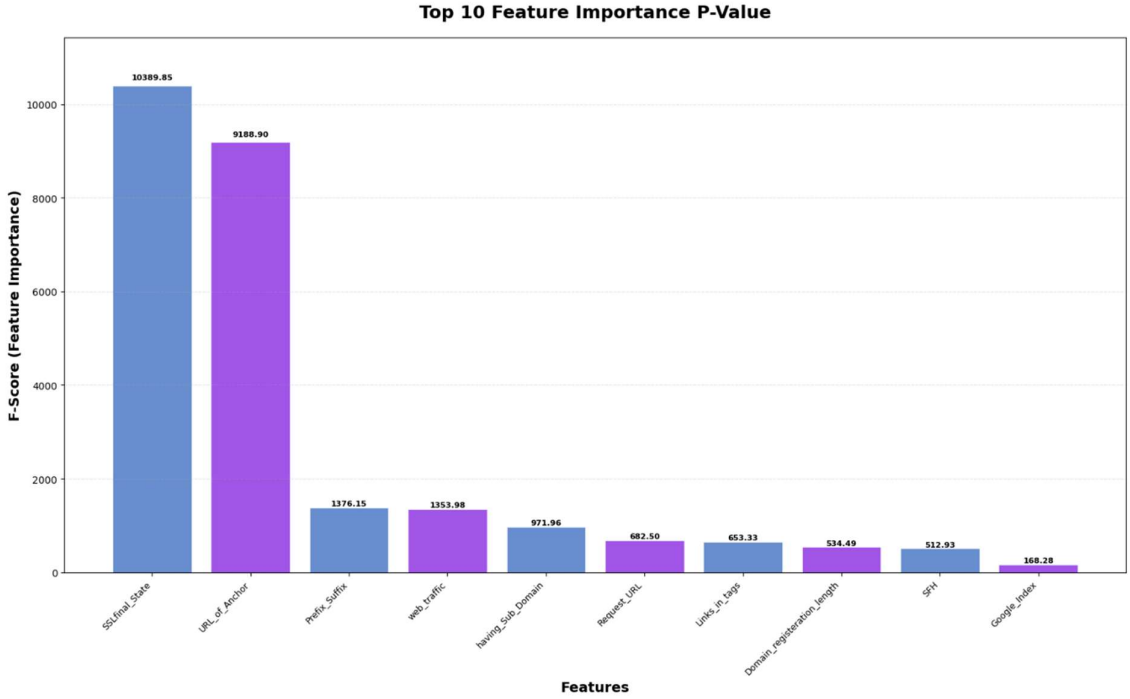

Figure 3. Top 10 p-values of feature importance

 https://doi.org/10.25077/TEKNOSI.v11i3.2025.377-386

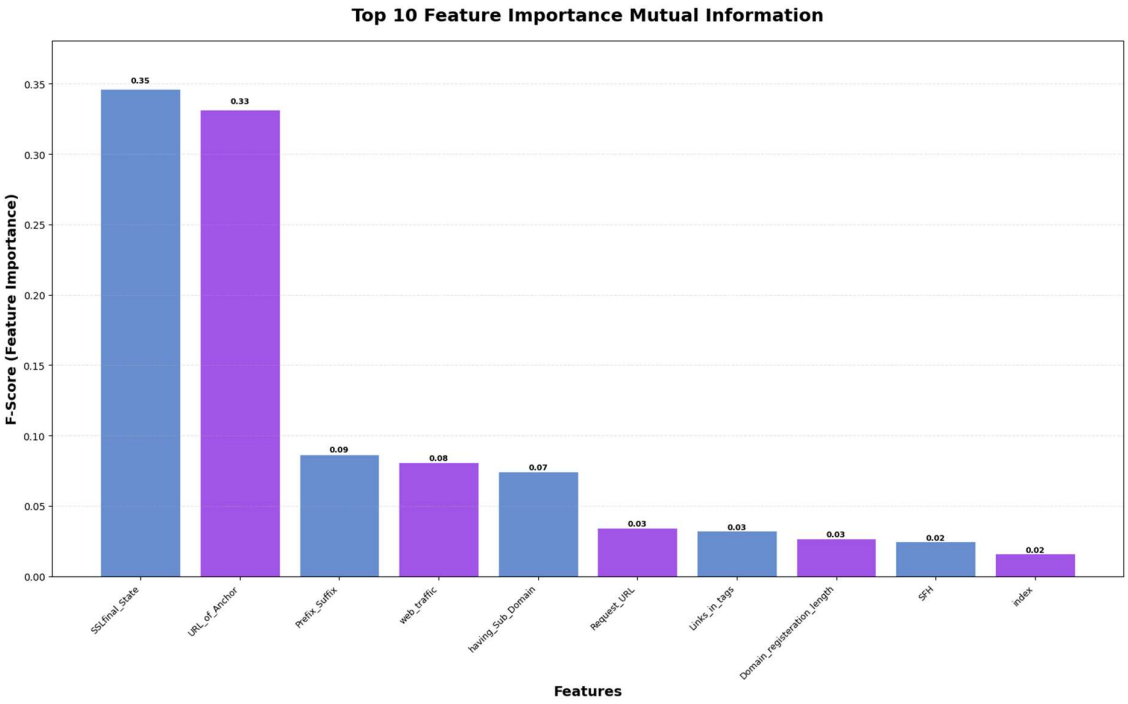

Figure 4. Top 10 Important Mutual Information

An analysis of the accuracy comparison in Figure 5 shows that both feature selection methods can improve classification performance. However, the combination of Random Forest with P-value-based feature selection (RF + P-value) consistently shows significant performance advantages compared to Random Forest with Mutual Information (RF + Mutual Info), especially when the number of features is close to optimal. The RF+P-Value accuracy curve reaches a higher peak