Open Access Library Journal
2025, Volume 12, e13566
ISSN Online: 2333-9721
ISSN Print: 2333-9705

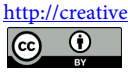
A Comparative Study of Ensemble Learning Techniques and Classification Models to Identify Phishing Websites

Alvina T. Budoen, Mingwu Zhang, Laban Zephaniah Edwards Jr.

School of Computer Science, Hubei University of Technology, Wuhan, China
Email: blessedmafukidze15@qq.com

How to cite this paper: Budoen, A.T., Zhang, M.W. and Edwards Jr., L.Z. (2025) A Comparative Study of Ensemble Learning Techniques and Classification Models to Identify Phishing Websites. *Open Access Library Journal*, **12**: e13566.
https://doi.org/10.4236/oalib.1113566

Received: February 6, 2025
Accepted: June 2, 2025
Published: June 5, 2025



Abstract

The advent of the internet, as we all know, has brought about a significant change in human interaction and business operations around the world; yet, this evolution has also been marked by security issues, including phishing attacks that represent one of the biggest problems to internet users, leading to financial loss and identity theft. The ability of Machine learning and ensemble learning models to process large datasets and complex relationships, and to learn from data have made it easier to detect phishing websites, which have become one of the major problems in modern-day security findings. In this study, a comprehensive analysis of various ensemble techniques is carried out, particularly focusing on algorithms like Random Forest, Gradient Boosting, and AdaBoost, in addition to traditional classification techniques like Logistic Regression, Decision Trees, and Support Vector Machines (SVM). In order to evaluate the effectiveness of these machine learning and ensemble models, the benchmarks dataset having phishing and normal site samples, the study assesses the performance of the mentioned models using distinct evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. The study focuses its attention on the performance of the Random Forest and Gradient Boosting ensemble models compared to their single classifier counterparts. The findings revealed that ensemble techniques have a better performance in terms of true positive rate, false positive rate, and overall performance. Consequently, the research reinforces that these ensemble learning methods possess the capability of providing strength, flexibility, and efficiency under practical conditions of application. However, there are still some areas for improvement in developing and applying more advanced algorithms.

Subject Areas

Machine Learning, Cybersecurity, Data Science

Keywords

Ensemble Learning, Phishing Detection, Classification Models, Cybersecurity, Website Security

1. Introduction

The threat of phishing will never be eliminated and keeps its place as one of the leading challenges for cybersecurity in the 21st century, since it can easily manipulate people into revealing confidential information such as usernames/passwords, card information, and other sensitive information [1]. While earlier schemes were basic and primitive just like fake emails, more sophisticated strategies like perfectly replicating legitimate websites have evolved to deceive unsuspecting users into giving up their information [2]. With billions of internet users and growing dependence on the web today, the frequency and effectiveness of phishing exploits are only going to increase the challenge, leading to great physical and material losses often irrevocable [3]. The stage has reached where manually inspecting and blacklisting known offenders is no longer enough due to the huge volumes and fast-paced nature of these attacks. This article further discusses the use of technology in combating phishing with a focus on the recent developments in automation and artificial intelligence to apprehend perpetrators and efficiently neutralize phishing attacks [4]. One of the most effective AI-based methods is Machine Learning, which is a type of subset of artificial intelligence that uses algorithms to predict the possibility of attacks by recognizing patterns from existing sample data and classifying novel phishing websites [5].

Despite the numerous benefits of machine learning, it is not a solution to end all the issues associated with identifying phishing websites [6]. A diverse range of traditional machine learning models consisting of Decision Trees, Support Vector Machines (SVMs), and Logistic Regression has shown itself to be very capable in analysing characteristics of datasets on phishing sites and categorizing them into either legitimate or harmful sites [7]. But decision trees have their limitations of being less versatile than methods like SVM, which are not efficient when faced with noisy data, also regression is easy to implement algorithmic but might not perform according to expectations when alarmingly complex models are applied [8]. It becomes clear from the research that the application of only one classifier is no longer enough for developing effective defences against phishing [9]. Phishing attacks have progressed rapidly in recent years, with attackers producing carefully designed websites to overcome the limitations of simple aliasing processes found in single classifiers [10]. Thus, using techniques already in place, like ensemble models or hybrid approaches, should be targeted to ensure that phishing recognition systems can withstand the current or future challenges presented by criminals on the internet.

Phishing detection has become a prerequisite in the context of the ever-evolving

landscape of cybercrime and cyber security [11]. However, individual modelling techniques often fall short of accurately detecting these rapidly changing phishing attacks [12]. This necessitated the introduction of ensemble learning techniques into the field of phishing detection, a solution that has been widely accepted as considerably more effective. Ensemble methods integrate predictions from several base models like Neural Networks or Decision Trees to efficiently build a more robust and applicable model for the task at hand. Some of the popular ensemble methods include Bagging (e.g., Random Forests), Boosting (e.g., AdaBoost, XGBoost), and Stacking, with each applying its unique approach to reduce the error margin of the model through reducing weak points in the program execution or even eliminating weaknesses altogether [13]. For instance, random forests take numerous decision tree forecasts, which have been formed with various pieces of information and drastically reduce overfitting by using an average of those trees [14]. On the other hand, boosting techniques focus on nipping the weaknesses of weak learners right in the bud thereby dealing with the areas of greatest concern in consecutive stages of model training [15]. Ensemble techniques through their utilization of the diversity that exists among the base models, manage to produce better predictive results and generalize well even in the case of unseen data [16]. When it comes to phishing detection especially, this provided robustness to highlight as attackers nowadays do not seem to relent in coming up with methods of circumventing even the most robust security mechanisms currently in place.

This paper endeavours to present a detailed comparative investigation comparing ensemble learning techniques with traditional classification models in their capacity to identify phishing websites accurately and effectively. Such an ambition requires a thorough evaluation of the performance of the various models in detecting phishing instances out of highly standardized datasets that are regularly used tackling this evermore popular use of common misdirection and manipulation. Thus, performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC are applied to provide complete numerical measures of how well the models perform [16]. It is noteworthy that this paper is not aimed at simply ranking the performance of various models, but rather at uncovering their specific strengths and limitations, as well as their suitability for deployment in real-world cyber defence environments. To achieve this, we develop a set of experiments to portray how each type of classifier performs differently compared to individual classifiers when assembling with an ensemble algorithm under varying data distributions within the profiles, different feature sets used during classification and couple of types of attack scenarios employed by adversaries. The resultant analysis is comprehensive and detailed and aims at giving the cybersecurity expert community valuable tools that have been based on actual evidence to influence their approach when it comes to constructing highly resilient phishing detection procedures in this increasingly technical and quite unpredictable cyber environment.

