Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Cyber Threat Detection Using an Ensemble Model Approach for Phishing Website Identification

Dani Rofianto[1], Egi Safitri[2], Khusnatul Amaliah[3], Jaka Fitra[4], Astria Hijriani[5]*

[1,3,4]Politeknik Negeri Lampung, Bandar Lampung, Indonesia
[2]Institut Informatika dan Bisnis Darmajaya, Bandar Lampung, Indonesia
[5]Ulsan National Institute of Science and Technology (UNIST), South Korea

[1]danirofianto@polinela.ac.id, [2]egisafitri@darmajaya.ac.id, [3]khusnatul@polinela.ac.id, [4]jakafitra@polinela.ac.id, [5]astria.hijriani@unist.ac.kr

ARTICLE INFORMATION

Article History:
Received: August 26, 2024
Last Revision: September 26, 2024
Published *Online*: September 30, 2024

KEYWORDS

Cyber Security,
Phishing,
Ensemble Learning,
Phishing Detection

CORRESPONDENCE

Phone: 085396491095
E-mail: danirofianto@polinela.ac.id

ABSTRACT

The development of digital technology has had a significant impact on various aspects of life, including an increase in cybersecurity threats, especially phishing attacks. Phishing is a method of cyber fraud that manipulates victims to provide sensitive information by posing as a trusted entity. This research aims to develop and evaluate the effectiveness of several machine learning algorithms in detecting phishing websites. The methods used in this research include the application of Random Forest, Extra Trees, Multiple Layer Perceptron, Ada Boost, and Decision Tree algorithms on website datasets containing the characteristics of phishing and non-phishing sites. Performance evaluation is performed by measuring the accuracy, precision, recall, and F1 value of each algorithm. In addition, a voting technique is applied to combine the results of the best-performing algorithms with the aim of improving the overall detection accuracy. The results showed that the voting technique was able to provide superior results compared to the use of a single algorithm, with significant improvements in accuracy and recall values. These findings reinforce the importance of ensemble approaches in machine learning to improve phishing detection capabilities, which in turn contributes to improved cybersecurity.

1. INTRODUCTION

During rapid technological advances, the internet and mobile devices have become an essential part of everyday life [1], [2]. Innovations in these technologies enable easy and quick access to global information. However, along with these great benefits come significant security risks, especially in the form of cybercrime such as phishing [3], [4]. Phishing is a form of online fraud that uses fake email addresses or websites to obtain users' personal information [5], [6]. Stolen information includes personal data (such as name, address, gender, and date of birth), account information (such as username and password), or financial information (such as credit card and account data) [7]. Phishers, often called phishers, use a variety of methods, including the sending of fake emails that appear to come from a particular bank or service, as well as the spread of malware [8].

One form of phishing attack is domain phishing, where the perpetrator obtains sensitive information without authorization by using a domain that mimics the original website[9]. In this attack, users are redirected to fake websites that look like legitimate ones or forced to provide personal information through blackmail. When users enter personal data, they unwittingly give attackers access to information that can be used for identity theft [8].

As cyberattacks become more intricate and complex, the challenges in accessing, assessing, and responding to cybersecurity issues continue to increase [10], [11], [12]. Data from the APWG (Anti-Phishing Working Group) shows that there were more than 51,000 different phishing sites in 2016 [13]. According to an analysis by Rivest Shamir Adleman (RSA), phishing attacks caused $9 billion in losses to global companies that same year, with over one million phishing attacks recorded, a 65% increase over the previous year. This increase in the number of phishing attacks has negatively impacted consumer trust in online platforms[14]. Various types of online fraud with phishing websites are often one of the common methods in social engineering on the internet [15]. Hackers create web pages that mimic trusted sites and then disseminate suspicious URLs through spam chats, messages, or social media. Unsuspecting users may think the URL is real. If they enter personal information, such as bank account numbers or government savings numbers, into the link, their data may be jeopardized [16]. Assembling models offers a promising solution to address the threat of phishing. Machine learning allows systems to learn from existing data and become more intelligent without the need for explicit definitions. However, the application of machine learning also faces challenges, such as the need for large memory, complicated labelling processes, and sometimes less accurate results.

Based on the above, this research aims to develop and implement an assembling model that combines various machine-learning techniques to improve accuracy and effectiveness in detecting phishing websites. By focusing on analyzing critical features such as URLs and HTML, this research aims to evaluate the contribution of each feature in improving phishing detection capability. In addition, this research will measure the performance of the ensemble model in terms of accuracy, sensitivity, and specificity and ensure its reliability under various conditions and datasets.

2. Related Work

Several previous studies have been conducted related to phishing website detection, one of which was conducted by Tang & Mahmoud [17]. The paper reviews phishing detection methods, emphasizing the limitations of traditional techniques like blacklists and highlighting the role of machine learning in improving prediction accuracy. It discusses the lifecycle of phishing attacks and compares various machine learning-based approaches for detecting phishing websites, focusing on data collection, feature extraction, modeling, and performance evaluation. The next research was conducted by Ozcan et al. [18] this study proposed a hybrid deep learning model based on Long Short-Term Memory and Deep Neural Network algorithms to detect Uniform Resource Locator (URL) phishing and evaluate the performance of the model on phishing datasets. The experimental results show that the proposed model achieves superior accuracy compared to other phishing detection models. Kurniawan et al. [19] study the risk of attacks on machine learning models that use IoT sensor-based architectures, specifically adversarial instance attacks that can cause the system to produce incorrect outputs. Previous studies have assumed that an attacker must access all features, but the impact of hacking on only a few sensors has not been addressed. This research explores the possibility of attacks on deep neural network (DNN) models by hacking several sensors. Experiments were conducted on a human activity recognition model with three sensors mounted on the user's chest, wrist, and ankle, and the results show that attacks can be carried out by hacking a limited number of sensors.

