Comparison of Machine Learning Methods in Performance of Phishing Website Classification

Jiahan Xie[a]
Department of Statistical Science, University College London, London, NW1 1ST, U.K.

Keywords: Decision Tree, Random Forest, K-Nearest Neighbor.

Abstract: In this article, three models using Decision Tree, Random Forest, and Gradient Boosting methods are built and evaluated in performance of phishing website classification. The model training and comparisons are based on a relevant dataset from Mendeley Data, on which a thorough data preprocessing and feature selection process is applied to ensure the quality of model evaluation, including handling of erroneous features and encoding for domain-related variables. Afterwards, grid search and a hybrid two-stage searching approach based on cross-validation are used for hyperparameter tuning. The Gradient Boosting model achieves the best performance regarding multiple evaluation metrics on the test set, with Random Forest being a close alternative. This result demonstrates that the use of ensemble learning methods can build more efficient classifiers compared to traditional machine learning methods. The study provides guidance for the selection of classification models for phishing websites and is expected to be helpful in future research concerning other ensemble machine learning models and deep learning models.

1 INTRODUCTION

A phishing website refers to a false site created by attackers to resemble legitimate websites visually and semantically. In common web-based phishing attack scenarios, such websites are utilized as a deception technique to gain the trust of users, persuade them to perform needed actions, and therefore obtain their sensitive private information such as identity data and financial account credentials (Varshney et al., 2024; Naqvi et al., 2023). By hosting replicas of HyperText Markup Language (HTML) contents of authentic websites on web servers, attackers can easily generate such websites and lure victims to them through various redirecting techniques, such as typosquatting, cross-site scripting (XSS), and link manipulation on websites (Varshney et al., 2024). Furthermore, phishing is not only a standalone threat but sometimes also a vector for other cyberattack mechanisms. Hence, phishing websites can be employed to execute additional attacks, for instance, ransomware attacks, thereby posing even sterner security challenges (Naqvi et al., 2023).

With the rapid digitization of services and the evolution of technologies, phishing websites have been an increasingly serious threat to cybersecurity. According to the Anti-Phishing Website Group (APWG), the second quarter of 2023 saw the third-highest quarterly total of phishing attacks, with more than 1.28 million attacks recorded (Kawale et al., 2024). Besides, International Business Machines Corporation (IBM) identified phishing as the attack vector attributed to the largest average amount of financial loss, with phishing attempts being reported across different sectors, including financial institutions, educational institutions and governmental organizations (Naqvi et al., 2023). In this context, it is urgent to develop effective classification methods in order to detect phishing websites and prevent potential loss.

One useful method for phishing detection implements machine learning (ML) techniques. Their usage is based on the idea that certain features of legitimate websites cannot be spoofed by phishing websites (Varshney et al., 2024). For example, there are features based on Uniform Resource Locator (URL) text analysis that expect to differ in safe sites and phishing sites, since two servers cannot run on the same URL via internet (Varshney et al., 2024; Hannousse and Yahiouche, 2021). The recent studies

[a] https://orcid.org/0009-0004-0671-866X

Xie, J.
Comparison of Machine Learning Methods in Performance of Phishing Website Classification.
DOI: 10.5220/0013827200004708

In *Proceedings of the 2nd International Conference on Innovations in Applied Mathematics, Physics, and Astronomy (IAMPA 2025)*, pages 434-444
ISBN: 978-989-758-774-0

concerning ML modelling in phishing website classification can be categorized by the type of features included in their datasets. Gupta et al. provided a parsimonious URL-based classification model of 9 variables extracted from URL text using Random Forest (RF), with an excellent performance of 99.57% accuracy (Gupta et al., 2021). Karim et al. designed a hybrid ML model combining Logistic Regression (LG), Support Vector Machines (SVM) and Decision Tree (DT) methods, which outperforms (accuracy of 95.23%, precision of 95.15%, recall of 96.38%, specificity of 93.77%, and F1-score 95.77%) other traditional ML methods (Karim et al., 2023). Ahammad et al. compared the performance of ML models such as DT, RF and Gradient Boosting Machines (GBM). Their GBM model achieved a training accuracy of 0.895 and a testing accuracy of 0.860, while the RF approach scored 0.883 and 0.853 and DF scored 0.850 (Ahammad et al., 2022). In addition, Almomani et al. collected semantic features including URL structure, HTML, JavaScript behaviors, and WHOIS metadata, and therefore evaluated 16 classifiers. The top accuracy scores for the models are around 97% for Gradient Boosting (GB) and RF (Almomani et al., 2022). Wei and Sekiya introduced deep learning methods into modelling and found that the models using Ensemble ML techniques outperformed others when they are applied to their datasets, even with reduced feature sets (Wei and Sekiya, 2022). Najjar-Ghabel et al. compared six ML models using a dataset of 47 features (content, behavior, domain info), which suggests RF model performing the best with 96.7% accuracy and high F1-score (Najjar-Ghabel et al., 2024).

Therefore, the objective of this paper is to evaluate the classification performance of DT, RF, and GB methods in the classification of phishing websites. The models are compared in terms of predictive accuracy, robustness across training and testing sets. The findings contribute to the development of more accurate ML-based phishing detection methods and their future application to web security tools.