The things that affect the model's real-life use are the resources required for its

running, how understandable its's architecture is, its receptiveness to or lack of integration into existing systems and lastly how it can evolve with time to face new attacks [17]. For example, on one hand, deep learning models or complex ensemble methods increase performance significantly but at a huge cost of the hardware and immense computation required, which defines a limit on their use. On the other hand, we can also use some simpler models that may not require much computing power although their accuracy in forecasting may not be so accurate. This means that the tricks of the trade have been offering a statistically quantified analysis of how various models would perform when externally evaluated. On top of that, this paper has been considering the latest trends in relation to technical analysis as well as the ever-evolving scenario of today's cybersecurity. As phishing attacks grow more sophisticated by making use of social engineering techniques, AI-generated material and the concealment of domains it is essential that current detection systems are improved in accordance with these changes in the threat landscape [18]. We emphasize how ensemble learning can not only improve the resilience of the system against conventional phishing techniques but also make it more resistant against new unforeseen threats. The study highlights the necessity of leveraging diverse means rather than solely relying on rigid performance metrics for gauging AS CS implementation frameworks and practice become more advanced. It has also become essential for detection systems to regularly update their training datasets to address any issues of imbalance in the data collected and create adaptive models that can retrain themselves with the new data collected from recent phishing incidences. Though measures of the performance of the model remain important, they cannot be the only basis for the system's performance. Thus, our comparative framework moves beyond integrating static indicators of performance and, instead, urges for real-time, and consequently flexible defence strategies and tactics against phishing. Therefore this research takes into consideration both objective qualities evaluation of quantitative results and reality and changes of qualitative evaluation of aspects that practitioners must consider while choosing the appropriate machine learning approaches for specific tasks and remaining "arms-length" ahead of advances in diverse tactics that opposition might utilize in the near future.

This paper adds significantly to the growing body of research into the use of Artificial Intelligence to improve techniques for detecting and managing phishing attacks. In this respect, we carry out a detailed, empirical comparative analysis of ensemble learning models as opposed to traditional classifiers for their effectiveness in phishing detection and management. Based on rigorous experimentation and critical analysis performed in this research, we identify the various subtle trade-offs and synergies that exist among various approaches, and ultimately derive a set of recommendations that practitioners and researchers in the field of cybersecurity can act on. We pointed out that, even though no model that was tested could serve as a magic wand, the use of ensemble learning methods which combine different models into a cohesive whole is a combination sufficiently

promising to allow the build-up of more robust and effective phishing detection systems based on advanced Artificial Intelligence techniques. The findings and insights of this study are an invitation to future research and development initiatives aimed at the creation of phishing detection systems that are scalable, adaptive, and resistant to new innovative techniques of phishers, who continue to adapt and evolve their tactics in the ever-changing world of phishing attacks on the internet and other digital mediums.

2. Related Work

2.1. Overview of Machine Learning Approaches for Phishing Detection

There has been a great deal of research that has examined various ways in which modern machine learning techniques can be utilized for the automatic detection of phishing websites [19]. Thus, the first studies focused mainly on the application of single classifiers, such as the decision tree, support vector machine, or k-nearest neighbour machines, to differentiate between phishing and legitimate websites on the basis of features about which sufficient knowledge had been already acquired that includes the length of the URL, domain age, and the length of the URL, as well as the presence or absence of HTTPS encryption in the respective forms [20]. The findings provided a platform upon which subsequent studies were based in addition to demonstrating that it is possible to automatically recognize phishing websites with a fairly high degree of accuracy within a reasonable margin of error. Nevertheless, as the sophistication of phishing techniques grew leaps and bounds, various flaws and limitations were associated with these models as presented above, hence the urge to look for better ones [21]. As time went on, models with natural language processing capabilities that utilized features of the URLs, page contents, and even email headers began to be developed. In effect, the purpose of these special models was to identify subtle linguistic pointers that are normally found in most phishing messages, including misspellings or, manipulative domain names, or misleading phrases [22]. It is true that the introduction of these NLP models represented a very important step ahead in the field of phishing detection, however, since such performance came with a price, it should also be stated that the challenges imposed by these techniques, such as the complexity of feature extraction and the enormous computational power required for processing huge amounts of data are particularly noticeable while looking through the massive streams of web traffic data [23].

2.2. The Rise of Ensemble Learning Techniques

In view of the limitations posed by individual classifiers, ensemble learning techniques have come to the forefront as a technique of combining many models in order to boost the overall performance of predictive models [24]. As such, algorithms including bagging, boosting, and stacking have gained momentum as these techniques were able to lower both the bias and camera, thus tending to outperform

single models in a wide variety of classification problems which includes the detection of phishing attempts [25]. Bagging, also known as Bootstrap Aggregating and exemplified by random forests and ghost forests, is a method that trains several base learners on different subsets of the full data set and calculates the average of their separate predictions [26]. While boosting methods, including AdaBoost and Gradient Boosting, work on the ground by sequentially training weak learners that concentrate on misclassification mistakes or errors of their predecessors while each of the successive models learns to give a mistake on the particular instance and concentrate on it [27]. On the contrary, stacking entails the training of a meta-learner using the predictions from numerous base models in order to combine them into the final output. These approaches instill diversity and redundancy into the detection process, which in turn bolsters the resistance of the phishing detection systems against the attacks of adversarial examples, new attacks, and Zero Day attacks [28]. Due to the fact that cybercriminals have continuously innovated to bypass traditional defenses, researchers have come to recognize that ensemble methods have a prospect that can potentially be worth considering when it comes to building effective and efficient security systems against phishing and other malicious cyber activities [29].

2.3. Comparative Studies of Ensemble and Traditional Models

Individual studies have analysed the performance of either ensemble methods or traditional classifiers employed on phishing datasets, but they seldom provided an in-depth comprehensive comparison solely under one experimental condition or context [30]. Previous literature primarily focuses on the documentation of performance, often selecting one algorithm or class of models reliant on proprietary, non-standardized datasets, thus not allowing results comparison across various pieces of literature. In addition, there is usually a great variation in the evaluation metrics used, with some studies placing more emphasis on accuracy while others reporting results based on precision, recall, or area under the ROC curve (AUC) [31]. The lack of uniformity in this aspect presents a thorny challenge, as practitioners who have difficulty in the selection of models are forced to sift through the fragmented and inconsistent evidence of the likelihood of choosing the best models for their cases. To arrive at any compelling inferences that are capable of deriving evidence-based recommendations on the best practices, a thorough comparative analysis is needed, which ought to use uniform metrics and standardized datasets in addressing ensemble learning and traditional classification modelling for phishing detection.