Karim et al. [20] study addresses the growing threat of phishing attacks, one of the most severe forms of cybercrime. By leveraging a dataset of over 11,000 phishing and legitimate URLs, the research applies machine learning algorithms, including decision trees, random forests, and a proposed hybrid LSD model (combining logistic regression, support vector machine, and decision tree) to identify and prevent phishing attempts effectively. The study employs feature selection, cross-validation, and hyperparameter optimization techniques to enhance model performance. Evaluation metrics such as accuracy, precision, recall, F1-score, and specificity demonstrate that the proposed approach outperforms existing methods, offering superior protection against phishing attacks.

Furthermore, Subhashini and Narmatha [21] evaluates how prediction accuracy in unbalanced datasets using the Synthetic Minority Over-Sampling Technique (SMOTE). The proposed model performs better than popular binary classification techniques such as Random Forest and XGBoost. The results showed that CatBoost achieved detection accuracy of up to 97%, making it a much better classifier than Random Forest and XGBoost. Another research was conducted by Ahasan et al. [22]. This research proposes an optimized Fuzzy Multi-Criteria Decision-Making (OFMCDM) and Improved Random Forest (IRF) based phishing detection model. The model utilizes Uniform Resource Locator (URL) and Hypertext Markup Language (HTML) features to prevent the sharing of sensitive user information such as username, password, social security number, or credit card number. Experiments show that the model provides competitive results compared to existing models, including Naive Bayes (NB), Logistic Regression (LR), K-Nearest Neighbor (KNN), and Decision Tree.

3. Methodology

The methodology contains the technical stages that will be carried out at the research stage. The experimental phase was conducted thoroughly to ensure the validity of the research results. The data acquisition process is a crucial first step, given that the quality of the data greatly affects the performance of the model. Next, the data is pre-processed to ensure that the data used by the model is clean, structured, and ready for processing. The next step is exploratory data analysis, which helps understand patterns in the data. Finally, model fitting is performed.

At this stage, various popular algorithmic models such as SVC, Decision Tree, Ada Boost, XGBoost, Random Forest, Extra Trees, Multiple Layer Perceptron (Neural Network), KNN, Logistic Regression, and Linear Discriminant Analysis are tested and fit to the training dataset. This fitting process allows the model to learn patterns and relationships in the data to make predictions on new data. Evaluation of the results is an important stage in assessing the quality of the model. Metrics such as accuracy, precision, recall, and F1-score are evaluated using the training and test datasets to understand the extent to which the model can predict correctly and consistently. Further analysis of the confusion matrix and precision-recall curve provides an in-depth understanding of the strengths and weaknesses of each model. Finally, a conclusion is drawn to summarize the research results obtained. The classification levels on the data for the algorithm are as follows [23]:

1. Excellent classification = 0.90 - 1.00
2. Good classification = 0.80 - 0.90
3. Fair classification = 0.70 - 0.80
4. Low classification = 0.60 - 0.70

The research method stages involve data collection, preprocessing, feature selection, model training, validation, and testing. This ensures robust evaluation of the model's accuracy, precision, recall, and F1-score. Afterward, a comparison is made between different models based on performance metrics.

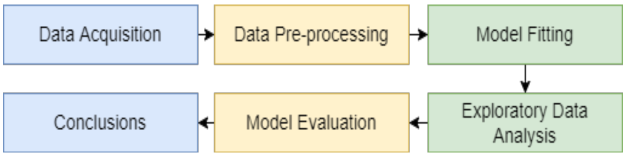

FIGURE 1. RESEARCH PROCESS FLOW

4. RESULT AND DISCUSSION

In this section of results and discussion, we discuss the experimental results of using various machine learning models in phishing website identification. This exploration involves a comprehensive set of steps, from data acquisition, data pre-processing, data visualization, model fitting, model evaluation, and inference. Each of these steps is conducted thoroughly to ensure the validity and accuracy of the results and to provide an in-depth understanding of the effectiveness of each model in detecting cyber threats through phishing website identification.

4.1 Data Acquisition

This step covers the collection of information used in the research, which involves the use of secondary data. The dataset in this study was obtained from the official website https://www.kaggle.com/datasets, a public data source. The dataset for phishing websites consists of 100,076 data samples with a total of 20 attributes (19 attributes and one target attribute). In this research, the attributes used to analyze and classify URLs reflect various characteristics of the URL structure that can help detect whether the URL is phishing. The first attribute is url_length, which measures the URL length and can indicate the complexity or disguise potential of a phishing site. Next, attributes such as n_dots (number of dots), n_hyphens (number of hyphens), n_underline (number of underscores), and n_slash (number of slashes) reflect the character usage pattern in the URL, where phishing URLs tend to have more special characters to trick users.

