Journal of Science and Technology 22 (December 2024): 122-129
ISSN 1994-0386 (Print), 2959- 3239 (Online)

PHISHING WEBSITE IDENTIFICATION: UNLEASHING THE POTENTIAL OF MACHINE LEARNING AND STACKING ENSEMBLES TECHNIQUES

Md. Ziaul Hassan*, Md. Humayun Kabir and Md. Amir Hamja
Department of Statistics, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh
*Corresponding author: Email: nirstathstu@gmail.com, Cell Phone: +88-01768932296
DOI: https://www.doi.org/10.59125/JST.22112

ABSTRACT
As the internet continues to play an integral role in our daily lives, the proliferation of phishing websites poses a significant threat to online security which needs to be addressed. This paper predicts phishing websites using Machine Learning (ML), and ensemble ML algorithms by utilizing an open dataset from Kaggle Competition. The study utilizes a varied array of ML algorithms, encompassing Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) to classify features indicative of phishing activities. Additionally, we propose two Stacking ensemble learning by combining ML models. The first one namely ESM1 which combines DT, KNN, SVM as base learners and LR as meta learner and second one namely ESM2 uses GBM, KNN, SVM as base learners and LR as meta learner. The performance of the used models are compared based on several matrices including accuracy, precision, recall, F1 score and AUC values. Among the trained models the ESM2 model outperformed others with a value of 0.98 for all performance measures and closely followed by individual RF and GBM with a value of 0.98 for all metrices. This research contributes to the ongoing efforts to enhance cybersecurity measures by leveraging the power of machine learning and stacking ensemble learning for the proactive identification and mitigation of online phishing threats.

Keywords: Decision tree, gradient boosting, online security, phishing websites, stacking ensemble

INTRODUCTION
In today's interconnected world, the rise of cybercrime, particularly phishing websites, presents a significant challenge to online security. As cybercriminals adopt more sophisticated tactics, identifying and defending against these threats has become increasingly important. The financial impact of cyberattacks is severe, with losses reaching $1.4 billion in 2019, highlighting the need for strong cybersecurity measures. Phishing attacks, which exploit email and social media to steal sensitive information, are especially widespread and damaging, underscoring the critical need for effective protection strategies. Researchers such as Zabihimayvan and Doran (2019), Nakamura and Dobashi (2019), and Paliath *et al.* (2020) have recognized phishing as the primary technique

employed by cybercriminals to gain unauthorized access to personal details, including usernames, passwords, and credit card information. ML-based solutions have proven effective, especially in identifying zero-hour phishing attacks. Faris and Yazid (2021) emphasize the power of these systems in detecting and preventing phishing attempts before they are officially recognized in the cybersecurity landscape. Phishing detection through ML has gained attention for its adaptability to emerging threats, with algorithms analyzing features like URL structure and domain age (Sindhu *et al.* 2020). These systems are effective in real-time detection, making them ideal for combating sophisticated phishing attacks that bypass traditional methods. Abedin *et al.* (2020) employed a limited number of algorithms, including RF, KNN, and LR, to analyze features extracted from legitimate and phishing URLs. RF emerged as the best performer in this context, achieving a recall rate of 99.0%, an accuracy of 97.0%, and an F1 score of 97.0%. Saha *et al.* (2020) explored phishing detection using ML, comparing two algorithms and applying PCA for feature selection. While they briefly examined CNN, RF proved most effective with 97.0% accuracy. Recent studies emphasize the potential of ensemble models to improve detection accuracy by combining multiple algorithms. Qutub *et al.* (2021) highlight stacking models as particularly effective. For example, Chung *et al.* (2023) developed a stacking model combining LR, XGB, RF, ANN, and SVM, achieving 97.6% accuracy. Similarly, Eroglu *et al.* (2022) used a stacking model with SMOreg, DT, LR, and M5P, attaining 86.07% accuracy in predicting product return rates. Building on this body of research, the present study aims to contribute to the ongoing efforts in phishing websites detection by creating stacking ensemble model.