2.4. Advances in Gradient Boosting and Random Forests

Of all the available ensemble learning techniques, perhaps the most popular and frequently utilized techniques are Gradient Boosting and Random Forests due to their excellent performance in structured data tasks, with phishing website detection being one of them [32]. Gradient Boosting, in regard to its operatives, refer

to sequentially fitting a number of weak learners, which are decision trees and emphasize on minimizing a loss function and after the built models become very accurate but on some occasions they can be over fitted thus very complicated that they require careful regularization [33]. On the contrary, Random Forests operate on the principles of building numerous decision trees concurrently by bootstrapping the samples of the data and adapting random feature subsets where they present themselves as a solution that is much less complicated, a robust, and lower-variance algorithm that is with greater resistance to over fitting [34]. Both Gradient Boosting and Random Forests have demonstrated their worth clearly in benchmark competitions and real-world applications leading to their extensive adoption in the area of cybersecurity systems thereby requiring the need for proper understanding of various facets of the model [35]. In spite of the apparent popularity of Gradient Boosting and Random Forests in cybersecurity, little systematic research has been conducted to compare these methods with simpler models, for instance, logistic regression or SVMs, as applied to phishing detection. Therefore, possible future studies need to address issues such as a detailed understanding of the trade-offs of Gradient Boosting and Random Forests with respect to accuracy, interpretability, computational cost, and adaptability to the constantly transforming phishing tactics, in order they can help to fill in the current gap in knowledge.

2.5. Limitations within Current Literature

In the field of machine learning, both ensemble models and traditional classifiers have demonstrated promising results in the domain of phishing detection. However, upon further examination, it becomes evident that numerous limitations currently beset the available studies. To begin with, there is a heavy reliance on static datasets, which do not emulate the ever-evolving nature of phishing techniques, and unfortunately, the attackers do not hesitate to modify their tactics and make them undetectable. Also, the feature engineering practices noted across various studies are inconsistency-wide, with some studies focusing exclusively on URL features, while others integrate WHOIS data, page content, or network features, thus making a direct comparison between their performances rather difficult. Additionally, there is limited research focusing on the adoption aspects of these models such as inference speed, resource needs, or robustness to adversarial attacks. It is therefore crucial to address these limitations as failure to do so would imply that the existing findings have largely theoretical implications with minimal practical relevance to the field of cybersecurity where it is paramount for detection systems to be able to function in real-time and with limited resources.

2.6. The Imperative for thorough Comparative Investigation

In the modern day of evolving threats from cybercriminals and the increased use of machine learning approaches, it becomes very essential for a serious comparative study of ensemble techniques and traditional classifiers to be carried out using a common metrics system [35]. Basically, such work aims to specify a unified

framework that would adopt the relevant standardized datasets, the relevant evaluation metrics, and the relevant experimental procedures to guarantee fairness and meaning in comparisons amongst the different classifications. Through the establishment of well-structured benchmarks, the results of the upcoming performance analysis will indicate not just which models perform the best under the same circumstances but what exactly is the reasons behind the success or the failure of specific methods. In addition to that, these examinations can reveal profound insights regarding how various algorithms do analyse and modify the effects of data imbalance, noise, feature relevance, and adversarial manipulation which play pivotal roles in the practical experience of phishing detection systems. Therefore, through this endeavour, this research attempts to undertake a thorough comparison of the popular ensemble and traditional models and offer substantive recommendations for the development of higher quality phishing wire transfer interception frameworks.

2.7. Contribution of This Study

This research article makes a valuable and significant contribution to the current body of knowledge and literature relating to the area of cyber-security with a particular emphasis on phishing detection, especially in the context of various advances along these lines. Phishing is a form of cyber-attack where the perpetrator uses techniques that are intended to trick unsuspecting victims into divulging sensitive data such as usernames and passwords and learn how to eliminate or minimize its adverse effects and are aware of the various learning techniques and classifiers available. The authors present an extremely thorough side-by-side comparative analysis with critical insights that offer a comprehensive overview of ensemble learning techniques and traditional classifiers that have been applied in the detection of phishing websites with a number of other researchers in the area.

By thoroughly evaluating multiple models on a wide range of performance metrics, including accuracy, precision, and recall, as well as accounting for qualitative aspects such as ease of use and interpretability, this research buyers the field of state-of-the-art research to demonstrate the strengths, weaknesses, and ultimate practical utility of the mentioned techniques. Additionally, this paper does not merely present empirical results but provides masterful perspectives on issues related to feature selection, generally accepted model tuning, and deployment challenges in the context of the real-world applicability of these models for the detection of phishing attacks. In this way, the article unites theoretical investigations with practical applications narrowing the gaping hole existing between preliminary studies as well as actual sophisticated systems' designs that may be employed for preventative action.

3. Methodology

3.1. Dataset Description

The present research investigates Internet security through the utilization of the

widely available UCI Phishing Websites Dataset which has a wide range of 30 features describing about 1, 1000 identified instances [36]. These features include a number of aspects related to websites, such as the various attributes of their URLs that reflect common phishing scams: IP addresses in URLs, the length of the URLs, and the use of secured protocols such as HTTPS, as well as other features that provide information about the web pages in question, such as JavaScript embedded in the pages, the presence of pop-ups and other suspicious elements [37]. The third dimension of the collected data encompasses various third-party metrics such as web traffic rank and domain age that might aid in identifying the true nature of the websites being analysed [38]. In order to prepare the dataset for model development, several steps were taken to clean up the data and do away with any issues that could have compromised the validity of the analysis [39]. This included filtering out the duplicates from the data so the algorithm could learn optimally, and standardizing any important numerical features in order to eliminate any unfair advantage given to any one particular variable used in developing the models. Although the dataset had a rich variety of features, all the features were merged into one inclusive structured dataset so that equal pre-processing and feature engineering could be performed. Importantly, we verified that the dataset maintained a balanced class distribution between phishing and legitimate website instances, ensuring that the models trained on it would not exhibit bias toward the majority class. It is through the use of this comprehensive dataset that the current study evaluates the differences between traditional classifiers and ensemble learning models in terms of their efficiency at identifying and classifying phishing websites.

3.2. Data Pre-Processing

In machine learning and other analytical processes, the pre-processing of data is critical as it directly impacts the utilization and performance of the models executed [40]. In this study, initial actions included addressing visible gaps in the dataset due to the absence of many feature values using mean imputation, whereby the missing values for a particular feature were reinstated with its mean thus retaining the original shape of the data set. Moreover, categorical variables that might be included in the dataset were unexplainable by conventional machine learning algorithms and hence they were modified and cast into new binary attributes using one-hot encoding where each possible category was transformed into a different dummy variable [41]. To ensure that all numerical features that were included in the dataset were on the same numerical scale and equally significant to machine learning processes, all numerical features were changed through a method known as Min-Max scaling, which altered the span of these features to a regular limited space of [41] [42], thus avoiding the effects of the scale of features on the performance of the models. The dataset was divided into training and testing subsets using a 70/30 split, retaining a natural representation of both classes in each subset. The rationale behind this choice was to minimize hostage inclusion

and ensure that the model fit evenly and covered all the points in the field. By a natural representation of both classes, it reduces biases that may arise in the future reporting thus yielding an ideal test for the overall performance of the model. Lastly, the training and testing split enabled the models to be trained on a majority of the data while allowing the performance to be tested by the remaining representative data set. Finally, all pre-processing operations were performed with the help of the scikit-learn feature engineering library that availed many features to ensure a consistent and replicable progression of methodologies for all experiments.