2 METHODS

2.1 Data Source

This study utilizes a dataset from the Mendeley Data website, named Web Page Phishing Detection and published on 25 June 2021. The original dataset is built by Hannousse and Yahiouche based on the proposed guidelines. It contains 11430 groups of data and 89 variables, including raw URLs, status labels, and 87 extracted features (Hannousse and Yahiouche, 2021). Features are sorted into three categories: URL-based features, features based on page contents, and features extracted via external services. The original dataset is in .csv format.

2.2 Data Preprocessing

Due to the large size of original dataset, only 5000 observations (2500 legitimate, 2500 phishing) selected by stratified sampling is used for this study instead. To ensure the repeatability of the selection, the random seed is set to 42, while the same setting is employed for later feature selection and model tuning.

Several data preprocessing steps are therefore applied to the extracted sample. The statuses of websites are encoded (0 for legitimate, 1 for phishing). Features with constant values or no longer relevant (such as the raw URL and the web traffic feature based on the now-defunct Alexa ranking service) are removed. For external features concerning the domain registration length and the domain age, there are erroneous values found which are represented by negative values (-1/-2). To properly deal with them, the two features are first manually sorted into four categories: "error" (value < 0), "zero" (value = 0), "small_positive" (0 < value≤365, namely within a year), and "large_positive" (value > 365, namely longer than a year). Afterwards, these external-based features are encoded by the one-hot encoding (OHE) method.

2.3 Feature Selection

After cleaning and encoding the data, feature selection is performed via a model-based approach. The process contains the training of three individual classifiers based on DT, RF and GB techniques on the preprocessed dataset to extract feature importance. Afterwards, the top 25 informative features from each model are identified, and the final subset of selected features is determined by the intersection of the three selections to only remain features that are consistently important for all three models.

Therefore, the sample dataset used for this study contains 5000 observations and 18 variables (17 features and a label). The names and explanations of features belonging to three categories are shown in Table 1.

Table 1: Attribute Information

Variables	Explanation
length_url	full length of URL
length_hostname	hostname length of URL
nb_qm	number of occurrences of "?" in URL
nb_slash	number of occurrences of "/" in URL
nb_www	number of occurrences of "www" in URL
ratio_digits_host	ratio of digits in the hostname
length_words_raw	length of the shortest word in the hostname
char_repeat	number of character repeats in URL
longest_word_raw	length of the longest word in URL
avg_word_path	average length of words in path
phish_hints	total occurrence of sensitive words ("wp", "login", "includes", "admin", "content", "site", "images", "js", "alibaba", "css", "myacccount", "dropbox", "themes", "plugins", "signin", "view")
nb_hyperlinks	number of links in url web page contents
ratio_intHyperlinks	ratio of internal hyperlinks of web page
ratio_extRedirection	ratio of external redirections of web page
safe_anchor	number of unsafe anchors (e.g. "#", "javascript", "mailto")
google_index	Whether the webpage is indexed by Google, 1 for yes, 0 for no
page_rank	value of page rank via Openpagerank

2.4 Machine Learning Models

In this paper, the performance of three ML classification models based on Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) are compared. While all three methods are tree-based, DT is a more traditional approach in comparison to ensemble ML techniques RF and GB. The training and testing set for modelling are established with a ratio of 8:2. In Figure 1, Figure 2, and Figure 3, the general working flow charts of three tree-based models are illustrated.

For the DT classifier, which is less complex, grid search is applied for model tuning. However, as an ensemble learning method combines multiple learning algorithms, it is inherently more complicated than traditional methods. Thus, the employment of an exhaustive grid search for hypermeter tuning of RF and GB is time-consuming and unpractical.

Instead, for the RF and GB models, a two-stage hyperparameter tuning method combining grid search, randomized search and cross-validation is designed to balance efficiency and accuracy. In Stage 1, rather than testing all possible combinations as in grid search, RandomizedSearchCV samples a fixed number of random candidates (set to 250 in this study) to effectively save time. This stage aims to identify a relatively promising hyperparameter range that can be refined in the subsequent step. In Stage 2, based on the settings identified in Stage 1, a focused GridSearchCV is conducted to refine the tuning. In this stage, certain parameter ranges are narrowed around the best values found previously, while grid search is employed for less-sensitive or undefined parameters such as max features and criterion.

For both tuning methods, the searching criteria is set to be model accuracy in classification. Besides, 5-fold cross-validation is used for all tuning processes. By splitting the training data into five parts and loops through them, each training fold can be validated by five times. This helps avoid potential overfitting to a specific fold and can ensure a more generalized model.

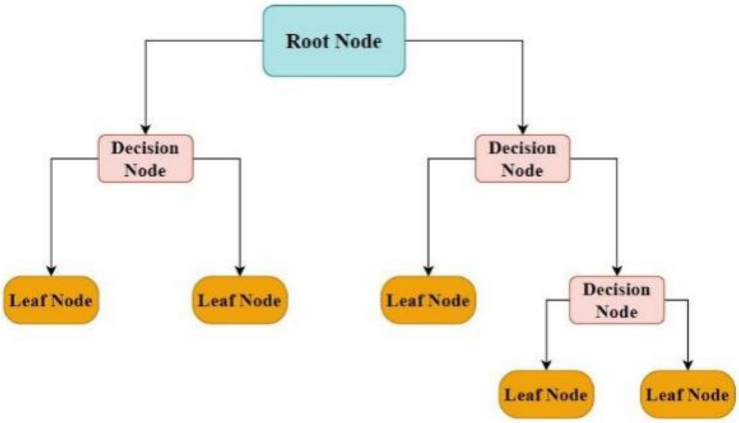

Figure 1: Flowchart of Decision Tree Classifier (Myles et al., 2010).