Other attributes, such as n_questionmark (number of question marks), n_equal (number of equal signs), n_at (number of @ signs), and n_and (number of & signs), frequently appear in suspicious URLs and may indicate disguised login pages or fake forms. Additionally, the presence of symbols such as n_exclamation, n_space, n_tilde, n_comma, n_plus, n_asterisk, n_hashtag, n_dollar, n_percent, and n_redirection (the “//” redirection mark) provide additional indications about the nature of the URL, as phishing URLs often have unusual patterns in the use of these characters. Finally, the phishing attribute serves as a target label that indicates whether the URL is phishing (1) or not (0). Analyzing the combination of these attributes allows the machine learning model to recognize the patterns and characteristics of suspicious URLs more accurately, thus helping to effectively distinguish between phishing and legitimate URLs.

4.2 Data Preprocessing

To improve the effectiveness and accuracy of machine learning models for detecting phishing websites, a series of data preprocessing stages are required. Data processing is an important first step before the data can be used to train and test the model. In this preprocessing stage, various crucial steps are taken to ensure good data quality and representation.

1. Data Segregation: The initial data is divided into attributes (X) and target variables (y), where the target variable is the 'phishing' to be predicted.
2. Train and Test Data Split: Using the train_test_split function of scikit-learn, the data is divided into training data (X_train, y_train) and testing data (X_test, y_test). This division is done to train the model on the training data and test the model performance on the testing data.
3. Class Distribution Check: Displays the class distribution on the training data (y_train) and testing data (y_test). This class distribution is important to ensure that the division of training and testing data reflects the same class distribution.
4. Standard Scaling: Preprocessing of attributes using StandardScaler. StandardScaler scales the features in the data to have mean=0 and standard deviation=1. It is especially useful for linear and KNN models that are sensitive to variable scaling.

By preprocessing, the data is ready to be used in training and testing machine learning models to detect phishing websites.

4.3 Data Visualization

Data visualization analysis was conducted to answer important questions regarding the class distribution and relationships between attributes in the dataset. Figure 3 illustrates the percentage class imbalance between phishing websites and legitimate websites. This visualization helps understand the class distribution in the dataset, with important information such as the percentage of phishing and non-phishing websites. Analyzing this class distribution is important to understand its potential impact on the performance of the prediction model and to take appropriate steps to address class imbalance. The bar chart in the image compares the percentages of phishing websites (represented by the label "1") and legitimate websites (represented by the label "0"). Based on the visualization, there is a clear class imbalance, with legitimate websites (0) having a higher percentage compared to phishing websites (1). Phishing websites make up approximately 40%, while legitimate websites account for around 60%. This class imbalance is essential to note because it may affect the performance of a prediction model. When training a model on imbalanced data, the model may be biased toward the majority class (legitimate websites), leading to reduced accuracy in predicting phishing websites.

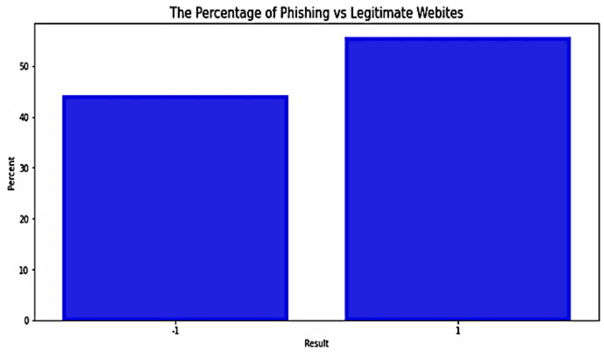

FIGURE 2. PERCENTAGE OF PHISHING WEBSITES AND LEGITIMATE WEBSITES

Next, we analyzed the correlation of the dataset to understand the relationship between the variables in the dataset. Correlation heatmaps provide a clear visual picture of the extent to which certain attributes correlate with each other[24]. This information is useful for identifying features that have a close relationship with phishing web page categories, which can provide valuable insights in the development of predictive models. Correlations between variables also provide additional insights into the factors that influence the likelihood of web page phishing. The results of the correlation heatmap visualization are presented in Figure 3. The displayed correlation heatmap illustrates the relationship between various attributes in the web page-phishing dataset, which consists of 20 URL features and a target label indicating whether the URL is phishing (1) or not (0). The colors on the heatmap range from blue to yellow, where dark blue indicates a strong negative correlation (close to -1) and bright yellow indicates a strong positive correlation (close to 1). Green color gradations indicate a weaker or no correlation (close to 0). The number in each cell of the heatmap indicates the value of the correlation coefficient between the two attributes, with values ranging from -1 to 1, which provides information about the linear relationship between the attributes.

The attributes used in this correlation analysis include various features that describe URL characteristics, such as URL length (url_length), number of dots (n_dots), number of hyphens (n_hyphens), and various other symbols such as question marks (n_questionmark), equals (n_equal), exclamation marks (n_exclamation), as well as other attributes that reflect the structure and complexity of the URL. The phishing target label serves as an indicator to determine if the URL is potentially phishing. This correlation analysis is critical in understanding how certain URL characteristics, such as many hyphens or question marks, can relate to the likelihood of a URL being phished. This interpretation provides valuable insights for further analysis, particularly in developing machine learning models that can detect phishing more accurately based on relevant URL features.