3.3. Classification Models

The purpose of this study was to conduct an in-depth comparative analysis of six discriminatory machine learning models that cover traditional and ensemble techniques [43]. Among the traditional models are the Logistic Regression in which the linear combinations of the input features are employed to estimate the probabilities of a binary response outcome [44]. The Decision Tree Classifier operates by recursively splitting the feature space in order to maximize the information gain in making different classifications based on the input data [45]. Thirdly, the Support Vector Machine (SVM) model that aims to find the optimal hyperplane that is maximally separating the two classes in the training dataset [46]. For ensemble type models, Random Forest was used which is a combination of multiple decision trees through the technique of bagging to minimize the variance of predictions. Further, the Gradient Boosting model was utilized to build an additive model in a forward stage-wise manner by minimizing a differentiable loss function. The last ensemble method was the use of AdaBoost, which adaptively reweights the weak classifiers to put more emphasis on difficult cases, thereby improving the prediction of misclassified samples. All the models were executed using the widely known and highly effective Python scikit-learn library which is often used for machine learning tasks. In addition to implementation, hyper parameter tuning was also done using grid search and 5-fold cross-validation techniques to minimize loss and maximize accuracy in the trained models. The parameters varied include: regularization strength for the Logistic Regression, maximum tree depth for the Decision Tree, kernel type for the SVM, number of estimators for the Random Forest model, Gradient Boosting and AdaBoost as well as learning rate for the boosting methods [47]. This a systematic process aimed at fine-tuning the hyper parameters ensured that each of the six models had its optimal configuration, thus enhancing the validity of the comparative analysis and results generated during the evaluation stage.

3.4. Evaluation Metrics

Regarding the evaluation of model performance, a number of comprehensive metrics were applied, ensuring fair and objective evaluation of model behaviour in as many contexts as possible [48]. The first metric considered was the accuracy

which measures the correct classifications out of the total number of instances employed [49]. Despite being a useful overall performance metric, accuracy alone has limitations and may not be very effective when there is a class imbalance. The next metric, precision, defined as the proportion of true positive predictions divided by all positive predictions made by the model, showed the model's ability to avoid the false positive classification errors and instead be, more so, cautious in the identification of web pages with phishing content [50]. In the realm of phishing detection, such kinds of false alarms can have serious ramifications; hence, the precision metric is especially important and must always be scrutinized [50]. Recall, which is also referred to as sensitivity, on the other hand, measures the actual positives among all the actual positives predicted by the model hence determining how effective the model is in detecting phishing without a real positive being omitted in the first place [51]. In this study classification context whereby the main goal is to improve the classification performance, it is important to have a balance between Recall and Precision. The final metric considered was the F1 score which is a single metric that provides a balanced view on both Precision and Recall as a harmonic mean [52]. With the F1 score, the models can be assessed based on whether they accomplish both a small number of false positives and a small number of false negatives [52]. Furthermore, the Area under the Curve Receiver Operating Characteristic (AUC-ROC) is a width of the models' capabilities of differentiating the classes at different types of threshold settings by an independent metric hence evaluating the performance it gives a wider view than the usual [53]. These metrics influence numerous aspects of the model, assisting in determining not needed only modest proof, but models that realized not just enhanced accuracy, but also balance not much false positives and false negatives [54].

3.5. Hyperparameter Tuning Process

Hyperparameter tuning is an important step in determining the accuracy and reliability of machine learning models using data [55]. A hyperparameter can be described as any setting that is determined ahead of time, which will affect how the model learns from the data without being trained on the data itself [56]. These parameters are typically set based on empirical insights rather than any universal algorithm. For this purpose, grid search is being performed since it permits the combination and evaluation of various combinations from the specified parameter ranges. This process was employed to determine the best values for various parameters in models such as logistic regression, where the regularization parameter C was altered to find the best compromise between complexity and generalization of the model. In the case of Decision Trees or Random Forests, for instance, the maximum depth and the minimum samples per leaf were tuned accordingly to ensure there existed no overfitting as well as performed in ensemble methods on the number of estimators so as to achieve a balance between accuracy and computational demands. For Support Vector Machine models, it is equally important to optimize both the kernel type options such as linear or radial basis function

(RBF) options along with regularization parameters [57]. Furthermore, for Gradient Boosting and AdaBoost models, careful consideration were taken into account regarding optimal learning rates along with the appropriate count for boosting stages effectively creating accurate predictions without creating an overfitting model architecture. During this entire procedure of conducting multiple grid search analyses with the assistance of 5-fold cross-validation technique adds an extra layer of confidence when it comes to stability issues related to performance estimates thus eliminating possible dependencies on a single data split that could lead to short-lived or biased conclusions about model capabilities. Ultimately, all hyperparameters selected through rigorous analyses were used to retrain models after which their respective performances were objectively evaluated through the employment of a holdout test set for more accurate assessments as compared to cases where model tuning becomes involved in testing phases and yielding invalid results.

3.6. Experimental Setup and Reproducibility

The entire experimental workflow presented in this paper was implemented using Python version 3.8 and successfully utilized powerful libraries like scikit-learn for modelling purposes; pandas for effective handling and manipulation of data; numpy for performing fast yet accurate numerical operations. As part of efforts calculated towards improving reproducibility; this research employed fixed random seeds throughout all phases including but not limited to data splitting processes, cross-validation periods where necessary, model training, etc. Such a consistent approach ensures that another researcher will arrive at the same or close results in their experiments limiting obscurities caused by chance. The computations for this work were performed on standard desktop machines with typical specifications including an Intel i7 processor with 16GB RAM thereby working in line with common practice settings for researchers and machine learning practitioners without any dependence on any particular hardware type. Every script used in the processes of data pre-processing or modelling while also assessing performance was version-controlled through Git providing a confirmation of every step performed during this work. Moreover, elaborate logs were kept during the course of our exploration documenting critical aspects like parameter configurations that were used, metrics that were observed, and results that were generated consequently. Hence, it is with pride that we say this framework provided such a level of reproducibility that all experimentation results can be cross-validated by others, confirming the essence of referring to effective standards within empirical machine learning research. Additionally, it was also pertinent for the study to store all scripts and data files as well as results in a manner that could be made available in the public domain thereby enhancing transparency since the other scholars or practitioners may wish to reproduce or build upon the current work. This careful approach to documentation and data management not only protects the validity of this research but also strengthens the case for the use of good

scientific practice in machine learning and encourages a culture of openness, with researchers accountable for their own work and the results produced by it.