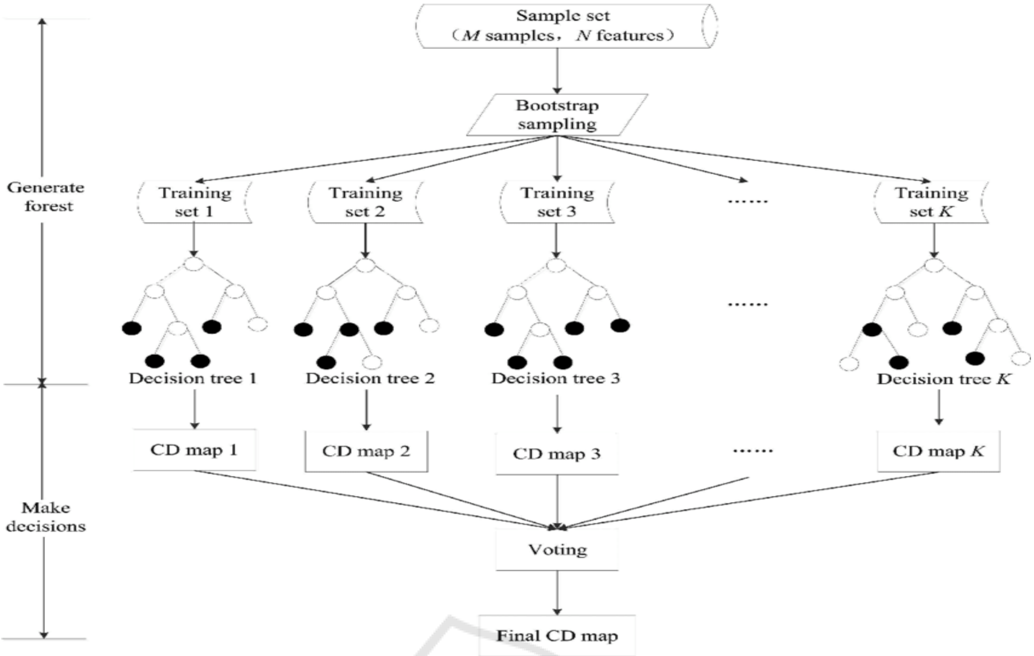

Figure 2: Flowchart of Random Forest Classifier (Feng et al., 2018).

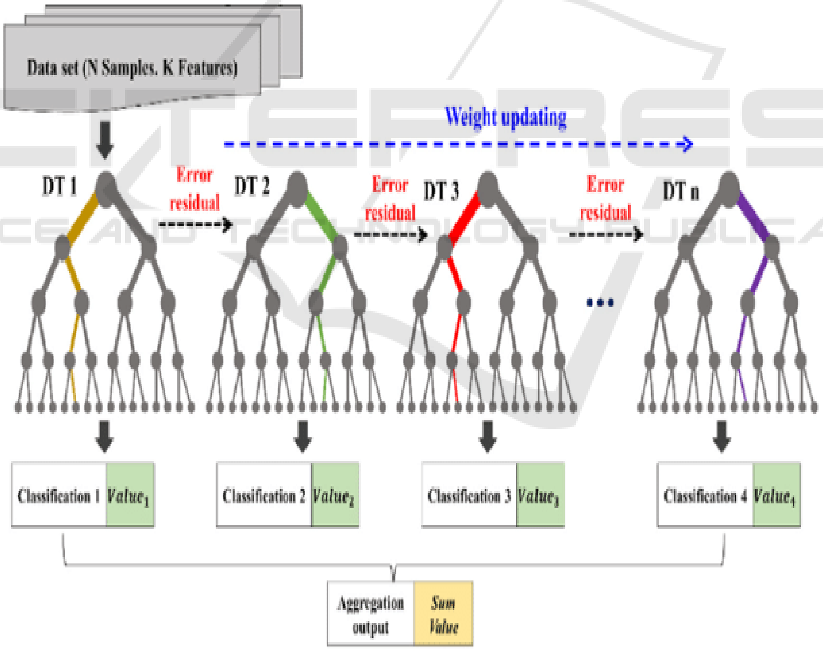

Figure 3: Flowchart of Gradient Boosting Classifier (Chen et al., 2022)

3 RESULT AND DISCUSSION

3.1 Descriptive Analysis

Figure 4 represents 6 of the total 17 histograms of the selected features. It can be seen these features display clear distinctions between phishing and legitimate websites. For instance, phishing URLs tend to be longer (Figure 4A, length_url), with more question marks and slashes (Figure 4B and 4C, nb_qm, nb_slash) and less likely to include “www” (Figure 4D, nb_www). Besides, the phishing websites exhibit more extreme values in some features such as phish_hints (Figure 4E) compared to legitimate ones, and the binary feature google_index (Figure 4F) can effectively distinguish the site for the majority of data. These distributional differences indicate that the selected features are suitable inputs for classification models. In addition, the diversity in feature types of the three categories suggests that the model can capture both superficial and structural aspects of phishing behavior.

