

---

# Journal of Informatics and Web Engineering

Vol. 4 No. 1(February 2025)

eISSN: 2821-370

---

## Optimising Phishing Detection: A Comparative Analysis of Machine Learning Methods with Feature Selection

**Mohamad Asraf Daniel<sup>1</sup>, Siew-Chin Chong<sup>2\*</sup>, Lee-Ying Chong<sup>3</sup>, Kuok-Kwee Wee<sup>4</sup>**

<sup>1,2,3,4</sup> Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Melaka, Malaysia

\*corresponding author: (chong.siew.chin@mmu.edu.my; ORCID: 0000-0003-0421-4367)

*Abstract* - Phishing is an act of cybersecurity attack that tricks people into sharing sensitive data. Due to the inefficiency of the current security technologies, researchers have been paying much attention to employing machine learning methods for phishing detection lately. In our proposed solution, the effectiveness of machine learning techniques with feature selection techniques for phishing detection is investigated. To be specific, Random Forest (RF) and Artificial Neural Network (ANN) are integrated with feature selection techniques, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). The goal was to identify and classify the model with the highest accuracy. The experiments were evaluated using a dataset of 4,898 phishing sites and 6,157 legitimate sites, with the phishing data sourced from Kaggle.com. Our experiments demonstrate that the combination of RF model with PCA achieved 95.83% accuracy, while the ANN model with PCA reached 95.07% accuracy. The incorporation of PCA and RFE not only optimised the models' predictive performance but also improved computational efficiency. Overfitting can also be reduced. The experimental results also demonstrate that the proposed ANN with PCA method outperforms the state-of-the-art methods. Consequently, this research highlights the potential of combining advanced feature selection techniques with machine learning algorithms to develop robust solutions for phishing detection. Yet, this undoubtedly contributes to a safer internet environment.

*Keywords*— Machine Learning, Phishing Detection, Feature Selection, Dimension Reduction, Cyber-attacks.

*Received: 15 August 2024; Accepted: 24 December 2024; Published: 16 February 2025*

*This is an open access article under the [CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/) license.*



---

### 1. INTRODUCTION

In recent years, the growth of electronic trading has led to a significant increase in cyber-attacks. Many strategies have been prompted to overcome these challenges [1],[2],[3]. Among the challenges, phishing attacks are serious concerns for individuals and organizations. These attacks cause the loss of sensitive data such as user credentials and finance information. Phishing attacks often involve the use of fraudulent websites which are hard to detect by the users easily. Therefore, detecting phishing websites has become a crucial task for cybersecurity professionals.

One way to address the issue is by using machine learning (ML) techniques. ML algorithms can be trained to automatically identify phishing websites through analyzing various features of the website, such as the domain name, content, and layout. However, ML models [4], [5], [6] may struggle to generalize across new, unseen phishing tactics that differ significantly from the training data. This often leads to potential false negatives. Our research work aims to investigate the use of ML for detecting phishing websites and to propose an optimal model to overcome the weaknesses. The proposed work includes collecting a dataset of websites, extracting features, developing ML algorithms, and evaluating the performance of the model. The expected outcome is an automated approach for detecting phishing websites with higher accuracy.

This research introduces the implementation of machine learning techniques, specifically RF and ANN algorithms, to be combined with PCA and RFE. RF is chosen due to its robustness, ability to handle non-linearity, and feature importance insights. For ANN, it is selected for its complex pattern recognition capabilities and adaptability. PCA and RFE are employed for dimensionality reduction and feature selection. These methods collectively improve phishing detection by enhancing feature extraction, increasing accuracy and precision, and ensuring real-time detection capabilities.

## 2. LITERATURE REVIEW

Machine learning algorithms have been proven to work well in phishing websites detection. The works of Dutta [1] investigated the use of machine-learning approaches in identifying phishing sites. For instance, a framework was developed using recurrent neural network to classify URLs and identify which ones are phishing. From this research, the authors demonstrated that their works are more efficient compared to the conventional phishing detection systems. Their research findings stated that RF generates the highest accuracy of 97.40%.

Yahya et al. [7] applied three machine learning algorithms for the identification of phishing websites which are K-Nearest Neighbour (KNN), RF and Decision Tree. They used a data set of 11,055 observations with 15 variables and they applied “result” as the independent variable. The authors concluded that RF attained the highest accuracy rate of 98.5% among the rest.

In the study by Fan et al. [8] their works compared five machine learning approaches for spotting phishing websites. The machines are decision tree, RF regression, logistic regression, support vector machine (SVM), as well as K-Nearest Neighbour (KNN). In the next step, they extract the most important features from those URLs of phishing websites as well as legitimate websites. After that, these features are used to train and test the machine learning models. Lastly, an evaluation of every model is conducted based on accuracy, precision, recall, and F1 score. The RF performed the best with accuracy of 97.0%.