The rest of the paper organized as in Section 2 the methods and materials, the findings are in Section 3 and finally in Section 4 the conclusion and future research directions.

MATERIALS AND METHODS

Data Sources and Description

The dataset used in this study, "Phishing_Legitimate_full," was sourced from Kaggle and created by IBM data scientists. It contains 10,000 entries and 51 variables, including a 'phishing website' dependent variable and 50 related factors, grouped into categories like URL structure, protocol information, and real-time features. Using this data, we trained machine learning models in Python 3.7 to detect phishing websites, with the objective of developing an effective detection system.

Predictive Model for Detecting Phishing Website

In this research, we present a framework for building and evaluating a phishing website detection model which is in Figure 1. Recognizing the importance of data quality, we prioritize the acquisition and pre-processing of high-quality data to optimize model performance. Feature engineering is key to selecting and creating impactful features that improve prediction accuracy. The dataset is split into training, validation, and testing sets to ensure robust model training and mitigate overfitting. Five individual machine learning models and two stacking model combining them are compared based on performance metrics like precision, recall, F1-score, accuracy, and AUC to select the best-performing model.

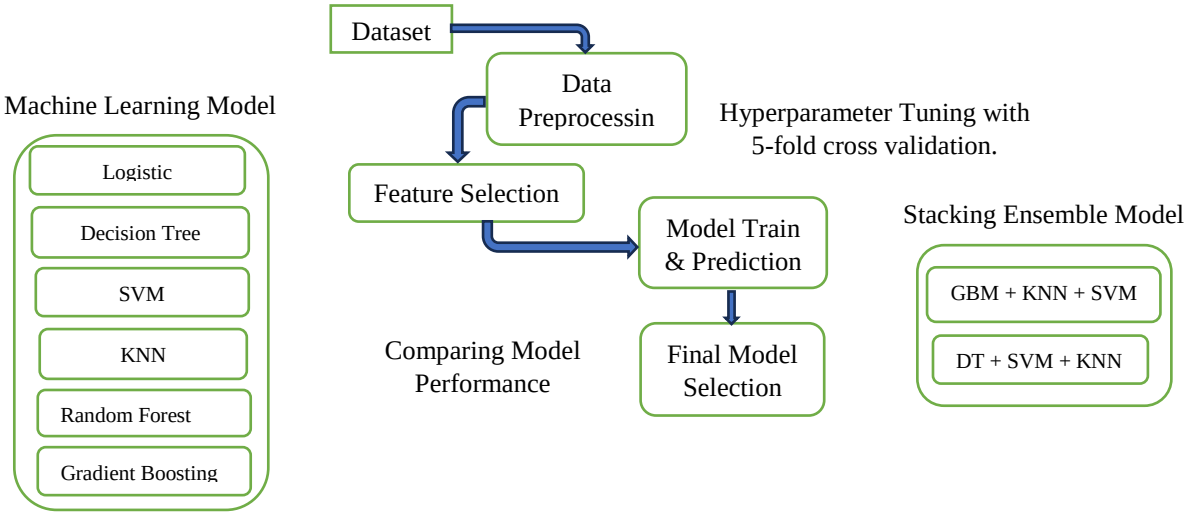

Figure 1. Overview of the proposed phishing websites detection system.

Feature Selection

Feature selection stands as a important step in ML, encompassing the identification and selection of a subgroup of pertinent factors from the main dataset for bolstering the model performance. The goal is to enhance efficiency and effectiveness of the models by retaining the most informative and discriminative features while discarding those that are irrelevant or redundant. In this particular instance, the chi-square test is employed to identify and retain the significant features.

Hyperparameter Tuning

Hyperparameter tuning, commonly referred to as hyper tuning, stands as a most important part in optimizing the performance of almost all ML models. It involves systematically exploring the hyperparameter space to find the most effective combination that maximizes a chosen performance metric. Hyperparameters, which are external configuration settings not learned from the data, significantly impact a model's behavior, generalization, and predictive accuracy.