Figure 5 illustrates the correlation heatmap. Several feature pairs, such as length_url and length_word_raw (with a correlation coefficient 0.79), have a significant correlation. While multicollinearity can be problematic for linear models due to its impact on coefficient stability, tree-based classifiers can effectively handle them without compromising predictive performance. Moreover, these models can capture non-linear interactions between features, making them suitable for modelling complex patterns in phishing detection.

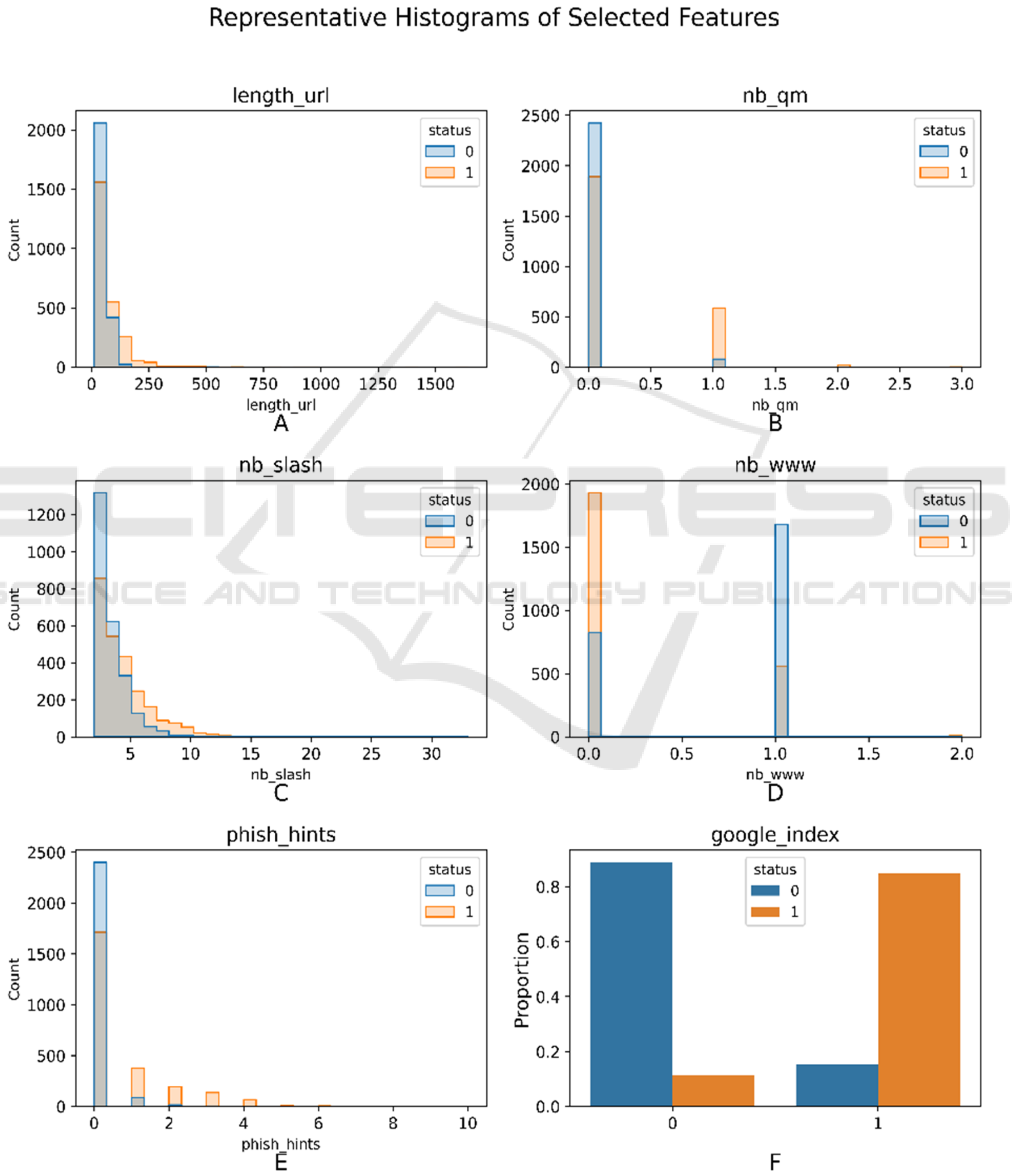

Figure 4: Representative Histograms of Selected Features (Picture credit: Original)