In the research outlined in Ramireddi et al. [9] the authors used various machine learning algorithms to study the characteristics of phishing URL detection. Like SVM, RF Classifier, Logistic Regression and Naïve Bayes classifier. The RF again outperformed the others, as evidenced by its accuracy of 98.5 and F1 score of 0.986 respectively.

On the other hand, Mandalik et al. [10] used RF and Decision Tree algorithms to develop machine learning models. The trained models achieved satisfactory results, with RF achieving 87.0% accuracy and Decision Tree achieving 82.4% accuracy.

Zamir et al. [11] conducted a comprehensive evaluation of various algorithms for phishing URL detection and classification. The study found that the RF technique is crucial in eliminating meaningless features. In their work, the Stacking approach had been applied to combine RF, Neural Network, and Bagging techniques. The experiment results exhibited outstanding performance of 97.4% accuracy in phishing URL classification.

Alnemari et al. [12] emphasizes the significance of using only URL attributes in identifying phishing websites. The authors applied datasets obtained from both the Kaggle and Phishtank websites. A hybrid approach had been introduced by combining PCA with RF and SVM to reduce the dataset’s dimensionality. The model helps to retain all the important data. This approach scored accuracy of 96.8% compared to other benchmarked techniques.

Zhu et al. [13] utilizes an ANN for detecting phishing websites. The model featured 17 input neurons corresponding to 17 features, a single hidden layer, and two output neurons. The model achieved an accuracy of 92.48% on the training and testing subsets.

### 3. RESEARCH METHODOLOGY

The overall research design of this project involves implementing a solution through machine learning algorithm for the phishing websites detection. The choice of applying machine learning is motivated by the rise of reported phishing attacks and the limitations of conventional rule-based methods. Machine learning offers the advantage of adaptability and the ability to learn patterns from data. This is needed for handling dynamic and evolving phishing tactics.

The proposed research framework, as illustrated in Figure 1, involves several key methods to enhance the detection of phishing websites. Initially, a dataset of phishing websites will be collected and preprocessed to ensure data quality and relevance. The data is split into train and test sets. Two approaches will be adopted: one incorporating feature selection processes and one without. For feature selection, PCA and RFE will be used to reduce dimensionality. Only the most informative features will be selected. The selected features will subsequently be utilized to train machine learning models, specifically RF and ANN. The performances of the models, both with and without feature selection, will be compared and analyzed to determine the most effective approach. The results and conclusions will provide insights into the efficiency of detecting phishing websites.

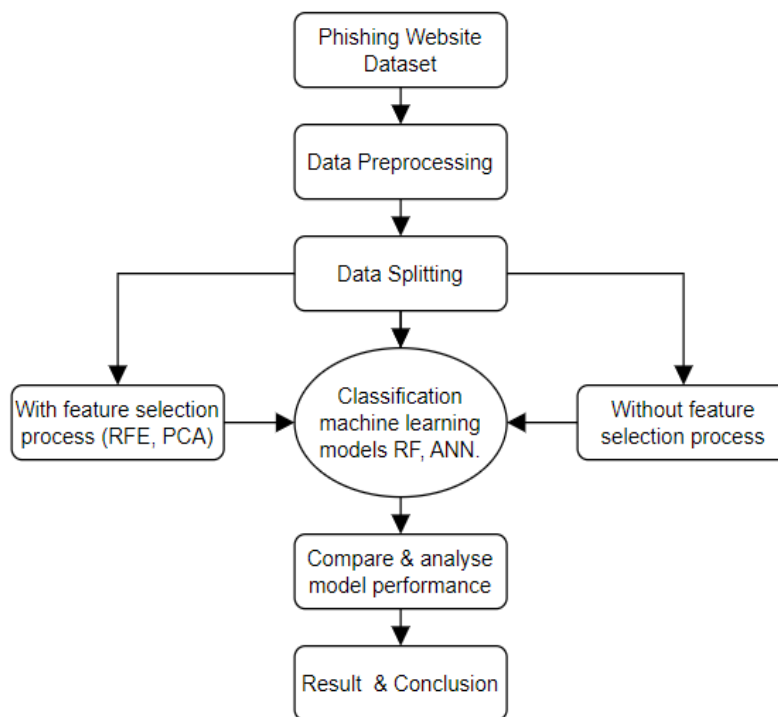


Figure 1. Research Framework

#### 3.1 Dataset

The experiments were evaluated using a dataset of 4,898 phishing sites and 6,157 legitimate sites, with the phishing data sourced from Kaggle.com [14]. The dataset contains a total of 11,005 records with 20 features. Table 1 lists the characteristics of the phishing website. The amount of legitimate and phishing websites is displayed in Table 2.