Cross Validation

Cross-validation is a widely adopted technique to evaluate a model's performance and generalization capability. Firstly, this process includes dividing the dataset into several subsets namely folds, then training model on one-fold, and assessing performance based remaining records. This cycle is repeated multiple times, with each fold acting as both a training and testing set. Following cross-validation, various machine learning algorithms are employed for further analysis.

Logistic Regression (LR)

LR is a foundational machine learning algorithm used for classification tasks. It predicts the probability of an instance belonging to a class using the sigmoid function, transforming input features and weights into a value between 0 and 1. If this value exceeds 0.5, the model predicts the event's occurrence. LR is trained by minimizing cross-entropy loss through optimization techniques like gradient descent, as shown by Jin and Lee *et al.* (2017).

Decision Tree (DT)

DTs are hierarchical models used for both classification and regression tasks. Each node represents a condition based on independent variables, and decisions are made by evaluating conditions at each node until a leaf node is reached to predict the output. For classification, the Gini index is often used to measure the quality of splits, guiding the tree's construction.

Support Vector Machine (SVM)

SVM is a binary linear model widely used for classification and regression tasks. It creates a decision boundary to classify data by maximizing the margin between classes, as noted by Vijayarani *et al.* (2015). SVM is effective for both linear and nonlinear data, excelling in maximizing the margin to reduce classification errors. It is adaptable for various tasks, including prediction of continuous variables, thanks to modifications introduced by Vijayarani and Dhayanand (2015).

K-Nearest Neighbor (KNN)

The KNN algorithm is used for both classification and regression tasks, applying the local approximation principle. It is a simple, lazy learner model that doesn't require training or a loss function. Instead, it predicts by analyzing the nearest neighbors. In classification, it selects the neighbor with the shortest distance, as highlighted by Singh and Bhatia (2018).

Random Forest (RF)

RF is an ensemble learning algorithm that builds multiple decision trees during training and derives predictions by averaging (for regression) or taking the majority vote (for classification) from these trees. Its key strength is reducing overfitting and improving accuracy by combining predictions from diverse, uncorrelated trees. Each tree is trained on a random subset of data and features, introducing variability and enhancing generalization. Additionally, they provide feature importance rankings, making them valuable for many machine learning applications.

Gradient Boosting Machine (GBM)

GBM is a powerful ensemble method similar to RF, combining weak learners to create a more robust predictive model, as noted by Natekine and Knoll (2013). Unlike RF, which builds trees independently, GBM sequentially adds predictors until no further improvement is possible, as described by Chomas *et al.* (2001). It supports various loss functions and is known for its speed

and efficient memory usage. However, its complexity makes visualization and interpretation challenging compared to simpler models like LR or DTs.

Ensemble Method

The ensemble method utilizes various algorithms to enhance predictive performance and reduce errors, similar to consulting multiple experts rather than relying on a single opinion, as noted by Ahmad *et al.* (2020). A popular ensemble model is bagging-based RF, which trains multiple decision trees independently and combines their outputs. Boosting techniques, like AdaBoost and Gradient Boosting, sequentially build weak learners, with each correcting the errors of its predecessor.

In this research, we implement two stacking method for predicting phishing websites, involving three steps: fitting individual algorithms to the training dataset, using their outputs to train a meta-learner, and allowing the meta-learner to make predictions based on these values. Two combinations were created using KNN, DT, GBM, and SVM as base algorithms, with LR serving as the meta-learner.

RESULTS AND DISCUSSION

Data Pre-processing & Feature Selection

In this research, we utilized ML techniques to improve phishing website detection. The first phase involves meticulous data preprocessing to refine the dataset for analysis. We applied a filter method, specifically the chi-square test, to select 21 out of 50 initial features based on a threshold of 500 chi-square values. This selection ensures that only the most significant features are retained, forming the foundation for a robust model aimed at enhancing cybersecurity measures against phishing attacks. The hyperparameters and cross-validation details for the methods utilized in this investigation are presented in Table 1.