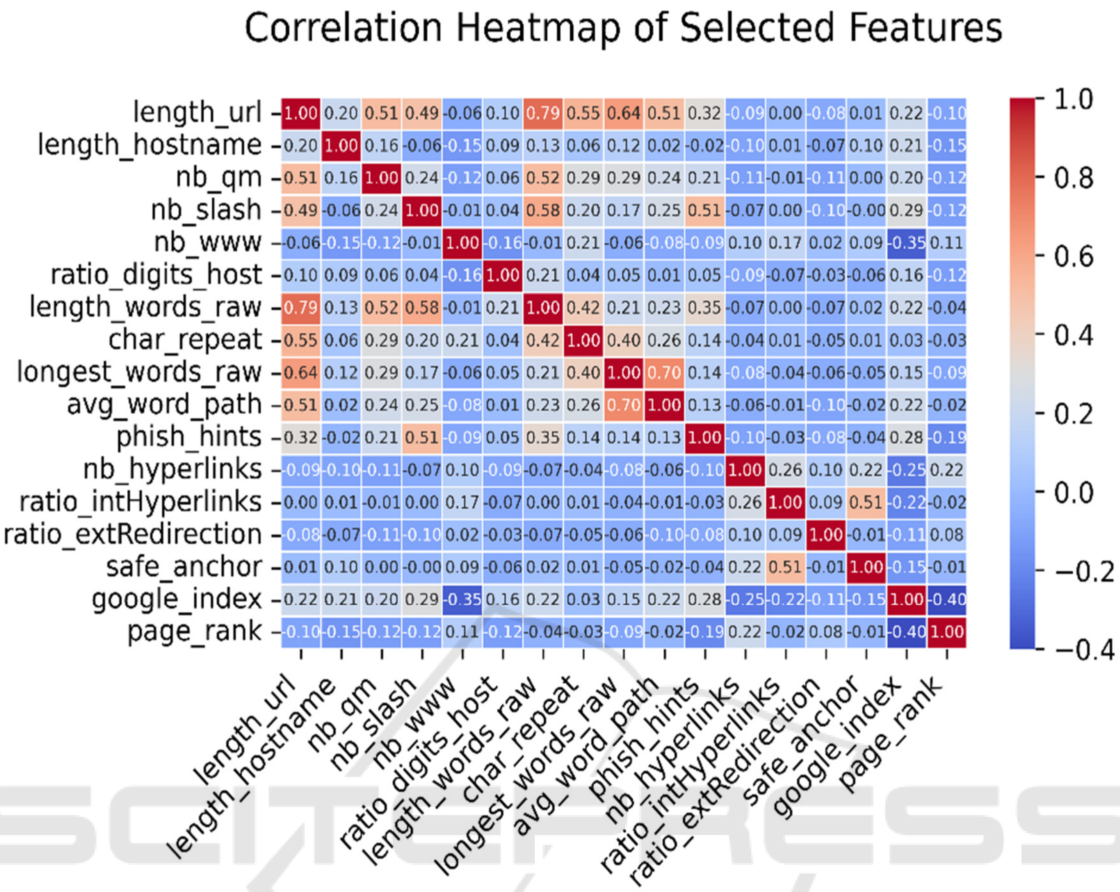

Figure 5: Correlation Heatmap of All Selected Features (Picture credit: Original)

3.2 Confusion Matrices and Evaluation Metrics

In this study, multiple metrics (accuracy, precision, recall and F1 score) for model performance evaluation are computed based on a 2 ∗ 2 confusion matrix. The structure of the confusion matrix and the calculation formula of each metric are provided in Table 2 and Equations (1) - (4) respectively.

Table 2: Confusion Matrix

		Prediction Label	
Actual Label		Positive (P)	Negative (N)
	Positive (P)	True Positive (TP)	False Negative (FN)
	Negative (N)	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TN+TP}{TP+FN+FP+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} = \frac{2TP}{2TP+FN+FP} \quad (4)$$

3.3 Model Evaluation

The confusion matrices on the train and test set corresponding to three ML models are illustrated in Figure 6 and Figure 7, with calculated evaluation metrics for train and test sets summarized in Table 3 and Table 4.

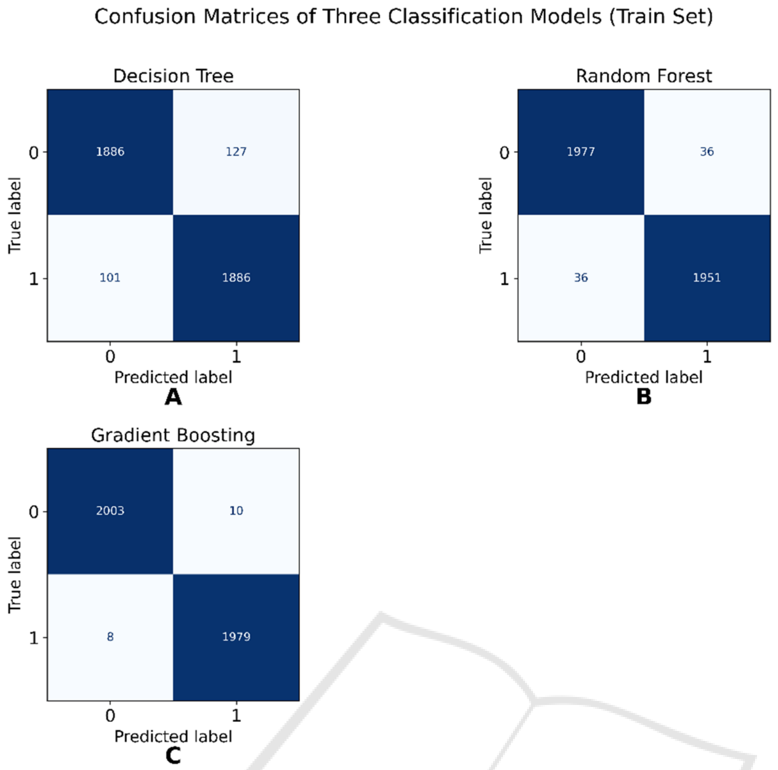

Figure 6: Confusion Matrices of Three Classification Models (Train Set) (Picture credit: Original)