It is essential to select the most relevant features and the most influential subset of features from a dataset. Relevant features are those that wield significant influence on classification outcomes, which can provide informative insights to bolster the predictive capabilities of the model. Among them, high-impact features show a strong correlation with the target variable and furnish substantial information to refine the model's predictive accuracy. For instance, in the realm of phishing detection, features such as URL length, the presence of special characters, SSL certificates, domain age, and PageRank emerge as noteworthy high-impact factors. In contrast, low-impact features exert minimal to negligible effect on classification outcomes and may even introduce noise into the model. Examples of such features

in phishing detection include HTML comments, the number of images on a page, font size variations, and background colour. To optimize model performance, the feature selection process involves identifying and retaining high-impact features while discarding low-impact ones. Techniques like PCA, RFE, and feature importance from RF aid in this endeavour. They allow the prioritization of features based on their impact on model performance. By homing in on high-impact features, machine learning models become more adept at accurate classification. Hence their efficiency and effectiveness in tasks like phishing website detection can be enhanced.

Table 1. Phishing Website Dataset

<b>Feature</b>	<b>Description</b>
<b>index</b>	A unique identifier or index for each entry in the dataset.
<b>having_IP_Address</b>	Indicates whether the URL uses an IP address.
<b>URL_Length</b>	The overall character count in the URL.
<b>Shortning_Service</b>	Specifies whether a URL shortening service (like bit.ly) is used.
<b>having_At_Symbol</b>	Indicates the presence of the "@" symbol in the URL, which can be used in phishing URLs.
<b>double_slash_redirecting</b>	Indicates if the URL has "/" after the protocol (http/https), which can be used to redirect URLs.
<b>Prefix_Suffix</b>	Indicates the presence of a prefix or suffix separated by a hyphen in the domain.
<b>SSLfinal_State</b>	Indicates the presence and state of an SSL certificate.
<b>Domain_registration_length</b>	The length of time the domain is registered for, typically measured in years.
<b>popUpWidnow</b>	Indicates whether the page generates pop-up windows.
<b>Iframe</b>	Indicates whether the page uses iframes, which can be used to embed content from other sites.
<b>age_of_domain</b>	The age of the domain, typically measured in years since it was first registered.
<b>DNSRecord</b>	Indicates the presence of a valid DNS record for the domain.
<b>Links_pointing_to_page</b>	The total number of external links directed to the page.
<b>web_traffic</b>	An estimate of the web traffic to the domain, possibly measured using metrics like Alexa rank or similar.
<b>Page_Rank</b>	The PageRank of the URL, a measure of its importance or quality as determined by Google's algorithm.
<b>Google_Index</b>	Indicates whether the URL or domain is indexed by Google.
<b>Result</b>	The final classification or status of the URL
<b>Statistical_report</b>	Data from statistical reports, possibly including metrics on user visits, bounce rates, etc.

Table 2. Phishing Website Dataset

<b>Website</b>	<b>Count</b>
<b>Legitimate</b>	6157
<b>Phishing</b>	4898

### 3.2 Data Preprocessing

Data preprocessing steps are undertaken to ensure the quality and effectiveness of the machine learning models. This includes handling missing data, removing duplicate entries, and standardizing features. The dataset undergoes normalization and scaling to bring all features to a consistent range and optimize the learning process.

### 3.3 Data Splitting

Data splitting is performed on the dataset to form training and testing sets. A standard split ratio of 80:20 and 60:40 is employed to maintain a balance between training and evaluation. The training dataset consists of a diverse collection of labelled examples, including both legitimate and phishing websites. This dataset aims to capture the various characteristics and patterns associated with phishing attacks. The testing dataset is arranged to assess the performance of model generalization on unknown data. Result is assigned based on the website's legitimacy. '-1' is used to denote phishing websites, while '1' is used to denote legitimate websites.

### 3.4 Features Selection

The proposed solution by combining machine learning algorithms with feature selection techniques can significantly enhance the effectiveness and efficiency in phishing website detection. High-dimensional datasets pose challenges such as overfitting, increased computational complexity, and reduced model interpretability due to the curse of dimensionality [16]. PCA addresses these issues by extracting the essential information and transforming the original features into a lower-dimensional space. Similarly, RFE iteratively removes the least important features and selects a meaningful subset. By reducing feature redundancy and focusing on informative features, PCA and RFE help to improve model generalization and predictive accuracy. Furthermore, the reduced feature space leads to computational efficiency, as models trained on fewer features require less computational resources and exhibit faster convergence during training. Additionally, the simplified feature representation enhances model interpretability. The underlying relationships between features and target variables can be observed easily. By acting as regularization techniques, PCA and RFE also mitigate overfitting, and provide better generalization performance on unseen data. In a nutshell, incorporating PCA and RFE with RF and ANN models facilitates better solutions for handling various phishing attacks.

