Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

"Hybrid Machine Learning Model for Phishing Detection"

Perceval Maturure 1st
School of Computing
University of Portsmouth
United Kingdom
up2136698@myport.ac.uk

Asim Ali 2nd
School of Computing
University of Portsmouth
United Kingdom
asim.ali@port.ac.uk

Alexander Gegov 2nd
School of Computing
University of Portsmouth
United Kingdom
alexander.gegov@port.ac.uk

***Abstract*— Phishing threats have remained a long-standing information security issue for many years, causing billions of pounds in losses both in the United Kingdom and worldwide** [1]**. The aim of the study was to develop and evaluate the performance of the Machine Learning models that would detect and monitor phishing attacks more accurately. A single dataset with 42 features, a total of 247950 phishing and non-phishing emails was used to develop eight supervised machine learning models. The metrics used in evaluating the models show that the enhanced hybrid algorithm developed from combining two models (decision trees and the random forests) from the trained models generated the best classifier with an accuracy of 96%, precision 98%, f-measure 96%, sensitivity 94%, MCC 92% and ROC 96%. The enhanced hybrid voting model developed was integrated with a Django web application using 13 important features to build an accurate phishing detection and monitoring application. The model is proposed as a novel hybrid model because it demonstrated higher classification capabilities due to its inherent design to deal with complex patterns, overfitting issue and the presence of many features when compared to the other single analysis models in the experiments.**

Keywords-; phishing detection, supervised machine learning, classification algorithms

I. INTRODUCTION

The information security domain has become a very important element in meeting organizational business objectives focusing on achieving confidentiality, integrity and availability. Phishing is a fraudulent process where government and financial institution websites are replicated by hackers with the aim of obtaining sensitive information such as identities, usernames and passwords [2]. Criminals often exploit the weaknesses found in system processes caused by system users [3]. Other phishing attacks often involve heterogeneous communication platforms such as email, quick response codes (QR), social media and text messages engineered to persuade the victim to perform tasks that benefit the attacker. Results of a quantitative study by the UK government show that the number of phishing threats and impersonation, has risen sharply from 2017 to 2023 compared to that of viruses and other malware threats [4]. The major aims of phishing attacks have been identified as for financial gain, identity hiding, fame and peer recognition. The impacts of these attacks of leave digital trust issues, reputational damage and cost implications from regulators.

Humans in organizations are the weakest link when it comes to information security [4]. With or without a training background in phishing attacks human beings can make a mistake of clicking on a malicious link resulting in malware delivery. In [3], [5], the major challenge of phishing is that attackers take advantage of human ignorance and naivety when interacting with electronic devices. Criminals target the victim who uses technology vs technology. As an example, users may disclose their passwords if an attacker asked them to update their user accounts via a supplied Hypertext Transfer Protocol (HTTP) link, which ultimately threatens the overall security of the system. The danger of technical vulnerabilities such as domain name system poisoning are that they can be used by attackers to engineer more persuading attack messages.

The technology employed in phishing attacks is complex, very little effort is required to deploy phishing attacks and at the same time it is difficult for users to distinguish between malicious websites and genuine ones [6]. There seems to be no single solution to address the problem of phishing threats because attackers keep generating creative ways of launching phishing attacks[7]. Each time when researchers come up with solutions to combat phishing attacks, attackers often change their strategy by exploiting any new possible vulnerabilities. Building a cybersecurity resilient workforce through continuous training is quite challenging for each phishing attack that is launched [7].

Artificial intelligence (AI) offers a game changing solution to problems associated with phishing threats however, the reference model needs to be updated to detect these future attacks more accurately [8]. In [9], machine learning (ML) is a branch in artificial intelligence which combines statistics, advanced mathematics, and computer science to extract knowledge from data. The algorithms behind the AI and ML technologies can analyze vast amounts of data vs investment in training the human workforce on recently detected phishing strategies. Supervised machine learning algorithms often identify questionable patterns that show signs of phishing attempts and often this data is used to provide recommendations to security operations center teams [10]. In [11], phishing detection using machine learning techniques have limitations because the tedious work involved in picking up the hand-crafted features, knowledge of the statistics domain is a requirement and the complexities encountered by models to pick certain patterns of phishing in URLs and html content from suspected phishing domains. While ML serves as a game changing solution to the problem of phishing detection, in [12], existing machine learning techniques often fail at the feature selection stage resulting in potential inaccurate predictions. Potential misclassification means that phishing emails will find their way into the organization. In [13], adversarial machine learning techniques by attackers can also take advantage of the

979-8-3503-5098-2/24/$31.00 ©2024 IEEE

2024 IEEE 12th International Conference on Intelligent Systems (IS) | 979-8-3503-5098-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IS61756.2024.10705257

reference model flaws leading to the misclassification of phishing domains.

The rest of the paper is organized as follows. Section 2 provides a brief summary of important related works, section 3 described the methodology, section 4 presents the results and finally conclusion form the section 5 of the study.