and shows better stability. Given this performance superiority, the P-Value method was chosen for a more in-depth analysis. Therefore, further research focused on adjusting the model to optimize the P-value to find the best balance between accuracy and efficiency by comparing the results of using the top 25 and 29 features.

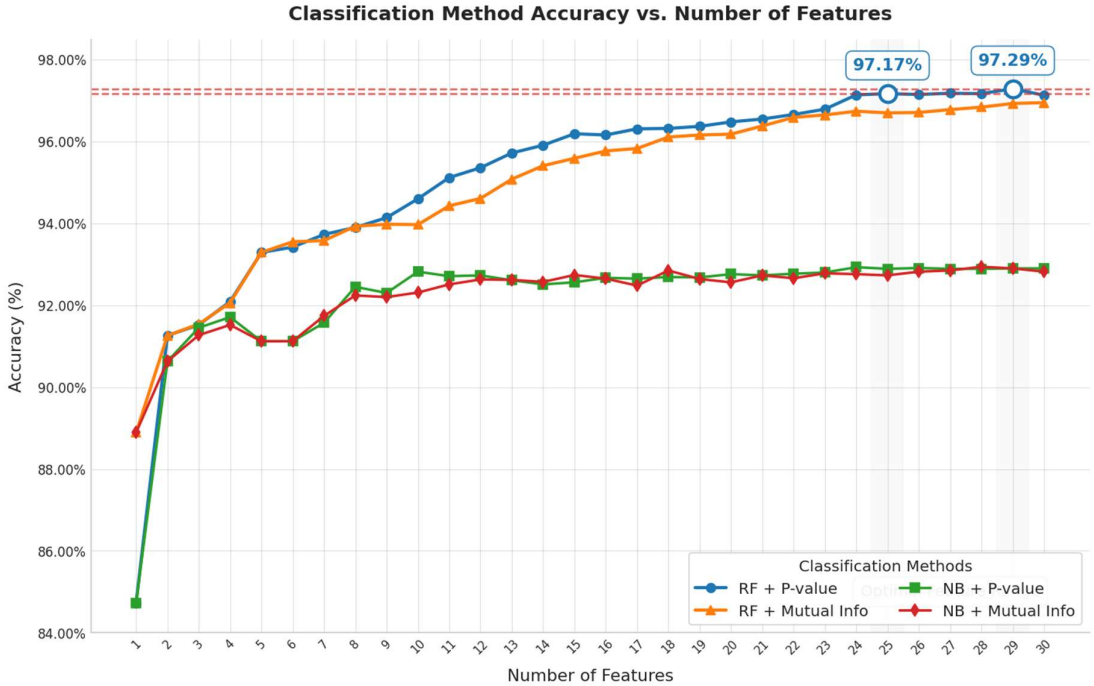

Figure 5. Classification methods, Accuracy, and number of features

Furthermore, the test showed a consistent increase in RF accuracy from 84.73% with one feature to 97.29% when 29 attributes were retained. In comparison, NB remained in the range of 90% to 93% without significant spikes, confirming RF's sensitivity to the