### 3.5 Principal Component Analysis (PCA)

PCA is a popular technique in machine learning process [17] to reduce the feature dimension of a dataset. The main objective is to transform a dataset with potentially correlated variables. This transformation creates a new set of uncorrelated variables, called principal components. These components are linear combinations of the original variables. They are ordered based on how much variance they explain in the data.

### 3.6 Recursive Feature Elimination (RFE)

This is a feature selection technique designed to select the important features in a dataset. RFE aims to enhance the model's performance by eliminating less significant features, thereby reducing overfitting. It has been adopted in phishing detection [18].

### 3.7 Artificial Neural Network (ANN)

ANNs are robust machine learning techniques inspired by the structure and functionality of the human brain. They consist of layers of interconnected nodes that process input data to produce an output [21]. Neural networks robust in capturing complex patterns and relationships within datasets. This is helpful in identifying phishing activities.

### 3.8 Random Forest (RF)

RF [20] is a famous approach in classification and regression tasks. It utilizes the decision trees during training and combines the outputs. Then the final decision can be made. For classification tasks, it outputs the mode of the classes predicted by the individual trees. For regression tasks, it outputs the mean prediction of the individual trees. RF is particularly effective in phishing detection due to its ability to harness multiple decision trees [15], which enhances accuracy and robustness in identifying phishing activities.

### 3.9 Model Evaluation

Evaluation is performed using the testing datasets based on the trained models. The performance metrics involved in assessing the performance are accuracy, recall, precision and F1 score, as shown in Equations (1), (2), (3) and (4) respectively. Model hyperparameters are fine-tuned through validation processes, including techniques like cross-validation.

$$Acc = \frac{TP + TN}{Total} \quad (1)$$

$$Rec = \frac{TP}{TP + FN} \quad (2)$$

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

$$F1 = 2 * \frac{Prec. Rec}{Pre + Rec} \quad (4)$$

True Positives (TP) are the number of correctly classified positive samples. False Positives (FP) are the number of incorrectly classified as positive samples. The samples which are incorrectly classified as negative are named as False Negatives (FN). For the samples that are correctly identified as negative is known as True Negatives (TN).

## 4. EXPERIMENTAL RESULTS

The computer system used in running the experiments is powered with an Intel Core i7-9750H processor with 6 cores and 12 threads, and 16 GB of memory, expandable up to 128 GB DDR4. The Google Colaboratory application is used along with various Python libraries for data processing and building machine learning models. 80% of the Phishing Website dataset is used for training and 20% is used for testing.

### 4.1 Experiment Result Analysis

In the experiments, two tests are conducted. The first test involves model classification without using feature selection techniques. While running the second test, feature selection techniques such as RFE and PCA are employed. Both outputs are evaluated to find the optimal algorithm combination with higher performance. Table 3 presents the classification result without feature selection in testing model. RandomizedSearchCV is used for hyperparameter tuning through various experiment settings and identify the optimal combination that maximizes the classifier's performance.

In the second experiment, RFE and PCA feature selection techniques are employed. Table 4 shows the classification result with feature selection. RF+PCA (80/20) method achieved the highest accuracy at 95.83%, precision at 95.00%, recall at 96.97%, and F1-score at 95.97%. It combines the RF classifier with PCA feature selection. RF+RFE (80/20) achieved slightly lower performance than RF+PCA, with an accuracy of 94.97% and other metrics in a similar range. ANN+PCA (80/20) method achieved an accuracy of 95.07% and ANN+RFE (80/20) method achieved an accuracy of 94.84%. (60/40) method performed well but not as effectively as the 80/20 split, the 60/40 split resulted in slightly lower accuracy compared to the 80/20 split. In addition, by comparing both the first and second experiments, it can

be noted that ANN performs well when combining with the feature selection method, be it PCA or RFE. This simplification of the input data allows the ANN to focus on the most significant features, improving its ability to learn and generalize, which leads to better performance.

Table 3. Classification Result Without Feature Selection

Method	Accuracy	Precision	Recall	F1-Score
<b>RF (80/20)</b>	96.78%	96.17%	98.24%	97.20%
<b>RF (60/40)</b>	96.58%	96.35%	97.69%	97.01%
<b>ANN (80/20)</b>	92.62%	92.00%	95.29%	93.62%
<b>ANN (60/40)</b>	87.72%	83.91%	97.01%	89.99%

Table 4. Classification Result with Feature Selection

Method	Accuracy	Precision	Recall	F1-Score
<b>RF+PCA (80/20)</b>	95.38%	95.00%	96.97%	95.97%
<b>RF+RFE (80/20)</b>	94.97%	95.03%	96.17%	95.60%
<b>RF+PCA (60/40)</b>	93.12%	93.64%	94.31%	93.98%
<b>RF+RFE (60/40)</b>	94.73%	95.31%	95.42%	95.37%
<b>ANN+PCA (80/20)</b>	95.07%	95.26%	96.09%	95.67%
<b>ANN+RFE (80/20)</b>	94.84%	94.84%	94.84%	94.84%
<b>ANN+PCA (60/40)</b>	94.61%	95.41%	95.11%	95.26%
<b>ANN+RFE (60/40)</b>	93.66%	93.66%	93.66%	93.65%

Table 5 presents a comprehensive analysis of the hyperparameters utilized in various machine learning classifiers employed for phishing detection. Hyperparameters are crucial as they govern the training process and influence the model performance. The table outlines the specific hyperparameters for each algorithm, including their values and the rationale behind their selection.