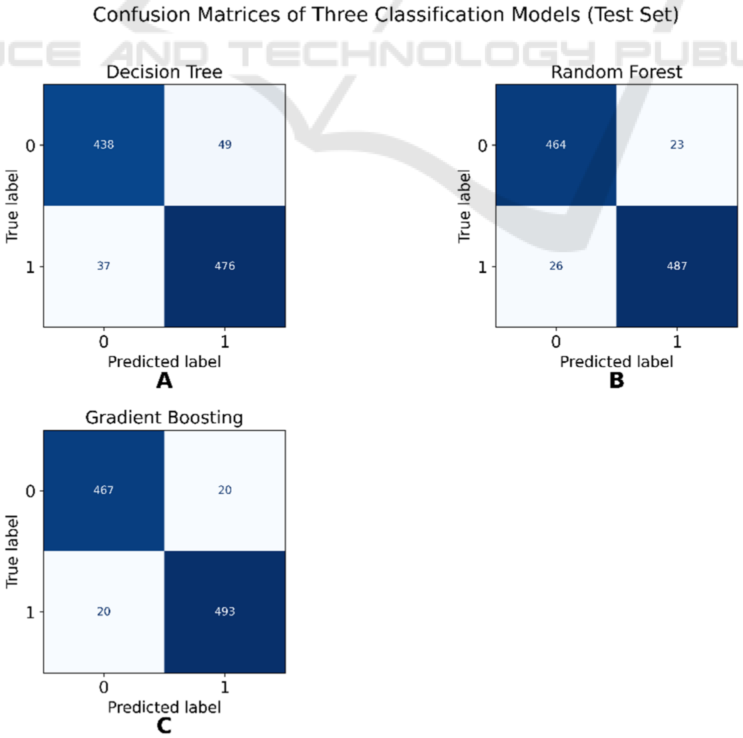

Figure 7: Confusion Matrices of Three Classification Models (Test Set) (Picture credit: Original)

Table 3: Evaluation Metrics (Train Set)

	Train Accuracy	Train Precision	Train Recall	Train F1 Score
Decision Tree	0.9425	0.9448	0.9391	0.9419
Random Forest	0.9800	0.9794	0.9804	0.9799
Gradient Boosting	0.9852	0.9874	0.9829	0.9851

From the calculated evaluation metrics presented in Table 3 and Table 4, the GB model achieves the best overall performance on both training and test sets. GB attains the highest accuracy (0.9600), precision (0.9592), and recall (0.9630) on the test set, indicating its capability to identify phishing websites accurately and capture most latent phishing websites effectively. The RF model follows closely behind, with comparable performance across all evaluation metrics (accuracy of 0.9540, precision of 0.9587 and recall of 0.9513). In contrast, the DT model performs significantly worse than the other two models. This illustrates the idea that ensemble learning methods are better classifiers in phishing website detection.

Table 4: Evaluation Metrics (Test Set)

	Test Accuracy	Test Precision	Test Recall	Test F1 Score
Decision Tree	0.9350	0.9462	0.9259	0.9360
Random Forest	0.9540	0.9587	0.9513	0.9550
Gradient Boosting	0.9600	0.9592	0.9630	0.9611

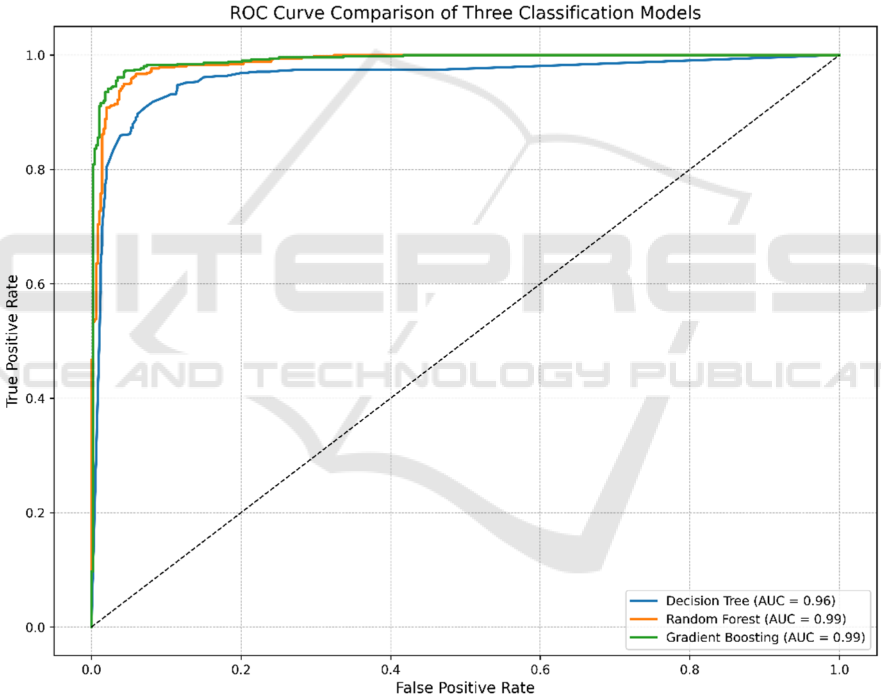

Figure 8: ROC Curve Comparison of Three Classification Models (Picture credit: Original)

These findings are further supported by the Receiver-operating characteristic (ROC) curves shown in Figure 8, where both GB and RF curves maintain closer to the top-left corner of the plot, demonstrating superior ability to distinguish between legitimate and phishing sites across various threshold settings, whereas the DT curve shows a relatively less steep pattern. The high AUC values (0.99 for GB and RF) confirm the strength of ensemble learning methods.