II. RELATED WORKS

In this section, the supervised machine learning classification algorithms, are discussed from other studies with a focus on concerns associated with classification algorithms, their flaws and reasons behind the success of each algorithm is discussed. In [14], majority of phishing detection systems generally run on single-analysis models which reduces their limitation to detect phishing attacks. While the hybrid models outperform the individual models from studies conducted, we cannot determine whether hybrid models have faster phishing detection speeds or not. Scalability, adaptability, and flexibility of hybrid models are all research areas that are yet to be explored.

A. Logistic regression

Logistic regression finds its application in regression, multi-classification, and binary classification which is useful in predicting whether there is a phishing detection or not in large datasets [15]. The regression model is known to have a stronger classification power when a few features are used during training. Given the ever-changing phishing strategies and phishing techniques this could be a problem when important features that could help in picking phishing attack features are left out. In such a scenario reducing the dataset features would have a negative effect on the model's predictive power.

B. K-nearest neighbour (KNN)

The KNN is a classification algorithm which depends on the k value where the data points to be classified are sorted closest to the k points and assumes that the dataset has been assigned some classes [16]. KNN model accuracy can be compromised by noisy data. There are other methods which have been developed to resolve the problem of noisy and useless [17]. Exclusion of noisy data in order improve the classification power can result in attack strategies being missed in detecting phishing attacks. Other supervised machine learning algorithms are not affected by noisy data but rather leverage to become better classifiers.

C. Support Vector Machines (SVM)

The SVM model has been widely used in solving pattern recognition and facial recognition problems [18]. One of the biggest drawbacks of the SVM is that it requires more time to train compared to other supervised machine learning problems and finding the perfect K value [19], [20]. While the SVM is excellent with classification, it requires more time to train which can be a problem when extremely large datasets are involved therefore limiting the dataset size is not a guarantee for a model with a high predictive classification capability.

D. Naïve Bayes

The bayes algorithm would have a better predictive capability when the training dataset is large [21]. One of the issues with the bayes algorithm is conditional dependence, ways of satisfying conditional dependence would mean adjusting the dataset features by selecting those with more weight than the others [22]. The challenge with this approach is computational overhead however, the model can generate faster results [23]. Higher model accuracy would require computational storage, and this can be challenge in resource limited settings.

E. Decision trees

In [24], describes the decision tree as a type of supervised machine learning approach where internal nodes are like a tree with testing nodes and the leaf represent the result of the decision. The decision tree approach is flexible because it can model non-linear relations however, its drawback is that it overuses data, and it is difficult to update decision tree samples [8]. While the decision tree algorithm is more accurate compared to the Naïve bayes its big disadvantage is that it is subject to overfitting. The decision tree is not affected by the volume of data, it can deal with noisy data while the Naïve bayes is less accurate on a small dataset. For the purposes of training the decision tree model, large datasets are preferred for easier detection of all phishing tactics embedded into a URL link however, the drawback with this is computational memory requirements.

F. Artificial Neural Networks (ANN)

The artificial Neural network is a set of interconnected nodes similar to the nerve system in a organism with a weight assignment to a node [17]. The neural networks algorithms are often restricted by aggressive assumptions of normality, feature dependence, linearity and are sensitive to the size of hidden layers [25] . The presence of irrelevant features makes the model training complicated, generally slower to train, this is contrary to the other supervised machine models like the random forest where the total presence of features with low relevance plays a significant role in the total classification capability. While time is a factor to consider in training an ANN model for use in a production environment, it is more important to have a model with the highest classification power than focusing on the time required for training. It is more advantageous to have a model which can handle lots of features and noisy data irrespective of the training period to deliver a highly accurate model.

G. Random Forests (RF)

The random forests are one of the most successful sets of combined supervised machine learning algorithms that are built on many decision trees, that are used in classification and regression. The accuracy of detecting phishing attacks is improved in the RF because it is not affected by having large datasets, noisy data, a common problem that compromises model performance as exhibited by classification by similar supervised machine learning classification algorithm such as the KNN and SVM. A limited dataset size would also

correspond to a limited phishing detection capability and so this is not the case with the RF algorithm.

H. Hybrid Solutions

Single analysis models are generally weak learners [26]. Each machine learning model either suffers a specific problem of overfitting, slow training, require high performance computers to quicken the training process, or some do not do well with large datasets. Some other algorithms have many benefits which would need to perfect other models. The ensemble models have different types of principles that they operate one to produce a more effective model. In this study the ensemble methods used a voting principle to provide a new vote to generate a more accurate prediction built on a multiple models termed as a hybrid solution.