Table 5. Performance of Selected ML Classifiers With Different Hyperparameter Settings

ML Classifiers	Hyperparameter Settings	Accuracy	Precision	Recall	F1-Score
<b>RF (80/20)</b>	min_samples_leaf = 1; min_samples_split = 2;	96.65%	96.02%	98.16%	97.08%
<b>RF (60/40)</b>	n_estimators = 200; max_depth = 30;	96.58%	96.35%	97.69%	97.01%

<b>ANN (80/20)</b>	hidden_layer_sizes = (10, 10, 10); Solver = adam; Activation = relu;	90.41%	87.59%	96.81%	91.97%
<b>ANN (60/40)</b>	Alpha = 0.01; learning_rate = adaptive.	90.72%	94.43%	88.95%	91.60%

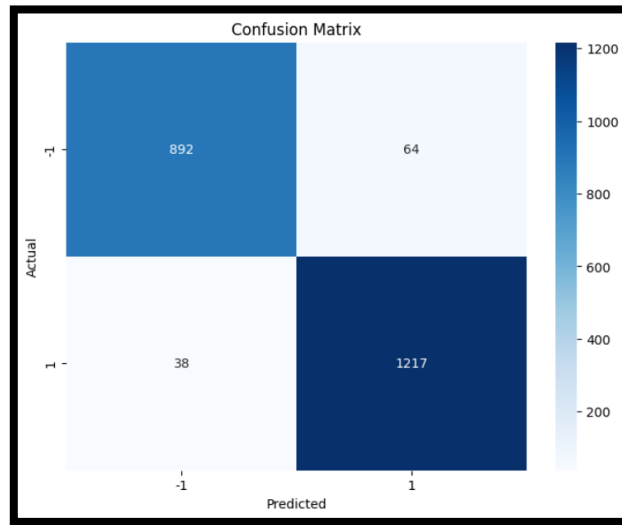


Figure 2. RF (80/20) Confusion Matrix Without Selection Feature

In Figure 3, this matrix represents the performance of the same RF classifier, but with feature selection using Principal Component Analysis (PCA). Comparing this matrix to Figure 2, the study can observe how feature selection impacts the model’s classification results. The diagonal cells (TN and TP) indicate correct class classifications, while off-diagonal cells (FP and FN) denote incorrect classifications.

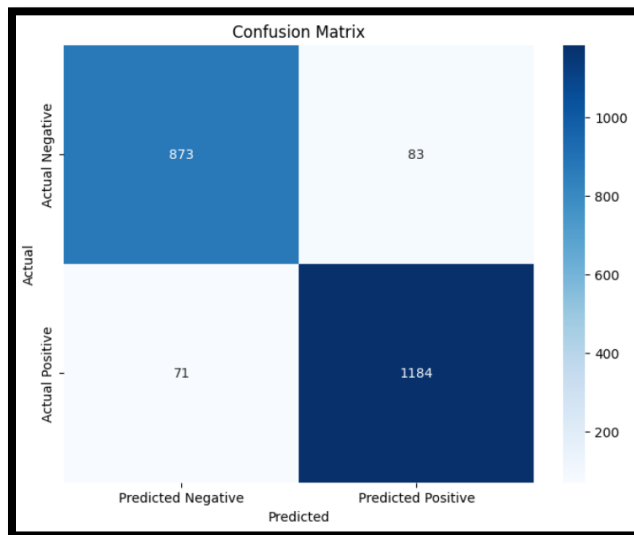


Figure 3. RF (80/20) Confusion Matrix With Principal Component Analysis



In Figure 4, This matrix corresponds to an ANN classifier without feature selection. Like the RF matrix, it shows the counts of TN, FP, FN, and TP. Finally, in Figure 5, this matrix represents the ANN classifier with PCA-based feature selection. By comparing this matrix to the previous one, the impact of feature selection on the ANN's performance can be evaluated. The study can analyze the model's performance based on these counts.