addition of predictive information. Although the 29-feature configuration achieved a peak accuracy of 97.29%, the difference was notable compared to the 25-feature model, which resulted in 97.17%, a very competitive margin. On the other hand, the average inference time test per URL showed that the 25-feature model was executed in 0.029950 seconds, which is faster than the 29-feature model, which required 0.031426 seconds, both of which were tested 5 times in Figure 6. This combination of near-equal accuracy and lower latency confirms the advantages of a 25-feature configuration for real-time phishing detection applications.

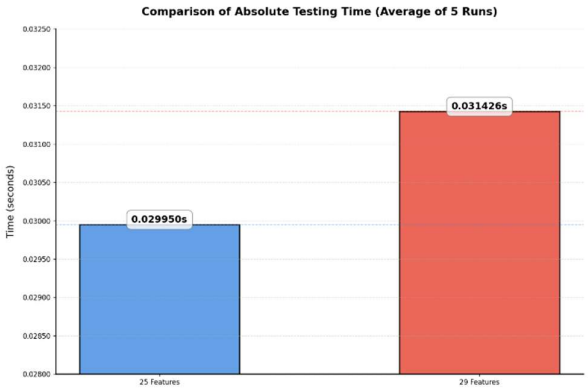

Figure 6. Absolute test time comparison

The Random Forest model optimized with the selection of P-value features in Table 3 demonstrated very consistent performance in both configurations tested, namely 25 and 29 features, with accuracies of 97.17% and 97.29%, respectively. Precision increased from 96.80% on 25 features to 96.98% on 29 features, while the recall risen from 98.16% to 98.20%, which simultaneously hoisted the F1 score from 97.48% to 97.58%. The marginal increase across these metrics indicates that the addition of four extra features strengthens the model's discriminating power, making a configuration of 25 features still reliable if computational efficiency is a priority. In contrast, 29 features are more recommended to maximize detection precision in environments that demand the highest accuracy.

Table 3. Comparison Best Feature

Model+Selection Feature	RF + P-Value	RF + P-Value
Feature	25	29
Accuracy	97,17	97,29
Precision	96,8	96,98
Recall	98,16	98,16
F1-Score	97,48	97,58

The analysis of the Confusion matrix in Figures 7 and 8 shows that the Random Forest (RF) model with 29 features is slightly superior to the one with 25 features. In the 29-feature configuration, the average per-fold results were recorded as 470.90 True Negatives (TN) and 604.60 True Positives (TP), while False Positives (FP) and False Negatives (FN) were 18.90 and 11.10, respectively. In contrast, on 25 features, TN dropped to 469.80 and TP to 604.40, while FP and FN rose slightly to 20.00 and 11.30. This slight difference indicates that the addition of four extra features decreases the number of positive and negative misclassifications, thereby improving the stability of RF predictions slightly on 29 features.

5. COMPARISON WITH OTHER APPROACH

Utilizing the P-value feature selection and Random Forest method, the study achieved a maximum accuracy of 97.29% with 29 characteristics, significantly surpassing several prior studies in phishing detection.

Comparison with similar studies reveals a significant performance improvement, as shown in Table 4. [7] with the Information Gain, OneR, and ReliefF methods achieved 96.1% accuracy. [8] without special feature selection reaches 96.89%. Other research, such as [9] with a Feature Importance of 95.25%, [6] with a PCA of 95.83%. The advantage of the P-value method lies in its simplicity of computation, which does not require complex algorithms such as ensembles or dimension reduction, yet provides superior results. This makes the approach practical for implementing an efficient phishing detection system.