3.4 Model Performance Based on Learning Curves

The learning curves of three classification models are shown in Figure 9, where training and cross-validation accuracy are compared as the training set size increases. The initial upward slope of the training score in the DT model's learning curve reflects the model's potential underfitting on small datasets. By contrast, ensemble models RF and GB maintain better performance under limited data conditions. Both ensemble models show high training accuracy and steadily improving cross-validation performance as larger amount of data is used, with the GB model achieves the highest final validation accuracy among the three models. All three plots of learning curves, however, display signs of overfitting, which is characterized by the visible gap between the training score and the cross-validation scores.

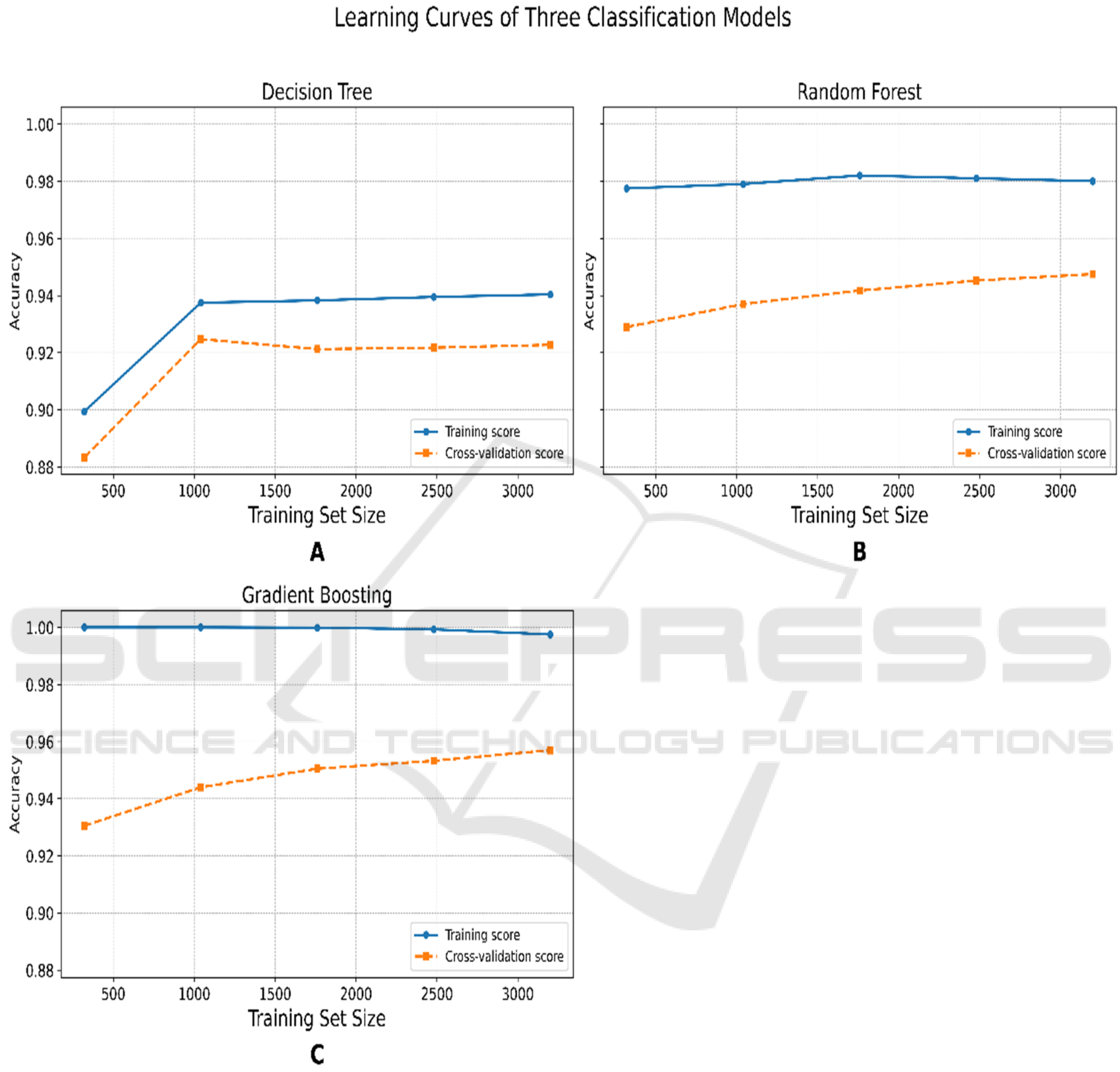

Figure 9: Learning Curves of Three Classification Models (Picture credit: Original)

3.5 Feature Importances of the Selected Model

For the most preferred GB model, which outperforms the other two approaches, external-based features of google index and page rank are highly important in the model according to Figure 10, which illustrates its feature importances. This leads to reduced efficiency of the model for off-line classification. Besides, external evaluation tools cannot always be fair and will turn out to be highly unreliable under deliberate manipulations. This result provides further insight into future studies, in which models less based on third-party resources can be trained and utilized for phishing websites classification.