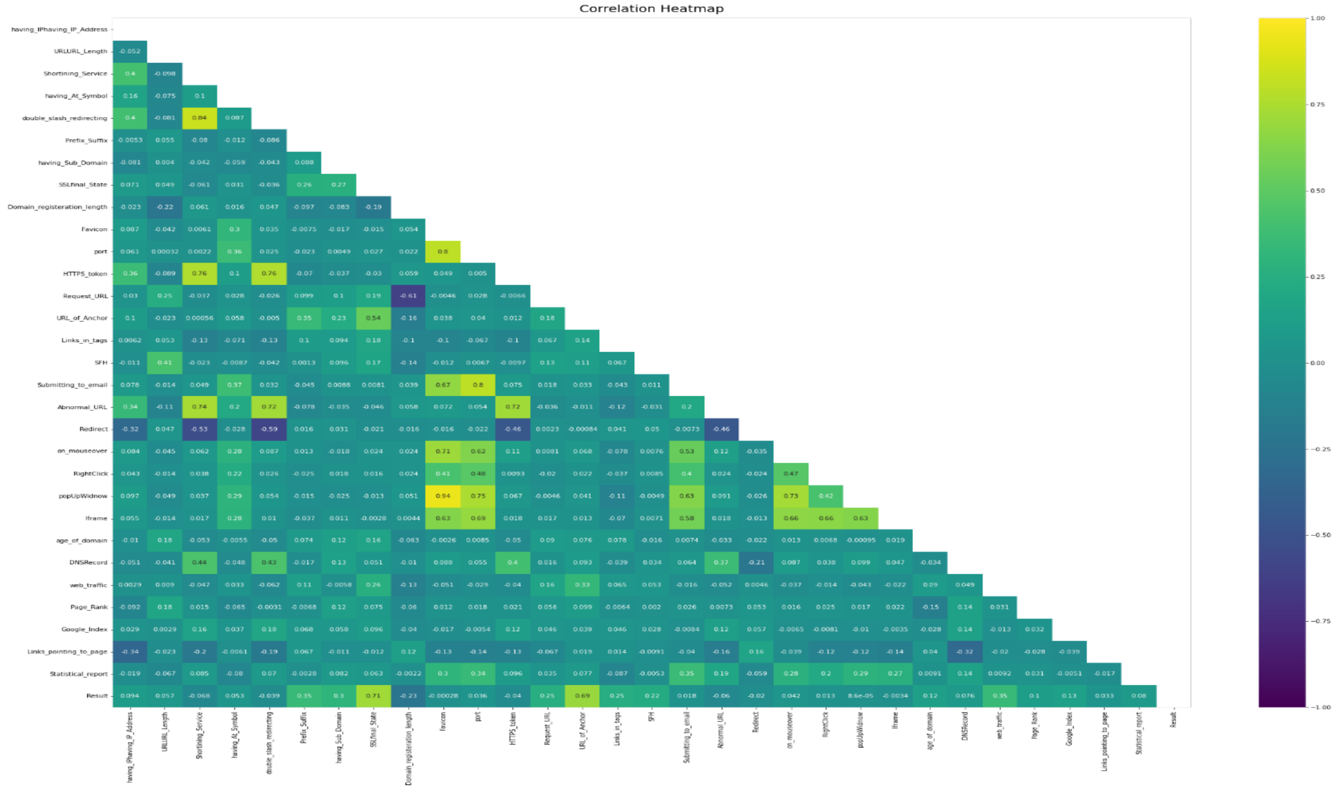

FIGURE 3. CORRELATION HEATMAP

4.4 Fitting and Model Evaluation

In this model testing phase, we used ten different Machine Learning algorithms, including SVC, Decision Tree, Ada Boost, XGBoost, Random Forest, Extra Trees, Multiple Layer Perceptron (Neural Network), KNN, Logistic Regression, and Linear Discriminant Analysis. Each algorithm is unique in handling datasets and phishing web page classification problems. Of course, the main goal of this test is to find and identify the best model based on the execution results. We will analyze the performance of each algorithm and compare them to determine the most optimal model in the context of phishing web page identification. The performance of each model will be evaluated based on the average value of cross-validation

(Cross-val Means) and error rate (Cross-val errors). Models with high cross-validation mean and low errors will perform well and effectively in classification tasks. Details of the experimental test results are presented in Table 1, which offers a complete picture of the effectiveness and efficiency of each algorithm in classifying phishing web pages.

TABLE 1. EXPERIMENTAL RESULTS USING SEVERAL MACHINE LEARNING ALGORITHMS

Algorithm	Cross Val Means	Cross Val Errors
Random Forest	0.971644	0.006727
Extra Trees	0.970563	0.007782
Multiple Layer Perceptron	0.966514	0.006524
Ada Boost	0.960301	0.007384
Decision Tree	0.956385	0.004848
Gradient Boosting	0.946803	0.008312
SVC	0.946803	0.005837
K-Neighbors	0.937212	0.007434
Logistic Regression	0.929250	0.007638
Linear Discriminant Analysis	0.921552	0.006087

The experimental results show that of the various algorithms tested, the five best algorithms for phishing website detection are Random Forest, Extra Trees, Multiple Layer Perceptron, Ada Boost, and Decision Tree. Random Forest and Extra Trees excel with the highest accuracy and lowest average error, making them highly effective with stable and consistent performance. Multiple Layer Perceptron also performed well with high accuracy and low average error, although it requires more data to achieve optimal results. Ada Boost provides good accuracy and the ability to correct previous model errors but has a slightly higher average error. Decision Tree, although slightly lower in accuracy compared to the other models, showed a low average error. Overall, Random Forest and Extra Trees stand out as the best algorithms in terms of accuracy and error consistency, while the other models also provide competitive performance.