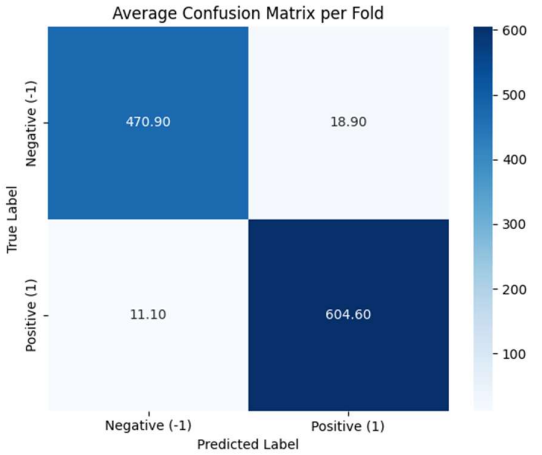

Figure 7. Confusion matrix with 29 features

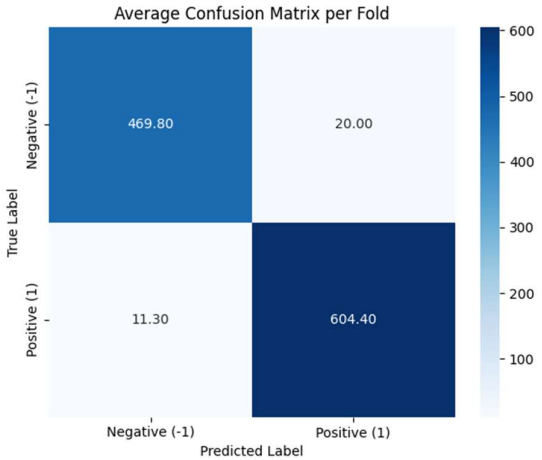

Figure 8. Confusion matrix with 25 features

https://doi.org/10.25077/TEKNOSI.v11i3.2025.377-386

Table 4. Comparison With Other Approach

Research Aspects	Model + Selection Feature	Dataset	Feature	Evaluation Method	Accuracy
Our Study	**RF + P-Value**	**Kaggle**	**29**	**10-Fold Cross Validation**	**97.29%**
	RF + P-Value	**Kaggle**	**25**	**10-Fold Cross Validation**	**97.12%**
[9]	RF + FI	Kaggle	-	Train-test split	95.25%
[8]	XGBoost	Kaggle	-	80-20 Split	96.89%
[7]	RF + IG	Kaggle	30	10-Fold Cross Validation	96.1%
[6]	ANN + PCA	Kaggle	30	80-20 Split	95.07%

6. CONCLUSION

The study's findings indicate that employing the P-value ANOVA F-test for feature selection in phishing detection using the Random Forest algorithm reveals that a configuration of 25 features yields the most rapid inference time, making it suitable for applications requiring high computational efficiency and responsiveness. However, if the user's primary priority is to achieve the highest level of detection accuracy, the 29-feature configuration is more feasible because it exhibits higher accuracy performance and better prediction stability. Thus, there is no absolute trade-off in choosing 25 or 29 features, but a customized solution can be found that suits the application's needs. This approach allows users to achieve the ideal balance between high performance and low inference time in a phishing detection system, depending on the implementation context and desired operational priorities. This is important, considering that in practice, speed of response and the accuracy of detection are crucial aspects that must always be regulated for the coordination of the cybersecurity system. Thus, this study successfully fills the gap in literature by demonstrating that a simple statistical approach, such as the P-value, not only competes but also provides superior results compared to more complex methods, offering a practical and efficient solution for real-world implementation.

This study demonstrates that the P-value is highly effective when combined with Random Forest. Further research can test the effectiveness of this P-Value feature selection on other advanced ensemble algorithms such as XGBoost, LightGBM, or CatBoost, as well as on Deep Learning models.

REFERENCES

[1] "APWG Trends Report Q1 2025", Accessed: Jul. 10, 2025. [Online]. Available: https://apwg.org/trendsreports

[2] W. Bambang Triadi Handaya, "Lukito, Deteksi Website Phishing Menggunakan Teknik Machine Learning 69 Deteksi Website Phishing Menggunakan Teknik Machine Learning."

[3] A. F. Mahmud and S. Wirawan, "Sistemasi: Jurnal Sistem Informasi Deteksi Phishing Website menggunakan Machine Learning Metode Klasifikasi Phishing Website Detection using Machine Learning Classification Method." [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[4] R. Aggrawal and S. Pal, "P-Value Feature Selection Technique for Prediction of Student Performance," 2021. [Online]. Available: www.ijrpr.com

[5] A. L. Young *et al.*, "Mutual information: Measuring nonlinear dependence in longitudinal epidemiological data," *PLoS One*, vol. 18, no. 4 April, Apr. 2023, doi: 10.1371/journal.pone.0284904.