Table 1. Hyper Parameter tuning and Cross validation.

Abbreviation	Model Parameter	Hyper Parameter	Cross Validation
LR	C= (0.001, 0.01, 0.1, 1, 10, 100)	C=10	5
DT	C= (3, 5, 7, 9)	C=9	5
SVM	C= (0.001, 0.01, 0.1, 1, 10, 100)	C=10	5
KNN	C= (3, 5, 7, 9)	C=7	5

Model Building

In this study, the prediction of phishing websites involved the use of DT, LR, RF, SVM, KNN, GBM, and two types of ensemble models. ESM1 incorporated DT, KNN, SVM as primary learners, with LR serving the role of meta model. On the other hand, ESM2 utilized GB, KNN, SVM as base models, with LR acting as the meta model.

Table 1. Performance measure of the trained models for phishing website detection

Model Name	Accuracy	F1 Score	Precision	Recall	ROC AUC
LR	0.93	0.93	0.93	0.94	0.93
DT	0.97	0.97	0.96	0.97	0.97
SVM	0.96	0.96	0.96	0.96	0.96
KNN	0.95	0.95	0.95	0.96	0.95
RF	0.98	0.97	0.97	0.96	0.98
GBM	0.98	0.96	0.97	0.97	0.98
ESM1	0.95	0.94	0.95	0.94	0.93
ESM2	0.98	0.98	0.98	0.98	0.98

The best-performing model in the evaluation is ESM2, which achieved an accuracy of 0.98, along with an F1 Score, Precision, and Recall of 0.98, demonstrating exceptional balance in classification performance. Close behind is the RF model, also with an accuracy of 0.98, but slightly lower F1 Score of 0.97 and Recall of 0.96. The ROC AUC score for RF matches that of ESM2 at 0.98, indicating strong discrimination ability. In contrast, Ensemble Model 1 (ESM1) ranked as the least effective, with an accuracy of 0.95, an F1 Score of 0.94, and a ROC AUC score of 0.93, reflecting challenges in precision and recall balance.

Table 3. Performance of Existing Model based on Phishing Websites Detection

Reference	Best Model	Performance	Type
Abedin *et al*. (2020)	Random Forest	Accuracy 0.97	Bagging Model
Abdul Razaque *et al.* (2020)	Logistic Regression	Accuracy 0.95	Single Model
Deshpande *et al.* (2021)	Random Forest	Accuracy 0.97	Bagging Model
Saha *et al*. (2020)	Random Forest	Accuracy 0.97	Bagging Model
This Study	Ensemble Model	Accuracy 0.98	Ensemble Model (stacking)

Several studies, including those by Abedin *et al.* (2020), Deshpande *et al.* (2021), and Saha *et al.* (2020), identified Random Forest as the best-performing model with an accuracy of 0.97 using bagging techniques. Abdul Razaque *et al.* (2020) found Logistic Regression with 0.95 accuracy to be the best single model. In contrast, this study's stacking-based ensemble model surpassed these with an accuracy of 0.98.

CONCLUSION

This study underscores the critical role of ML and ensemble methods in enhancing the detection of phishing websites, which remain a pervasive threat to online security. By employing a comprehensive analysis of various ML algorithms—including DT, KNN, LR, SVM, RF, and GBM—we demonstrated the effectiveness of these techniques in classifying features associated with phishing activities. The introduction of two Stacking ensemble models, ESM1 and ESM2, further illustrates the potential for combining multiple algorithms to improve performance. Notably, the ESM2 model achieved an impressive accuracy of 0.98 across key performance

metrics, surpassing other models, including RF and GBM, which also showed high efficacy. These findings highlight the importance of adopting advanced machine learning strategies to proactively address online phishing threats and contribute to the broader efforts in fortifying cybersecurity measures. Future research could focus on several key areas: First, exploring the integration of deep learning techniques, such as neural networks, to enhance detection accuracy further. Second, investigating the effectiveness of transfer learning in adapting models to new phishing tactics and trends could provide insights into real-time threat adaptation. Third, the development of more sophisticated ensemble methods, including stacking with additional algorithms or hybrid approaches combining traditional and modern techniques, may yield even better predictive performance.