To better understand the performance of the best algorithms obtained, in this case, Random Forest, Extra Trees, Multiple Layer Perceptron, Ada Boost, and Decision Tree on phishing website detection, we used learning curve plots. Learning curves help to evaluate how well the model learns from training data and how generalizable the model is to unseen data. *Learning Curves* are graphs that show the relationship between the size of the training dataset and the performance of the model, both on training data and test data. It helps identify whether the model is overfitting, underfitting, or having other issues in the training process. The learning curves of the five best experimental algorithms are presented in Figure 4.

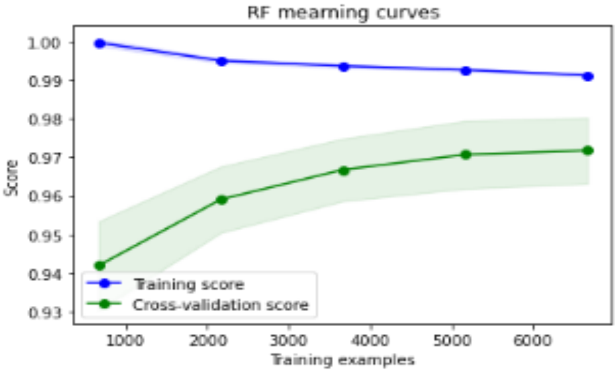

FIGURE 4. RF ALGORITHM LEARNING CURVES GRAPH

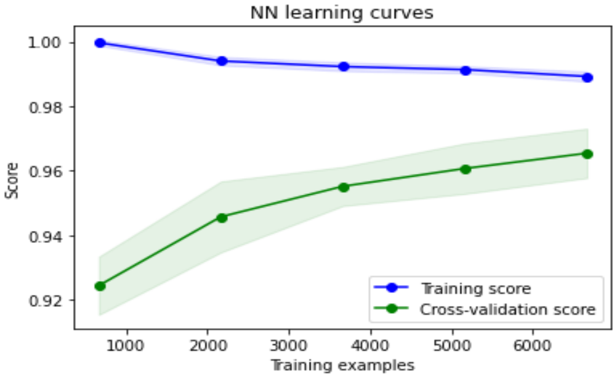

FIGURE 5. NN ALGORITHM LEARNING CURVES GRAPH

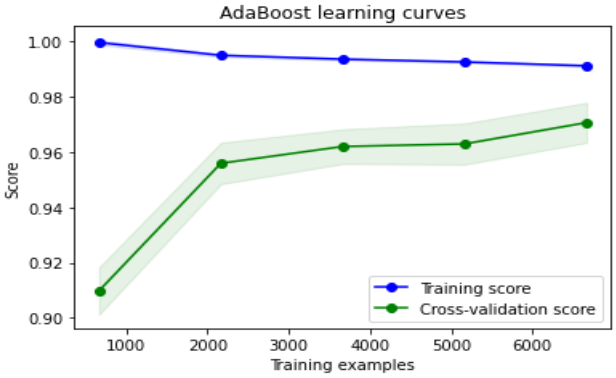

FIGURE 6. ADABOOST ALGORITHM LEARNING CURVES GRAPH

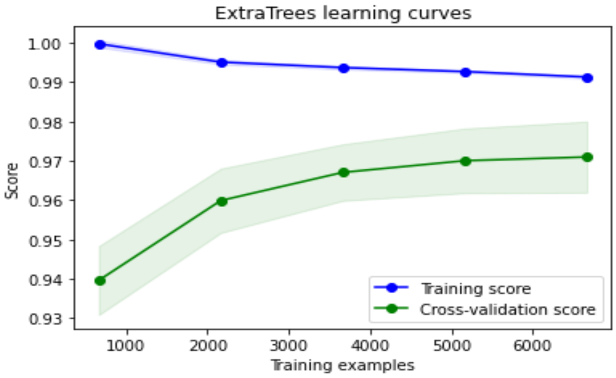

FIGURE 7. EXTRA TREES ALGORITHM LEARNING CURVES GRAPH

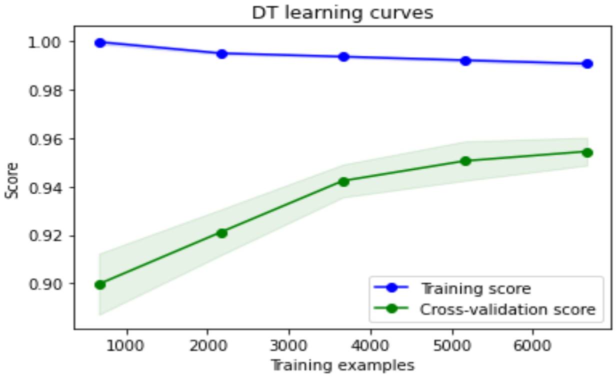

FIGURE 8. DT ALGORITHM LEARNING CURVES GRAPH

Based on the learning curves above, Random Forest and Extra Trees tend to show stable performance and good generalization, while MLP and AdaBoost may need more attention in terms of hyperparameters and data size. Decision Trees show potential for overfitting, which needs to be monitored carefully. This analysis underlies the selection of better models to be applied in phishing detection systems. Furthermore, voting or model aggregation approaches will be applied to improve the phishing website detection performance. The voting approach involves combining the predictions of multiple models to make a final decision, usually by taking a majority vote on the model predictions. This aims to utilize the strengths of each model and reduce their weaknesses.