[6] M. A. Daniel, S.-C. Chong, L.-Y. Chong, and K.-K. Wee, "Optimising Phishing Detection: A Comparative Analysis of Machine Learning Methods with Feature Selection," *Journal of Informatics and Web Engineering*, vol. 4, no. 1, pp. 200–212, Feb. 2025, doi: 10.33093/jiwe.2025.4.1.15.

[7] S. N. A. Kamarudin, I. R. A. Hamid, C. F. M. Foozy, and Z. Abdullah, "Feature Selection Approach to Detect Phishing Website Using Machine Learning Algorithm," in *AIP Conference Proceedings*, American Institute of Physics Inc., Nov. 2022. doi: 10.1063/5.0104347.

[8] M. A. Taha, H. D. A. Jabar, and W. K. Mohammed, "A Machine Learning Algorithms for Detecting Phishing Websites: A Comparative Study," *Iraqi Journal for Computer Science and Mathematics*, vol. 5, no. 3, pp. 275–286, 2024, doi: 10.52866/ijcsm.2024.05.03.015.

[9] Selvan K, "Prediction Of Phishing Websites And Analysis Of Various Classification Techniques," *International Journal Of Scientific & Technology Research*, vol. 9, p. 2, 2020, [Online]. Available: www.ijstr.org

[10] A. R. Omar, S. Taie, and M. E. Shaheen, "From Phishing Behavior Analysis and Feature Selection to Enhance Prediction Rate in Phishing Detection." [Online]. Available: https://apwg.org/

[11] L. Mat Rani, C. F. Mohd Foozy, and S. N. B. Mustafa, "Feature Selection to Enhance Phishing Website Detection Based On URL Using Machine Learning Techniques," *Journal of Soft Computing and Data Mining*, vol. 4, no. 1, pp. 30–41, May 2023, doi: 10.30880/jscdm.2023.04.01.003.

[12] L. Tang and Q. H. Mahmoud, "A Deep Learning-Based Framework for Phishing Website Detection," *IEEE Access*, vol. 10, pp. 1509–1521, 2022, doi: 10.1109/ACCESS.2021.3137636.

[13] Akash Kumar, "Phishing website dataset." Accessed: Jul. 19, 2025. [Online]. Available:

https://www.kaggle.com/datasets/akashkr/phishing-website-dataset

[14] S. K. Kwak and J. H. Kim, "Statistical data preparation: Management of missing values and outliers," Aug. 01, 2017, *Korean Society of Anesthesiologists*. doi: 10.4097/kjae.2017.70.4.407.

[15] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, Jun. 2022, doi: 10.1016/j.gltp.2022.04.020.

[16] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[17] N. O. F. Elssied, O. Ibrahim, and A. H. Osman, "A novel feature selection based on one-way ANOVA F-test for e-mail spam classification," *Research Journal of Applied Sciences, Engineering and Technology*, vol. 7, no. 3, pp. 625–638, 2014, doi: 10.19026/rjaset.7.299.

[18] V. Vajrobol, B. B. Gupta, and A. Gaurav, "Mutual information based logistic regression for phishing URL detection," *Cyber Security and Applications*, vol. 2, Jan. 2024, doi: 10.1016/j.csa.2024.100044.

[19] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, Jun. 2024, doi: 10.58496/bjml/2024/007.

[20] M. M. Alani and H. Tawfik, "PhishNot: A Cloud-Based Machine-Learning Approach to Phishing URL Detection," *Computer Networks*, vol. 218, Dec. 2022, doi: 10.1016/j.comnet.2022.109407.

[21] O. Peretz, M. Koren, and O. Koren, "Naive Bayes classifier – An ensemble procedure for recall and precision enrichment," *Eng Appl Artif Intell*, vol. 136, Oct. 2024, doi: 10.1016/j.engappai.2024.108972.

[22] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." [Online]. Available: http//roboticsStanfordedu

[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit Lett*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf Process Manag*, vol. 45, no. 4, pp. 427–437, Jul. 2009, doi: 10.1016/j.ipm.2009.03.002.

AUTHOR'S BIOGRAPHY

Fahmi Bahtiar Adi Nugroho
Is a student in the Informatics Engineering Study Program, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia.

Wildanil Ghozi
A lecturer in the Informatics Engineering Study Program, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia.

Fauzi Adi Rafrastara
A lecturer in the Informatics Engineering Study Program, Faculty of Computer Science, Dian Nuswantoro University, Semarang, Indonesia.