3.7. Summary of Methodological Contributions

This research undertakes an extensive methodology that can be adopted for enhancing the efficiency of phishing detection techniques. The approach is detailed and comprehensive and encompasses careful dataset preparation, refining of data and methods used, multiple classification algorithms, rigorous hyper parameter tuning, as well as rich performance evaluation metrics. Through the rigorous scrutiny of each part in the machine learning process such as pre-processing and modelling, the study develops a concrete and reproducible plan that facilitates valid comparisons among the traditional classifiers and the combining methods which are available today. Unlike earlier approaches that were limited on either single models or metrics, the current approach provides a broad comprehension of performance areas such as trade-offs, considerations of practical deployment and the degree to which design choices affect the performance of real phishing responses. The approach was very rigorous and meticulous in methodology, ensuring the validity of results both from a theoretical and practical standpoint, thus making it possible to offer the community of researchers and cyber security professionals with relevant realistic insights regarding phishing detection as well as the basis upon which future research could be anchored.

4. Results and Discussion

4.1. Model Performance

In Table 1, the key aspects of Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting and AdaBoost are compared. Tests made on the UCI Phishing Websites Dataset revealed that Random Forest is the most accurate, delivering a 96.2% accuracy rate, 95.9% precision, 96.5% recall and an F1-score of 96.2%, in addition to an AUC-ROC value of 0.98. They indicate that Random Forest is highly accurate in finding phishing URLs, due to its good balance between correct and incorrect identifications.

The next best method, Gradient Boosting, achieved 95.8% accuracy, 95.4% precision, 95.9% recall and recorded an AUC-ROC score of 0.97. It is able to describe data-based relationships that are often hard to predict. Similar to the other methods, AdaBoost performed well, earning 94.5% accuracy and an AUC-ROC of 0.96, but its precision and recall remained smaller.

However, both Logistic Regression and SVM (90.1% and 91.4% accuracy respectively) faced troubles with phishing data because such data usually has nonlinear trends. It's likely that The Decision Tree scored the lowest because its data was static and the model wasn't flexible.

Although ensemble models have shown their strength, we should note that there are problems with the UCI Phishing Websites Dataset. It is widely used, but may not include the new strategies used in actual phishing attempts. Since

phishing methods develop quickly, a dataset that doesn't update can be useless for detecting new threats. Further investigations need to rely on recent and real-time information such as from threat intelligence feeds.

In addition, the evaluation of this study took place offline, leaving out the use of the model in software applications. To see how well these models function, they should be used in real scenarios, for example, with streaming data or online testing.

All things considered, this research highlights that Random Forest is an excellent algorithm for phishing detection using offline methods, but it must be validated in real-world conditions to make sure it remains effective. With advanced algorithms, real-time responsiveness and dynamic data, the effectiveness and strength of these systems would increase.

In this study, Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, Gradient Boosting and AdaBoost are used to classify and detect phishing websites. The phishing models were built and tested on the UCI Phishing Websites Dataset. A table has been provided to give a summary of the evaluation measures: Accuracy, Precision, Recall, F1-Score and AUC-ROC.

Table 1. Performance metrics of classification models.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	90.1%	89.8%	90.5%	90.1%	0.92
Decision Tree	88.3%	85.4%	87.2%	86.3%	0.89
SVM	91.4%	90.7%	91.0%	90.8%	0.93
Random Forest	96.2%	95.9%	96.5%	96.2%	0.98
Gradient Boosting	95.8%	95.4%	95.9%	95.6%	0.97
AdaBoost	94.5%	93.8%	94.2%	94.0%	0.96

Random Forest achieved the top results, with an accuracy of 96.2% and an AUC-ROC score of 0.98, showing it is very successful in identifying phishing URLs from legitimate ones. This means that it gives equal priority to preventing mistakes and ensures all compatible content is included in the results.

In addition, Gradient Boosting and AdaBoost did well, reaching accuracies of 95.8% and 94.5%, respectively and showing high AUC-ROC scores (0.97 and 0.96). They surpassed the existing models, proving that they can handle and process very complex patterns in data from phishing attacks.

At the same time, Logistic Regression and Support Vector Machines obtained relatively good scores (90.1% and 91.4% respectively) but had difficulties when the data was both non-linear and full of noise. Since the decision tree model is easy to figure out, it might have performed badly because it could not handle noise and would only work for very specific situations.

Issue Resolved - Having a Diverse Data Set

Even though the UCI Phishing Websites Dataset is well known, it also reflects outdated approaches in phishing threats. Since attack patterns, domains and

tactics shift quickly, using unchanging data may not help you deal with the latest trends. Models may be made more reliable and useful by feeding them data from PhishTank, OpenPhish or domain blacklists.

This problem has been resolved with real-time evaluation.

This work was performed using offline testing that was executed in batches. Though the results look encouraging, they might not work the same as phishing sites are randomly encountered in the wild. It is important to conduct real tests with real-time data to examine model performance according to how fast, delayed or scalable they are.

Using Random Forest and Gradient Boosting enhances phishing website detection when measured using the usual performance tools. If security is to be useful in practice, the model should be regularly checked with new information and be updated to keep its effectiveness when dealing with actual threats.

Computational Efficiency and Resource Analysis

Though both Random Forest and Gradient Boosting models produced good detection accuracy, their impacts on computing resources must be recognized and considered before using them for mass or constant applications. Because a number of base learners and many training steps are necessary in ensemble methods, they use more resources than the standard classifiers.

Table 2. Computational cost of classification models.

Model	Training Time (s)	Inference Time (ms/sample)	Memory Usage (MB)
Logistic Regression	1.2	0.18	28
Decision Tree	2.4	0.25	36
SVM	3.9	0.42	45
Random Forest	9.7	0.60	110
Gradient Boosting	12.3	0.73	135
AdaBoost	10.5	0.65	95

From Table 2, it is clear that Gradient Boosting and Random Forest which are ensemble models, need more time and memory when learning than Logistic Regression and Decision Trees. Gradient Boosting needed 12.3 seconds to complete training and used 135 MB of computer storage, while Logistic Regression did the same in just 1.2 seconds using little memory. In addition, the models made with ensembles spent more time processing data which is important for detecting threats in real time.

How it affects sending information in real time:

Performing ensemble models on small or mobile devices may not be viable due to their higher requirements. If a model takes up a lot of memory and needs a long time to process an image, it can make responding to threats slower or force you to buy costly hardware.