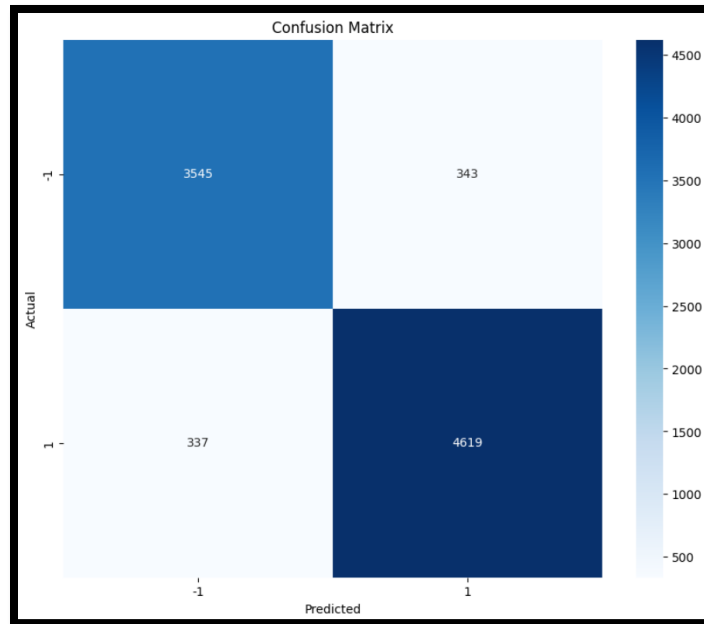


Figure 4. ANN (80/20) Confusion Matrix Without Selection Feature

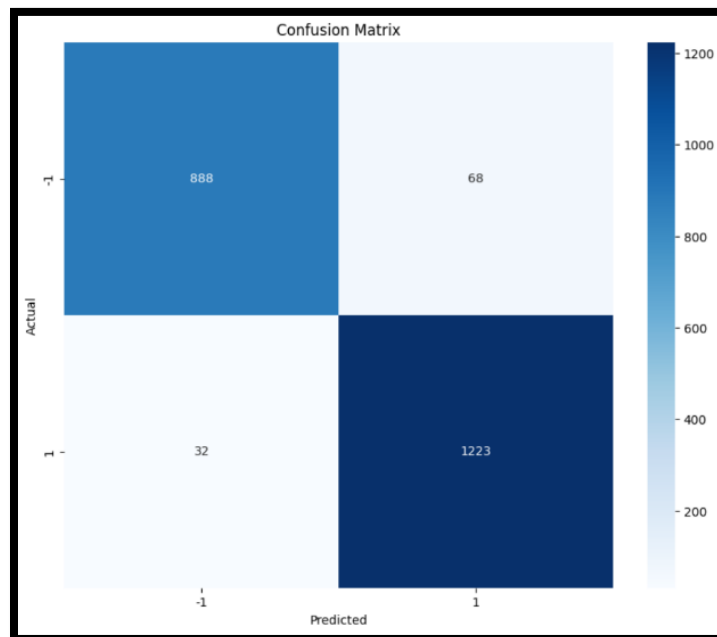


Figure 5. ANN (80/20) Confusion Matrix With Principal Component Analysis

#### 4.2 Comparative Analysis with Existing Result

A comprehensive comparative analysis was conducted to compare the proposed model with the state-of-the-art methods. Despite extensive experimentation, the test of RF model did not surpass the highest reported accuracy of 98.5% achieved by other researchers. The RF implementation achieved a satisfactory but not superior accuracy, highlighting the challenges and limitations of this approach for phishing detection in the dataset. Table 6 shows the comparison of the RF Model with the existing work.

Table 6. Comparison of The RF Model With The Existing Work

<b>Model</b>	<b>Accuracy</b>
<b>RF [9]</b>	98.50%
<b>Proposed RF</b>	96.78%

In contrast, the result of the ANN model demonstrated a significant improvement over the previous study. A 2021 study [19] reported a 92.48% accuracy using ANN for phishing detection. By optimizing the neural network architecture and incorporating PCA for feature reduction, our proposed work achieved a notable accuracy of 95.07%. Table 7 shows the result of the ANN+PCA model compared to the existing work,

Table 7. Comparison of The ANN+PCA Model With The Existing Work

<b>Model</b>	<b>Accuracy</b>
<b>ANN [19]</b>	94.50%
<b>Proposed ANN+PCA</b>	95.07%

## 5. CONCLUSION

The proposed solution in this paper has yielded noteworthy findings in the domain of website phishing detection. The implemented system, leveraging machine learning models, demonstrates a high level of accuracy in distinguishing phishing websites as opposed to legitimate ones. The machine learning models exhibit a outstanding ability to adapt to evolving phishing tactics, showcasing the system's resilience against dynamic and emerging threats. The inclusion of PCA and RFE as the feature selection techniques has significantly enhanced the overall performance. The proposed model is carefully designed and tested with a dataset comprising of legitimate and phishing websites. The main objective of our performance evaluation is to improve the accuracy of classification models. Compared with the state-of-the-art methods, the RF model test result did not excel the others. However, the experiment results show satisfactory result on the ANN+PCA (80/20) at accuracy of 95.07%. This demonstrates the potential of combining machine learning methods with feature selection for enhancing phishing website detection. In the future works, the proposed solution can be experimented with more datasets of phishing attacks. Other domains such as deepfake detection or fraud detection can be considered with this proposal. This promises a more realistic and applicable direction towards cybersecurity attacks.