1. Random Forest and Extra Trees are ensemble models that have excellent and stable performance. Using them in voting can improve accuracy and reduce the variability of results.
2. Multiple layer perceptron offers high accuracy and can capture non-linear relationships that may not be fully captured by ensemble models. Incorporating MLP in voting can enrich the model representation.
3. Ada Boost allows users to focus on difficult examples and correct previous model errors. This can add power to the voting model by handling more complex cases.
4. Decision Tree, while slightly lower in accuracy, offers high interpretability and can provide additional insights into the data that might help in the final decision-making.

After training the Voting Classifier with training data (X_train, y_train), the model was tested on test data (X_test), resulting in a prediction accuracy of 97.51%. To further evaluate the performance, a Confusion Matrix is used to visualize the model performance.

In this visualization, a heatmap is used to provide a clear picture of the model's prediction distribution. The blue color highlights the intensity of the number of predictions in each category, with annotations showing the corresponding percentage values. This helps in identifying areas where the model performs well and where improvements may be needed. The Confusion Matrix plot (Figure 5) provides an overview of the model's accuracy and error distribution, which is crucial for understanding the strengths and weaknesses of the Voting Classifier model in detecting phishing websites. This visualization reinforces the result that the voting approach improves prediction accuracy by leveraging the combined strengths of the five best algorithms.

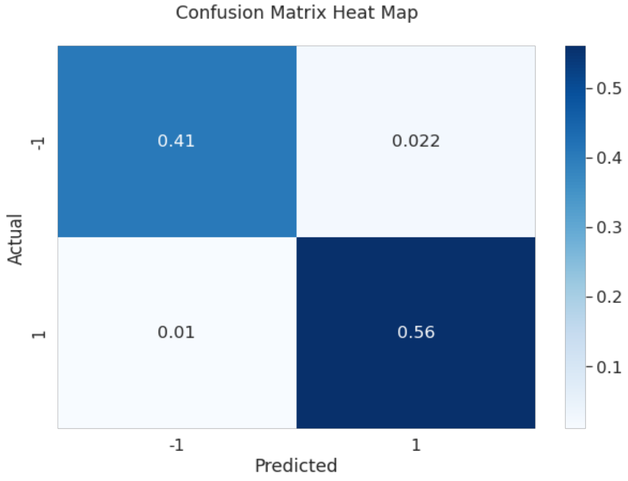

Figure 5. Confusion Matrix Heat Map

The following Classification Report details the evaluation metrics at the Train and Test stages. The Classification Report contains information on precision, recall, and F1-score for each class, as well as accuracy, which helps in evaluating the extent to which the model can correctly predict certain classes. More detailed results are presented in Table 2 below.

Table 2. Metrics Evaluation

	precision	recall	f1-score	support
Phishing	0.98	0.95	0.96	1565
Legitimate	0.96	0.98	0.97	2084
Accuracy			0.97	3649
Macro avg	0.97	0.96	0.97	3649
Weighted avg	0.97	0.97	0.97	3649

Table 2 shows the model's performance metrics evaluation results in classifying URLs as phishing or legitimate using the proposed machine learning model. The model is evaluated based on several metrics: precision, recall, f1-score, and support for each class (Phishing and Legitimate). In the phishing class, the model achieved a precision of 0.98, which indicates the model's accuracy in correctly identifying phishing URLs. A 0.95 recall indicates that the model could detect 95% of the total phishing URLs. The F1-score, the harmonic mean between precision and recall, reached 0.96, indicating a good balance between the two metrics. For the legitimate class, the precision and recall were 0.96 and 0.98, respectively, with an F1-score of 0.97, indicating that the model also effectively recognized legitimate URLs. The model's overall accuracy reached 0.97, indicating that 97% of the total predictions made were correct. The macro avg and weighted average of 0.97 for precision, recall, and F1-score, respectively, indicate that the model performs consistently and is superior in classifying both classes. With a total data set of 3,649, these results show that the model performs very well and can tackle the phishing classification problem with high accuracy and effectiveness.

4.5 Discussion

From the experimental results, the five best algorithms for phishing website detection are Random Forest, Extra Trees, Multiple Layer Perceptron, Ada Boost, and Decision Tree. Random Forest and Extra Trees algorithms show superior performance with the highest accuracy and lowest average error, making them highly effective in detecting phishing with stable and consistent performance. Multiple Layer Perceptron also performed well with high accuracy and low average error, although it requires more data to achieve optimal results. Ada Boost provides good accuracy and the ability to correct previous model errors but has a slightly higher average error. Decision Tree, despite having slightly lower accuracy compared to the other models, shows a low average error. The Voting Classifier approach, which combines these five algorithms, resulted in a prediction accuracy of 97.51%. Evaluation of model performance using the Confusion Matrix shows a clear distribution of predictions and helps identify areas where the model performs well as well as areas that require improvement.