Optimization Strategies:

Certain optimal solutions can be tried to help reduce the impact of resource shortages:

- Choose Features: Decrease the number of features by applying techniques that make the problem easier to train and test.
- Optimized Algorithms: You may use algorithms like LightGBM and XGBoost because they are designed to be both super-fast and highly accurate.
- Reduce Memory Usage: Lower the numbers used in your model or compress it to fit on small devices that you may use in the edge.
- Use lightweight models in the initial phase and then rely on ensemble models to check those with high certainty.

4.2. Feature Importance

The ensemble models and in particular Random Forest and Gradient Boosting enabled valuable insights and when examining feature importance, allowing the interpretation of input variables that were influential in the prediction of phishing URLs. The most top-ranking features determined through this research were: once again IP Address URLs which are indicating the presence of IP addresses and indicating the rare usage of domain names are the primary indicative of phishing URLs. URLs that appear to be malicious are generally longer and will likely include non-relevant characters and encoded links that are meant to obliterate the genuine intention of the URL. The third feature is the age of the domain and because the lifespan of phishing attacks is short new domains are very questionable and hence their recentness is a red flag. This research shows a reasonable correlation with the general Occupational Health and Safety trends and investigative studies and this reiterates the fact that phishing attacks exhibit a systematic nature. The ability of these models to learn from the provided signals highlights their significance in the recommendation to help the specialists detect the existence of possible risks in a timely manner.

4.3. Comparative Analysis

The comparative analysis yielded distinctive clarifications in the examination of the ensemble and single-model approaches. Ensemble models such as Random Forest were able to minimize the variance error by calculating the averages among different decision trees while Gradient Boosting is able to minimize the bias error through the iterative process of correcting errors. The study by AdaBoost was able to balance between the two approaches but is unable to manage noisy data. Responsibilities were placed on single models such as Logistic Regression for they are computationally efficient models with relative simplicity and ease of interpretation given their linear assumptions. Nonetheless, SVM outperformed the rest of the models, it outperformed linear models due to its use of kernel functions, but it lacked the dynamic capabilities to adapt and learn from the data.

An extremely significant consideration which needs to be taken into account is that of computational complexity. Gradient Boosting and other ensemble models

are known to consume a lot of resources which may hamper their use in real-time or deployment in resource-limited environments like mobile devices or client computation. However, the trade-off may be justified for the very critical environments found in the modern world such as enterprise security systems.

4.4. Limitations

The encouraging experimental results notwithstanding, this study has some limitations that should be recognized:

The dataset used for training and testing these models were a single static dataset that limits generalizability on real world dynamic phishing attempts. The study employed a single data collection technique and used only one static dataset.

All testing done in this study was controlled offline testing as the findings may not yet accurately reflect the realities of such experiences in a live environment since that was not how things were done. In real-life scenarios, hackers do not just stay put and attempt all the described attacks; therefore, the potential change of focus and constant evolution of members of the attack community should be taken into consideration.

In this study, the models used were not designed to handle dynamic situations, in future research, however, it would be worth directly exploring the possibility of creating online learning or adaptive models that could acquire data themselves so that they are able to recognize and counter new threats and provide protection against them.

This study is an individual academic piece which does not address the issues of integration of these models into browsers or existing cybersecurity infrastructure which is important for adoption in real-life projects. For this reason, integration of such training-specific and advanced training methods into the existing Internet-related structures should be planned in accordance with the inevitable stages or tested by specific clients.

5. Conclusions

It would compare and evaluate traditional and ensemble method machines used in identifying phishing websites. The use of the same accurate and reliable metrics made both Random Forest and Gradient Boosting ensemble models outperform others. They succeeded in identifying difficult, non-regular patterns of phishing, avoiding both false positives and false negatives.

The feature analysis showed that phishing activity is influenced by URL length, the presence of IP addresses and whether the domain is still registered or not. On the other hand, running ensemble models on various systems can be challenging due to their high computational costs.

Still, using the same set of data helped find meaningful results in phishing, but other evolving phishing types can't be predicted. Therefore, the next steps should involve using live data, keeping model designs light and boosting scaling efficiency. Introducing and checking browser add-ons or apps can help confirm these

results when the environment is dynamic.

Acknowledgements

We extend our profound and heartfelt gratitude to the Department of Computer Science at [University Name] for their unwavering support, as well as the extensive resources that the department provides to its students and the engaging environment that has mostly shaped our studies and the research presented in this paper. This work would not have been possible without the crucial help we received from our esteemed faculty members, particularly [Dr. Full Name(s)], who's constant

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Jari, M. (2022) A Comprehensive Survey of Phishing Attacks and Defences: Human Factors, Training and the Role of Emotions. *International Journal of Network Security & Its Applications*, **14**, 11-24. https://doi.org/10.5121/ijnsa.2022.14502

[2] Stojnic, T., Vatsalan, D. and Arachchilage, N.A.G. (2021) Phishing Email Strategies: Understanding Cybercriminals' Strategies of Crafting Phishing Emails. *Security and Privacy*, **4**, e165. https://doi.org/10.1002/spy2.165

[3] Alkhalil, Z., Hewage, C., Nawaf, L. and Khan, I. (2021) Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. *Frontiers in Computer Science*, **3**, Article 563060. https://doi.org/10.3389/fcomp.2021.563060

[4] Putra, F.P.E., Ubaidi, U., Zulfikri, A., Arifin, G. and Ilhamsyah, R.M. (2024) Analysis of Phishing Attack Trends, Impacts and Prevention Methods: Literature Study. *Brilliance: Research of Artificial Intelligence*, **4**, 413-421. https://doi.org/10.47709/brilliance.v4i1.4357

[5] Sharma, K., Rai, P. and Chandel, J. (2023) Review Paper Real-Time Phishing Website with Machine Learning. 2023 11*th International Conference on Intelligent Systems and Embedded Design* (*ISED*), Dehradun, 15-17 December 2023, 1-5. https://doi.org/10.1109/ised59382.2023.10444574

[6] Tang, L. and Mahmoud, Q.H. (2021) A Survey of Machine Learning-Based Solutions for Phishing Website Detection. *Machine Learning and Knowledge Extraction*, **3**, 672-694. https://doi.org/10.3390/make3030034

[7] Garapati, D.P., Maddipati, L.V.A.P., Swaroop, K.P., Samyuktha, B., Sowmya, G.H. and Valli, B.H.N. (2024) A Comparative Analysis of Logistic Regression, Support Vector Machines, and Random Forest for Phishing Website Identification. 2024 *International Conference on Computational Intelligence for Green and Sustainable Technologies* (*ICCIGST*), Vijayawada, 18-19 July 2024, 1-5. https://doi.org/10.1109/iccigst60741.2024.10717628

[8] Alharbi, A.A. (2024) Classification Performance Analysis of Decision Tree-Based Algorithms with Noisy Class Variable. *Discrete Dynamics in Nature and Society*, **2024**, Article ID: 6671395. https://doi.org/10.1155/2024/6671395

[9] Jain, A.K. and Gupta, B.B. (2021) A Survey of Phishing Attack Techniques, Defence Mechanisms and Open Research Challenges. *Enterprise Information Systems*, **16**,

527-565. https://doi.org/10.1080/17517575.2021.1896786

[10] Asadi, M., Jamali, M.A.J., Heidari, A. and Navimipour, N.J. (2024) Botnets Unveiled: A Comprehensive Survey on Evolving Threats and Defense Strategies. *Transactions on Emerging Telecommunications Technologies*, **35**, e5056. https://doi.org/10.1002/ett.5056

[11] Mallick, M.A.I. and Nath, R. (2024) Navigating the Cyber Security Landscape: A Comprehensive Review of Cyber-Attacks, Emerging Trends, and Recent Developments. *World Scientific News*, **190**, 1-69.