## ACKNOWLEDGEMENT

We thank the anonymous reviewers for the careful review of this article.

## FUNDING STATEMENT

This research was not funded by any grant.

## AUTHOR CONTRIBUTIONS

Mohamad Asraf Daniel: Conceptualization, Methodology, Validation, Writing – Original Draft Preparation;  
Siew-Chin Chong: Project Administration, Supervision, Writing – Review & Editing;  
Lee-Ying Chong: Project Administration, Writing – Review & Editing;  
Kuok-Kwee Wee: Data Curation, Methodology, Review.

## CONFLICT OF INTERESTS

No conflict of interests was disclosed.

## ETHICS STATEMENTS


The paper follows The Committee of Publication Ethics (COPE) guideline.

## REFERENCES

- [1] A. K. Dutta, “Phishing website detection by machine learning techniques,” *PloS One*, vol. 16, no. 10, p. e0258361, 2021, doi: 10.1371/journal.pone.0258361.
- [2] W.-H. Chong, S.-C. Chong, and L.-Y. Chong, “The assistance of eye blink detection for two-factor authentication,” *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 111–121, 2023, doi: 10.33093/jiwe.2023.2.2.8.
- [3] Y. J. Chew, S. Y. Ooi, K. S. Wong, Y. H. Pang, and S. O. Hwang, “Evaluation of black-marker and bilateral classification with J48 decision tree in anomaly-based intrusion detection system,” *Journal of Intelligent and Fuzzy Systems*, vol. 35, no. 6, pp. 5927–5937, 2018, doi: 10.3233/JIFS-169834.
- [4] F. Salahdine, Z. El Mirabet, and N. Kaabouch, “Phishing attacks detection: A machine learning-based approach,” *International Journal of Computer Science and Information Technology*, vol. 9, no. 3, pp. 1–8, 2022, doi: 10.48550/arXiv.2201.10752.
- [5] L. Torrealba Aravena, P. Casas, J. Bustos-Jiménez, G. Capdehourat, and M. Findrik, “Phish Me If You Can—Lexicographic analysis and machine learning for phishing websites detection with PHISHWEB,” in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, 2023, pp. 1–6, doi: 10.1109/NetSoft57336.2023.10175503.
- [6] A. Alswailem, B. Abdullah, and N. Almamary, “Detecting phishing websites using machine learning,” in *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, 2019, pp. 1–6, doi: 10.1109/CAIS.2019.8769571.
- [7] F. Yahya, M. B. Anai, R. I. W. Mahibol, S. A. Frankie, C. K. Ying, R. G. Utomo, and E. L. N. Wei, “Detection of phishing websites using machine learning approaches,” in *2021 International Conference on Data Science and Its Applications (ICoDSA)*, 2021, doi: 10.1109/ICoDSA53588.2021.9617482.
- [8] Z. Fan, “A joint feature selection and integrated learning algorithm for phishing website detection,” in *2021 International Conference on Applied Machine Learning (ICAML)*, 2021, pp. 1–6, doi: 10.1109/ICAML54311.2021.00018.
- [9] S. Ramireddi, T. N. Pandey, and V. A. Woonna, “Classification of phishing websites using machine learning models,” *International Journal of Computer Science and Information Technology*, vol. 9, no. 3, pp. 1–8, 2023, doi: 10.1109/AISP57993.2023.10134944.
- [10] A. Mandalik, R. Sankararajan, V. V. Raveendran, and P. K. Sivakumar, “Phishing website detection using machine learning,” in *2022 IEEE International Conference for Convergence in Technology (ICT)*, 2022, pp. 1–6, doi: 10.1109/I2CT54291.2022.9824801.