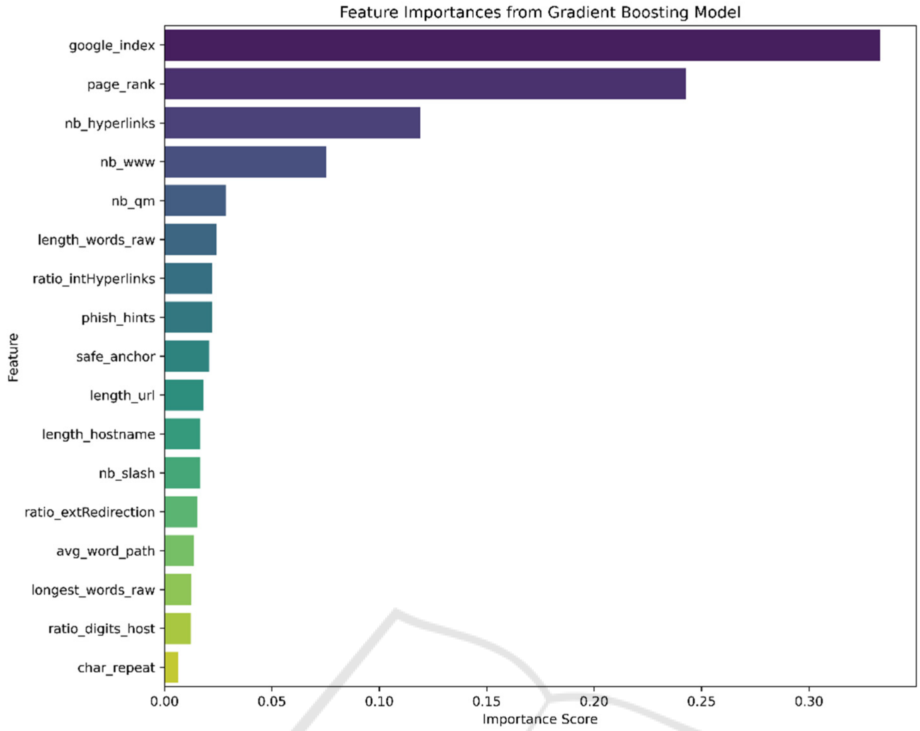

Figure 10: Feature Importance from Gradient Boosting Model (Picture credit: Original)

4 CONCLUSION

In conclusion, this study shows that the GB classification model has the best overall performance over the test set, with the RF model being a close alternative. Besides, the GB model exhibits better generalization with less severe overfitting. The results highlight the effectiveness of ensemble methods for classification tasks of phishing websites. Nonetheless, there still exist some limitations. All models show mild overfitting, which could reduce robustness in more generalized or real-world scenarios. Besides, the original dataset was last updated in 2021. Therefore, the model trained with older data does not necessarily remain valid in detecting and classifying up-to-date websites. In addition, the reliance on external-based features limits the model's use in offline detection settings and introduces potential risks of external manipulation. Therefore, future training of classification models needs to consider incorporating more up-to-date datasets, reducing dependence on third-party features, and applying pruning techniques to avoid overfitting. Furthermore, this study focuses solely on classical ML models based on the relevant Scikit-learn libraries. Thus, future studies can explore the building of alternative ensemble ML and deep learning models.

REFERENCES

Ahammad, S. H., et al. 2022. Phishing URL detection using machine learning methods. *Advances in Engineering Software,* 173, 103288.

Almomani, A., et al. 2022. Phishing website detection with semantic features based on machine learning classifiers. *International Journal on Semantic Web and Information Systems*, 18(1), 1-24.

Chen, J., et al. 2022. Machine learning-based classification of rock discontinuity trace: Smote oversampling integrated with GBT Ensemble Learning. *International Journal of Mining Science and Technology,* 32(2), 309-322.

Feng, W., et al. 2018. A novel change detection approach based on visual saliency and random forest from multi-temporal high-resolution remote-sensing images. *International Journal of Remote Sensing,* 39(22), 7998-8021.

Gupta, B. B., et al. 2021. A novel approach for phishing urls detection using lexical based machine learning in a real-time environment. *Computer Communications,* 175, 47-57.

Hannousse, A., Yahiouche, S. 2021. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Engineering Applications of Artificial Intelligence,* 104, 104347.

Karim, A., et al. 2023. Phishing detection system through hybrid machine learning based on URL. *IEEE Access,* 11, 36805-36822.

Kawale, M., et al. 2024. Machine learning based phishing website detection. *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom),* 833-837.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., Brown, S. D. 2010. An introduction to decision tree modeling. *Journal of Chemometrics,* 18(6).

Najjar-Ghabel, S., Yousefi, S., Habibi, P. 2024. Comparative analysis and practical implementation of machine learning algorithms for phishing website detection. *2024 9th International Conference on Computer Science and Engineering (UBMK),* 1-6.

Naqvi, B., et al. 2023. Mitigation strategies against the phishing attacks: A systematic literature review. *Computers &amp; Security,* 132, 103387.

Varshney, G., et al. 2024. Anti-phishing: A comprehensive perspective. *Expert Systems with Applications,* 238, 122199.

Wei, Y., Sekiya, Y. 2022. Sufficiency of ensemble machine learning methods for phishing websites detection. *IEEE Access,* 10, 124103-124113.