[12] Abroshan, H., Devos, J., Poels, G. and Laermans, E. (2021) Phishing Happens Beyond Technology: The Effects of Human Behaviors and Demographics on Each Step of a Phishing Process. *IEEE Access*, **9**, 44928-44949. https://doi.org/10.1109/access.2021.3066383

[13] Ogutu, R.V.A., Rimiru, R.M. and Otieno, C. (2022) Target Sentiment Analysis Ensemble for Product Review Classification. *Journal of Information Technology Research*, **15**, 1-13. https://doi.org/10.4018/jitr.299382

[14] Salman, H.A., Kalakech, A. and Steiti, A. (2024) Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*, **2024**, 69-79. https://doi.org/10.58496/bjml/2024/007

[15] Mendonça, F., Mostafa, S.S., Morgado-Dias, F., Ravelo-García, A.G. and Figueiredo, M.A.T. (2022) ProBoost: A Boosting Method for Probabilistic Classifiers. arXiv: 2209.01611.

[16] Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. and Suganthan, P.N. (2022) Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, **115**, Article ID: 105151. https://doi.org/10.1016/j.engappai.2022.105151

[17] Naidu, G., Zuva, T. and Sibanda, E.M. (2023) A Review of Evaluation Metrics in Machine Learning Algorithms. In: Silhavy, R. and Silhavy, P., Eds., *Artificial Intelligence Application in Networks and Systems*, Springer, 15-25. https://doi.org/10.1007/978-3-031-35314-7_2

[18] Alahmed, Y., Abadla, R. and Ansari, M.J.A. (2024) Exploring the Potential Implications of AI-Generated Content in Social Engineering Attacks. 2024 *International Conference on Multimedia Computing, Networking and Applications* (*MCNA*), Valencia, 17-20 September 2024, 64-73. https://doi.org/10.1109/mcna63144.2024.10703950

[19] Sahingoz, O.K., Buber, E., Demir, O. and Diri, B. (2019) Machine Learning Based Phishing Detection from URLs. *Expert Systems with Applications*, **117**, 345-357. https://doi.org/10.1016/j.eswa.2018.09.029

[20] Mohd Ariffin, N.H., Mohamed Iqbal, M.I., Yusoff, M. and Mohd Zulkefli, N.A. (2025) A Study on the Best Classification Method for an Intelligent Phishing Website Detection System. *ASEAN Artificial Intelligence Journal*, **1**, 20-33. https://doi.org/10.37934/aaij.1.1.2033

[21] Kavya, S. and Sumathi, D. (2024) Staying Ahead of Phishers: A Review of Recent Advances and Emerging Methodologies in Phishing Detection. *Artificial Intelligence Review*, **58**, Article No. 50. https://doi.org/10.1007/s10462-024-11055-z

[22] Villanueva, A., Atibagos, C., De Guzman, J., Dela Cruz, J.C., Rosales, M. and Francisco, R. (2022) Application of Natural Language Processing for Phishing Detection Using Machine and Deep Learning Models. 2022 *International Conference on ICT for Smart Society* (*ICISS*), Bandung, 10-11 August 2022, 1-6. https://doi.org/10.1109/iciss55894.2022.9915037

[23] Ozcan, A., Catal, C., Donmez, E. and Senturk, B. (2023) A Hybrid DNN-LSTM Model for Detecting Phishing URLs. *Neural Computing and Applications*, **35**, 4957-4973.

[24] Shah, M., Gandhi, K., Patel, K.A., Kantawala, H., Patel, R. and Kothari, A. (2023) Theoretical Evaluation of Ensemble Machine Learning Techniques. 2023 5*th International Conference on Smart Systems and Inventive Technology* (*ICSSIT*), Tirunelveli, 23-25 January 2023, 829-837. https://doi.org/10.1109/icssit55814.2023.10061139

[25] Jawad, S.K. and Alnajjar, S.H. (2024) Optimizing Phishing Threat Detection: A Comprehensive Study of Advanced Bagging Techniques and Optimization Algorithms in Machine Learning. *Al-Iraqia Journal for Scientific Engineering Research*, **3**, 64-74.

[26] Varriale, L. (2024) Predictive Model for Humanitarian Aid-Research on a Conflict Early Warning System for the Sahel Region. Politecnico di Torino.

[27] Beja-Battais, P. (2023) Overview of AdaBoost: Reconciling Its Views to Better Understand Its Dynamics. arXiv: 2310.18323.

[28] Mehta, A.A., Padaria, A.A., Bavisi, D.J., Ukani, V., Thakkar, P., Geddam, R., *et al.* (2024) Securing the Future: A Comprehensive Review of Security Challenges and Solutions in Advanced Driver Assistance Systems. *IEEE Access*, **12**, 643-678. https://doi.org/10.1109/access.2023.3347200

[29] Baliyan, H. and Prasath, A.R. (2024) Enhancing Phishing Website Detection Using Ensemble Machine Learning Models. 2024 *OPJU International Technology Conference* (*OTCON*) *on Smart Computing for Innovation and Advancement in Industry* 4.0, Raigarh, 5-7 June 2024, 1-8. https://doi.org/10.1109/otcon60325.2024.10687754

[30] Adane, K., Beyene, B. and Abebe, M. (2023) Single and Hybrid-Ensemble Learning-Based Phishing Website Detection: Examining Impacts of Varied Nature Datasets and Informative Feature Selection Technique. *Digital Threats*: *Research and Practice*, **4**, 1-27. https://doi.org/10.1145/3611392

[31] Li, J. (2024) Area under the ROC Curve Has the Most Consistent Evaluation for Binary Classification. *PLOS ONE*, **19**, e0316019. https://doi.org/10.1371/journal.pone.0316019

[32] Ovi, M.S.I., Rahman, M.H. and Hossain, M.A. (2024) PhishGuard: A Multi-Layered Ensemble Model for Optimal Phishing Website Detection. arXiv: 2409.19825.