Analysis of model performance evaluation metrics such as precision, recall, F1-score, and support for each class shows that the developed model can detect phishing and non-phishing websites with high accuracy. The macro average and weighted average of these metrics also show the model's excellent overall performance. Overall, the ensemble approach with Voting Classifier successfully improves the phishing website detection performance, provides a model with high accuracy and error consistency, and effectively solves phishing-related cybersecurity problems.

5. Conclusions

This research shows that the use of the five best algorithms, Random Forest, Extra Trees, Multiple Layer Perceptron, Ada Boost, and Decision Tree in the Voting Classifier, can improve accuracy and consistency in detecting phishing websites. Random Forest and Extra Trees stand out with superior performance, while Multiple Layer Perceptron, Ada Boost, and Decision Tree also make significant contributions. Evaluation using the Confusion Matrix and precision, recall, and F1-score metrics shows that the model can detect phishing very accurately. With a prediction accuracy of 97.51%, this ensemble approach proves effective as a solution to improve cybersecurity against phishing threats. The combination of these algorithms in the Voting Classifier improves model robustness by leveraging the strengths of each method. Random Forest and Extra Trees excel in handling complex data patterns, while Multiple Layer Perceptron, AdaBoost, and Decision Tree enhance model diversity and stability. This collaborative approach ensures comprehensive detection of phishing websites, minimizing false positives and negatives. The high prediction accuracy of 97.51% highlights its potential to mitigate evolving cybersecurity threats. Ultimately, this ensemble model offers a reliable and scalable solution to bolster online security and protect users from phishing attacks.

References

[1] J. Guaña-Moya, M. A. Chiluisa-Chiluisa, P. del C. Jaramillo-Flores, D. Naranjo-Villota, E. R. Mora-Zambrano, and L. G. Larrea-Torres, "Ataques de phishing y cómo prevenirlos Phishing attacks and how to prevent them," in *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, 2022, pp. 1–6. doi: 10.23919/CISTI54924.2022.9820161.

[2] N. Frevel, D. Beiderbeck, and S. L. Schmidt, "The impact of technology on sports – A prospective study," *Technol Forecast Soc Change*, vol. 182, 2022, doi: 10.1016/j.techfore.2022.121838.

[3] Kunal, M. Rana, D. Sharma, and Anurag, "Understanding Cyber-Attacks and their Impact on Global Financial Landscape," in *2023 International Conference on Circuit Power and Computing Technologies (ICCPCT)*, 2023, pp. 1452–1456. doi: 10.1109/ICCPCT58313.2023.10245828.

[4] Ö. Aslan, S. S. Aktuğ, M. Ozkan-Okay, A. A. Yilmaz, and E. Akin, "A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions," 2023. doi: 10.3390/electronics12061333.

[5] A. A. Hasegawa, N. Yamashita, M. Akiyama, and T. Mori, "Experiences, Behavioral Tendencies, and Concerns of Non-Native English Speakers in Identifying Phishing Emails," *Journal of Information Processing*, vol. 30, 2022, doi: 10.2197/ipsjjip.30.841.

[6] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing Attacks: A Recent Comprehensive Study and a New Anatomy," 2021. doi: 10.3389/fcomp.2021.563060.

[7] A. Redi and N. Ernasari, "Efforts to Overcome Web-Based Phishing Crimes in the World of Cyber Crime," in *Proceedings of the 3rd Multidisciplinary International Conference, MIC 2023, 28 October 2023, Jakarta, Indonesia*, EAI, 2023. doi: 10.4108/eai.28-10-2023.2341807.

[8] Bhuvana, A. S. Bhat, T. Shetty, and Mr. P. Naik, "A Study on Various Phishing Techniques and Recent Phishing Attacks," *International Journal of Advanced Research in Science, Communication and Technology*, 2021, doi: 10.48175/ijarsct-2094.

[9] S. Zhang, Z. Yan, K. Dong, H. Li, and X. Yuchi, "Phishing Domain Name Detection Based on Hierarchical Fusion of Multimodal Features," in *2022 IEEE 16th International Conference on Big Data Science and Engineering (BigDataSE)*, IEEE, Dec. 2022, pp. 1–6. doi: 10.1109/BigDataSE56411.2022.00010.

[10] S. Oh and T. Shon, "Cybersecurity Issues in Generative AI," in *2023 International Conference on Platform Technology and Service, PlatCon 2023 - Proceedings*, 2023. doi: 10.1109/PlatCon60102.2023.10255179.

[11] N. S. M. Mizan, M. Y. Ma'arif, N. S. M. Satar, and S. M. Shahar, "Cnds-cybersecurity: Issues and challenges in asean countries," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.4 S1, 2019, doi: 10.30534/ijatcse/2019/1781.42019.

[12] S. K. Khan, N. Shiwakoti, P. Stasinopoulos, and M. Warren, "Cybersecurity regulatory challenges for connected and automated vehicles – State-of-the-art and future directions," *Transp Policy (Oxf)*, vol. 143, 2023, doi: 10.1016/j.tranpol.2023.09.001.

[13] S. Chanti and T. Chithralekha, "A literature review on classification of phishing attacks," 2022. doi: 10.19101/IJATEE.2021.875031.