- [11] A. Zamir, H. U. Khan, T. Iqbal, N. Yousaf, F. Aslam, A. Anjum, and M. Hamdani, "Phishing website detection using diverse machine learning algorithms," *The Electronic Library*, vol. 38, no. 1, pp. 65–80, 2020, doi: 10.1108/EL-05-2019-0118.
- [12] S. Alnemari and M. Alshammari, "Detecting phishing domains using machine learning," *Applied Sciences*, vol. 13, no. 8, p. 4649, 2023, doi: 10.3390/app13084649.
- [13] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, "DFOB-ANN: An artificial neural network phishing detection model based on decision tree and optimal features," *Applied Soft Computing*, vol. 95, p. 106505, 2020, doi: 10.1016/j.asoc.2020.106505.
- [14] S. Akashkr, "Phishing website dataset," Kaggle, 2023. [Online]. Available: [www.kaggle.com/datasets/akashkr/phishing-website-dataset](http://www.kaggle.com/datasets/akashkr/phishing-website-dataset).
- [15] X. Yang, L. Yan, B. Yang, and Y.-F. Li, "Phishing website detection using C4.5 decision tree," in *DEStech Transactions on Computer Science and Engineering*, 2017, doi: 10.12783/dtcse/itme2017/7975.
- [16] Q. Zhang, "Practical thinking on neural network phishing website detection research based on decision tree and optimal feature selection," in *Journal of Physics: Conference Series*, vol. 2031, no. 1, p. 012062, 2021, doi: 10.1088/1742-6596/2031/1/012062.
- [17] E. A. Wibowo, "Phishing website detection using neural network and PCA based on feature selection," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 9, no. 2, pp. 1150–1153, 2020, doi: 10.35940/ijrte.2277-3878.
- [18] G. Alshammari, M. Alshammari, T. S. Almurayziq, A. Alshammari, and M. Alsaffar, "Hybrid phishing detection based on automated feature selection using the chaotic dragonfly algorithm," *Electronics*, vol. 12, no. 13, p. 2823, 2023, doi: 10.3390/electronics12132823.
- [19] F. Salahdine, Z. Elmabet, and N. Kaabouch, "Phishing attacks detection: A machine learning-based approach," in *2021 IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 2021, doi: 10.1109/UEMCON53757.2021.9666627.
- [20] P. Ganesh and S. Kalaiarasi, "SVM and random forest algorithm are used well to detect phishing attacks for enhanced accuracy," *Studies in Fuzziness and Soft Computing*, vol. 10, no. 1S, pp. 531, 2023, doi: 10.17762/sfs.v10i1S.531.
- [21] T. Shahzad and K. Aman, "Unveiling the efficacy of AI-based algorithms in phishing attack detection," *Journal of Informatics and Web Engineering*, vol. 3, no. 2, pp. 116–133, 2024, doi: 10.33093/jiwe.2024.3.2.9.

## BIOGRAPHIES OF AUTHORS

	<p><b>Mohamad Asraf Daniel Bin Mohammed Feisal</b> is a Student in Multimedia University. Completed his study in Bachelor of IT Hons. Security Technology. His research focuses on Automated Detection of Phishing Websites Using Machine Learning Techniques. He can be contacted at email: <a href="mailto:asrafDaniel02@gmail.com">asrafDaniel02@gmail.com</a></p>

	<p><b>Siew-Chin Chong</b>, an IEEE Senior Member, earned her B.IT in Software Engineering, M.Sc in Information Technology, and Ph.D. in Information Technology from Multimedia University in 2003, 2006, and 2018, respectively. Currently, she is the Deputy Dean of Student Experience &amp; Alumni at the Faculty of Information Science and Technology, Multimedia University, Malaysia. Her research interests include machine learning, biometric security, and mobile app development, and she has published extensively in these areas. Additionally, she has served as an Editorial Board Member for several journals and as Technical Chair for numerous international conferences. She can be contacted at email: <a href="mailto:chong.siew.chin@mmu.edu.my">chong.siew.chin@mmu.edu.my</a></p>
	<p><b>Lee-Ying Chong</b> received B. IT. (Hons.) majoring in Information System Engineering in year of 2003. She received her Master degree in Science, majoring Information Technology from Multimedia University, in the year of 2007. She obtained the degree of Doctor of Philosophy (Information Technology) in year 2018. Her current research interests include biometrics authentication, computer vision and machine learning. She is senior member of IEEE since 2013. Currently she is working as a senior lecturer in Faculty of Information Science and Technology, Multimedia University, Malaysia. She can be contacted at email: <a href="mailto:lychong@mmu.edu.my">lychong@mmu.edu.my</a></p>
	<p><b>Kuok-Kwee Wee</b> received his BSc in Computer Science and MSc in Networking from University Putra, Kuala Lumpur, Malaysia. He then completed his study in PhD (Engineering) from Multimedia University, Malaysia. He is currently Associate Professor at the Faculty of Information Science and Technology in Multimedia University, Melaka, Malaysia. He is also a member of Editor Board of an International journal, Senior Member of IEEE, reviewer for international conferences and journals, active member of several international bodies. With his vast knowledge, skills and experiences in academia and industrial R&amp;D, he is be consultant and trainer for government and private sectors. His research interests include Quality ofservice, broadband wireless access, networking and mobile communication. He can be contacted at email: <a href="mailto:wee.kuok.kwee@mmu.edu.my">wee.kuok.kwee@mmu.edu.my</a></p>