REFERENCES

Abedin NF, Bawm R, Sarwar T, Saifuddin M, Rahman MA and Hossain S. 2020. Phishing Attack Detection using Machine Learning Classification Techniques. 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India. pp. 1125-1130.

Ahmad I, Yousaf M, Yousaf S and Ahmad MO. 2020. Fake News Detection Using Machine Learning Ensemble Methods. Complexity. pp. 1–11.

Chomas JE, Dayton P, May D and Ferrara K. 2001. Threshold of fragmentation for ultrasonic contrast agents. Journal of biomedical optics. 6(2):141-150.

Chung D, Yun J, Lee J and Jeon Y. 2023. Predictive model of employee attrition based on stacking ensemble learning. Expert Systems with Applications. p. 215.

Deshpande A, Pedamkar O, Chaudhary N and Borde S. 2021. Detection of phishing websites using Machine Learning. International Journal of Engineering Research & Technology (IJERT). 10(1):430-434

Faris H and Yazid S. 2021. Phishing Web Page Detection Methods: URL and HTML Features Detection. 2020 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), BALI, Indonesia. pp. 167-171.

Jin SB and Lee JW. 2017. Study on Accident Prediction Models in Urban Railway Casualty Accidents Using Logistic Regression Analysis Model. Journal of the Korean Society for Railway. 20(4):482–490.

Nakamura A and Dobashit F. 2019. Proactive Phishing Sites Detection. 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Thessaloniki, Greece. pp. 443-448.

Natekin A and Knoll A. 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics. 7:1-21.

Paliath S, Qbeitah MA and Aldwairi M. 2020. PhishOut: Effective Phishing Detection Using Selected Features. 2020 27th International Conference on Telecommunications (ICT), Bali, Indonesia. pp. 1-5.

Qutub A, Al-Mehmadi A, Al-Hassan M, Aljohani R and Alghamdi HS. 2021. Prediction of Employee Attrition Using Machine Learning and Ensemble Methods. International Journal of Machine Learning and Computing. 11(2):110–114.

Razaque A, Frej MBH, Sabyrov D, Shaikhyn A, Amsaad F and Oun A. 2020. Detection of Phishing Websites using Machine Learning. 2020 IEEE Cloud Summit, Harrisburg, PA, USA. pp. 103-107.

Saha I, Sarma D, Chakma RJ, Alam MN, Sultana A and Hossain S. 2020. Phishing Attacks Detection using Machine Learning Approach. Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT. pp. 1180–1185.

Sindhu S, Patil SP, Sreevalsan A and Rahman F. 2020. Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation. 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India. pp. 391-394.

Singh R and Bhatia A. 2018. Sentiment analysis using Machine Learning technique to predict outbreaks and epidemics. International Journal of Advanced Science and Research. 3(2):19-24.

Tüylü ANA and Eroglu E. 2022. The prediction of product return rates with ensemble machine learning algorithms. Journal of Engineering Research. pp. 1-14.

Vijayarani S and Dhayanand S. 2015. Liver disease prediction using SVM and Naïve Bayes algorithms. International Journal of Science, Engineering and Technology Research (IJSETR). 4(4):816–820.

Vijayarani S, Dhayanand S and Phil M. 2015. Kidney disease prediction using SVM and ANN algorithms. International Journal of Computing and Business Research (IJCBR). 6(2): 1–12.

Zabihimayvan M and Doran D. 2019. Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection. 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA. pp. 1-6.