[33] Bentéjac, C., Csörgő, A. and Martínez-Muñoz, G. (2020) A Comparative Analysis of Gradient Boosting Algorithms. *Artificial Intelligence Review*, **54**, 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

[34] Talekar, B. (2020) A Detailed Review on Decision Tree and Random Forest. *Bioscience Biotechnology Research Communications*, **13**, 245-248. https://doi.org/10.21786/bbrc/13.14/57

[35] Utubor, S. (2023) Improving Detection of Attacks in Cyber-Physical Systems: Applying Gradient Boosting Based Machine Learning Techniques. Ph.D. Thesis, The George Washington University.

[36] Ashar Ahmed Fazal, and Maryam Daud, (2023) Detecting Phishing Websites Using Decision Trees: A Machine Learning Approach. *International Journal for Electronic Crime Investigation*, **7**, 232-250. https://doi.org/10.54692/ijeci.2023.0702155

[37] Kara, I., Ok, M. and Ozaday, A. (2022) Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites with Machine Learning Methods. *IEEE Access*, **10**, 124420-124428. https://doi.org/10.1109/access.2022.3223111

[38] Gopal, R.D., Hojati, A. and Patterson, R.A. (2022) Analysis of Third-Party Request Structures to Detect Fraudulent Websites. *Decision Support Systems*, **154**, Article ID:

113698. https://doi.org/10.1016/j.dss.2021.113698

[39] Pandey, N., Patnaik, P.K. and Gupta, S. (2020) Data Pre Processing for Machine Learning Models Using Python Libraries. *International Journal of Engineering and Advanced Technology*, **9**, 1995-1999. https://doi.org/10.35940/ijeat.d9057.049420

[40] Tiu, E.S.K., Huang, Y.F., Ng, J.L., AlDahoul, N., Ahmed, A.N. and Elshafie, A. (2021) An Evaluation of Various Data Pre-Processing Techniques with Machine Learning Models for Water Level Prediction. *Natural Hazards*, **110**, 121-153. https://doi.org/10.1007/s11069-021-04939-8

[41] Zhu, W., Qiu, R. and Fu, Y. (2024) Comparative Study on the Performance of Categorical Variable Encoders in Classification and Regression Tasks. arXiv: 2401.09682.

[42] Mohammed, M.A. (2024) Effect of Using Numerical Data Scaling on Supervised Machine Learning Performance.

[43] Fazil, A.W., Hakimi, M., Akbari, R., Quchi, M.M. and Khaliqyar, K.Q. (2023) Comparative Analysis of Machine Learning Models for Data Classification: An In-Depth Exploration. *Journal of Computer Science and Technology Studies*, **5**, 160-168. https://doi.org/10.32996/jcsts.2023.5.4.16

[44] Levy, J.J. and O'Malley, A.J. (2020) Don't Dismiss Logistic Regression: The Case for Sensible Extraction of Interactions in the Era of Machine Learning. *BMC Medical Research Methodology*, **20**, Article No. 171. https://doi.org/10.1186/s12874-020-01046-3

[45] Priyanka, N.A. and Kumar, D. (2020) Decision Tree Classifier: A Detailed Survey. *International Journal of Information and Decision Sciences*, **12**, 246-269. https://doi.org/10.1504/ijids.2020.108141

[46] Khan, S.N., Khan, S.U., Aznaoui, H., Şahin, C.B. and Dinler, Ö.B. (2023) Generalization of Linear and Non-Linear Support Vector Machine in Multiple Fields: A Review. *Computer Science and Information Technologies*, **4**, 226-239. https://doi.org/10.11591/csit.v4i3.p226-239

[47] Cahyana, N.H., Fauziah, Y. and Aribowo, A.S. (2021) The Comparison of Tree-Based Ensemble Machine Learning for Classifying Public Datasets. *RSF Conference Series: Engineering and Technology*, **1**, 407-413. https://doi.org/10.31098/cset.v1i1.412

[48] Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimarães, G.A.S., Cruz, G.O.R., *et al.* (2023) Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, **7**, Article 15. https://doi.org/10.3390/bdcc7010015

[49] Lalor, J.P., Abbasi, A., Oketch, K., Yang, Y. and Forsgren, N. (2024) Should Fairness Be a Metric or a Model? A Model-Based Framework for Assessing Bias in Machine Learning Pipelines. *ACM Transactions on Information Systems*, **42**, 1-41. https://doi.org/10.1145/3641276

[50] Haghish, E.F. and Czajkowski, N. (2023) Reconsidering False Positives in Machine Learning Binary Classification Models of Suicidal Behavior. *Current Psychology*, **43**, 10117-10121. https://doi.org/10.1007/s12144-023-05174-z

[51] Kopparaju, S.T., Chavarriaga, C., Galarreta, E. and Bhatia, S. (2024) Natural Language Processing-Enhanced Machine Learning Framework for Comprehensive Phishing Email Identification. 2024 15*th International Conference on Computing Communication and Networking Technologies* (*ICCCNT*), Kamand, 24-28 June 2024, 1-6. https://doi.org/10.1109/icccnt61001.2024.10723950

[52] Riyanto, S., Sitanggang, I.S., Djatna, T. and Atikah, T.D. (2023) Comparative Analysis

Using Various Performance Metrics in Imbalanced Data for Multi-Class Text Classification. *International Journal of Advanced Computer Science and Applications*, **14**, 1082-1090. https://doi.org/10.14569/ijacsa.2023.01406116

[53] Jin, J., Shen, Y., Fu, Z. and Yang, J. (2024) Few-Shot Open-Set Recognition via Pairwise Discriminant Aggregation. *Neurocomputing*, **602**, Article ID: 128214. https://doi.org/10.1016/j.neucom.2024.128214

[54] Li, J. (2023) An Exploration of Relationships between Prevalence, TPR, TNR and Model Performance Metrics. *SSRN Electronic Journal*, **152**, 1549-1556. https://doi.org/10.2139/ssrn.4530905

[55] Hossain, M.R. and Timmer, D. (2021) Machine Learning Model Optimization with Hyper Parameter Tuning Approach. *Global Journal of Computer Science & Technology*, **21**, 31. https://gjcst.com/index.php/gjcst/article/view/2059

[56] Yu, T. and Zhu, H. (2020) Hyper-Parameter Optimization: A Review of Algorithms and Applications. arXiv: 2003.05689.

[57] Ding, X., Liu, J., Yang, F. and Cao, J. (2021) Random Radial Basis Function Kernel-Based Support Vector Machine. *Journal of the Franklin Institute*, **358**, 10121-10140. https://doi.org/10.1016/j.jfranklin.2021.10.005