[14] "APWG: Phishing Activity Trends Report Q4 2018," *Computer Fraud & Security*, vol. 2019, no. 3, pp. 4–4, Jan. 2019, doi: 10.1016/S1361-3723(19)30025-9.

[15] S. Asiri, Y. Xiao, S. Alzahrani, S. Li, and T. Li, "A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks," *IEEE Access*, vol. 11, pp. 6421–6443, 2023, doi: 10.1109/ACCESS.2023.3237798.

[16] W. Li, S. Manickam, S. U. A. Laghari, and Y.-W. Chong, "Uncovering the Cloak: A Systematic Review of Techniques Used to Conceal Phishing Websites," *IEEE Access*, vol. 11, pp. 71925–71939, 2023, doi: 10.1109/ACCESS.2023.3293063.

[17] L. Tang and Q. H. Mahmoud, "A Survey of Machine Learning-Based Solutions for Phishing Website Detection," 2021. doi: 10.3390/make3030034.

[18] A. Ozcan, C. Catal, E. Donmez, and B. Senturk, "A hybrid DNN–LSTM model for detecting phishing URLs," *Neural Comput Appl*, vol. 35, no. 7, pp. 4957–4973, Mar. 2023, doi: 10.1007/s00521-021-06401-z.

[19] A. Kurniawan, Y. Ohsita, and M. Murata, "Experiments on Adversarial Examples for Deep Learning Model Using Multimodal Sensors," *Sensors*, vol. 22, no. 22, p. 8642, Nov. 2022, doi: 10.3390/s22228642.

[20] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari, and S. R. K. Joga, "Phishing Detection System Through Hybrid Machine Learning Based on URL," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3252366.

[21] K. Subashini and V. Narmatha, "Website Phishing Detection of Machine Learning Approach using SMOTE method," in *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, IEEE, Feb. 2023, pp. 1–5. doi: 10.1109/ICECCT56650.2023.10179745.

[22] M. A. Al Ahasan, M. Hu, and N. Shahriar, "OFMCDM/IRF: A Phishing Website Detection Model based on Optimized Fuzzy Multi-Criteria Decision-Making and Improved Random Forest," in *2023 Silicon Valley Cybersecurity Conference (SVCC)*, IEEE, May 2023, pp. 1–8. doi: 10.1109/SVCC56964.2023.10165344.

[23] P. Subarkah and A. N. Ikhsan, "Identifikasi Website Phishing Menggunakan Algoritma Classification And Regression Trees (CART)," *Jurnal Ilmiah Informatika*, vol. 6, no. 2, pp. 127–136, Dec. 2021, doi: 10.35316/jimi.v6i2.1342.

[24] V. Fey, D. Jambulingam, H. Sara, S. Heron, C. Sipeky, and J. Schleutker, "Biocpr–a tool for correlation plots," *Data (Basel)*, vol. 6, no. 9, 2021, doi: 10.3390/data6090097.

AUTHORS

Dani Rofianto

He is an Software Engineering professional, with a Bachelor's degree in Mathematics from Halu Oleo University (2016) and a Master's degree from IPB University (2020). As a lecturer, he is active in Mathematics research, Data Science and software development, focusing on knowledge and practical problem solving. With high dedication, Dani is committed to advancing and spreading knowledge, inspiring many individuals in the academic world.

Egi Safitri

She is an Applied Mathematics and Data Science professional. She began her academic journey by obtaining a bachelor's degree in mathematics from Halu Oleo University in 2016 and a master's degree from the University of Indonesia in 2020. She contributed to developing students' understanding of complex mathematical concepts. As a lecturer, she is also actively involved in Applied Mathematics and Data Science research, focusing on knowledge development and practical problem-solving. With a burning passion and high dedication, she has proven herself a committed professional who advances and disseminates knowledge. Her dedication to education and research inspires many aspiring individuals in the academic world.

Khusnatul Amaliah

She is a Software Engineering professional, holding a Bachelor's degree in Information Engineering from Universitas Amikom Yogyakarta (2016) and a Master's degree Universitas Amikom Yogyakarta (2018). As a lecturer, she is actively engaged in research software development. With a high level of dedication, Khusna is committed to advancing and sharing knowledge, inspiring many individuals in the academic world.

Jaka Fitra

He is a Software Engineering professional, holding a Bachelor's degree in Computer Science from UMITRA (2013) and a Master's degree from IBI Darmajaya (2020). As a lecturer, he is actively engaged in research on IoT Aquaculture, Data Science, and software development, with a focus on automation and practical problem-solving. With a high level of dedication, Jaka is committed to advancing and sharing knowledge, inspiring many individuals in the academic world.

Jaka Fitra

She is a PhD student at the Ulsan National Institute of Science and Technology, focusing on process mining value and maturity model development. She holds a bachelor's degree in computer science from Gadjah Mada University (2004) and a master's degree from the Informatics Institute of Technology Sepuluh November

(2013). Her research integrates process science, spatial data, application development, and artificial intelligence to optimize organizational performance through data-driven analysis. As a lecturer at the University of Lampung, she teaches software development, software design, and information systems. She also holds a postgraduate diploma in geoinformatics from the University of Twente and has completed advanced courses in database management, web programming, and total quality management.