III. METHODOLOGY AND IMPLEMENTATION

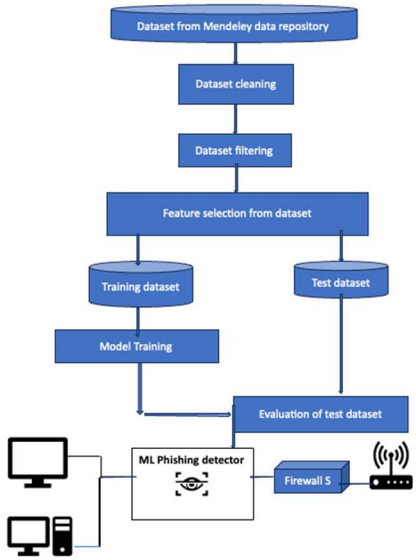

Figure 3.1 the method employed

The above figure 3.1 shows all the steps applied to each of the supervised machine learning algorithms trained. The dataset for the study has 42 features, 247950 records where 48.6% of the data resembled malicious domains and 51.84 genuine domains obtained from an online Mendeley data repository where other researchers submit their data [27]. A standard scaler was used to standardize values in the dataset so that they exist within the values of 0 to 1. The supervised machine learning techniques employed in this project started by primarily developing classification models using a single dataset that share common features as a substitute for exhibits of phishing urls and domains [28]. The supervised machine learning algorithms are then trained to learn the presence of phishing patterns from the dataset [29]. The dataset split ratio of training set to test will be interchanged to compare the performance of the model. In a typical supervised machine learning scenario, a classification algorithm is more appropriate as it generates a 99% accuracy [30], [31]. A series of confusion matrices were used to rate the accuracy, precision, f1-measure, and performance of each of the classification models. The best classifier was integrated with a Django application to develop a phishing detection and monitoring system as the proposed in Figure 3.2. Figure 3.3 is a list of important features as generated by the random forest model.

A. Proposed Enhanced Phishing Dectection and Monitoring System

The phishing detection system in figure 3.2 was built using python, bootstrap 4 front-end framework, Django 4, a web application framework integrated with the highly accurate Voting model developed from combining the random forests and the decision trees [32]. Important fields from figure 3.4 have been used in the development of the system. The model developed was trained on a scaled data using the standard scaling technique therefore input features will be values between 0 and 1.

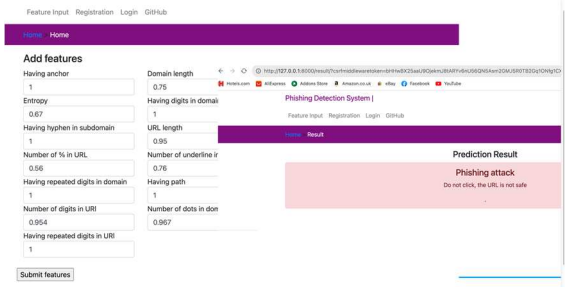

Figure 3.2 Phishing Detection and Monitoring system

B. Dataset

The dataset for this study was obtained from a Mendeley data repository where a study was done on detecting phishing attacks [27]. The dataset published 07 June 2023 consists of 247 950 data points records where 119 409 are from legitimate websites and 128 541 are from phishing website instances. The dataset is one of the most recent datasets consisting of 42 features.

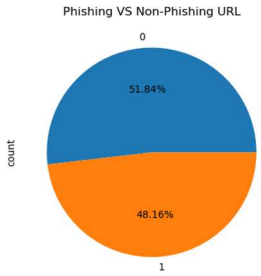

Figure 3.3 Dataset

C. Important features

In Figure 3.4, the important features necessary to detect phishing attacks are displayed with a blue background. The features have been extracted using the important feature method built within the random forest model because it is the only supervised machine learning model with high classification capability other than the hybrid model. The height of the blue

bars represents the extent to which each feature is important. Almost to 60% of the features selected for model training in this dataset are essential in the delivery of a hybrid phishing detection classifier.

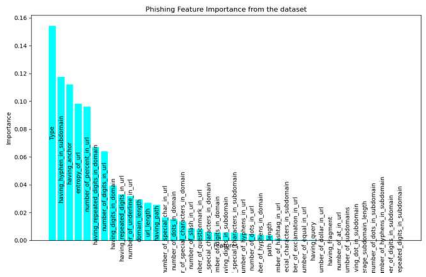

Figure 3.4 Important features

IV. RESULTS AND EVALUATION

A. *Summative matrix of results*

The results of the study are summarized in table 4.1 where all metrics used to measure model performance are tabulated. The ranking of metrics used to evaluate the model's classification capability were not limited to precision. accuracy and sensitivity. A few are mentioned in the bar graph exhibits in figures 4.1, 4.2 and 4.3 for brevity. One of the factors that attributing to the low classification capability of the logistic regression model is the presence of the large number of features selected for training [12]. The Naïve bayes algorithm is the weakest classifier because of reasons associated with conditional dependence which need to be satisfied by adjusting dataset features. In a more general way, the logistic regression, the KNN, SVM, Naïve Bayes and the neural networks ranked lowest in terms of classification capability. The ensemble voting model was developed by combining the decision trees and the random forest which were the highest performing models. The resulting model is an enhanced hybrid solution because it combines two best models from the experiment giving a more reliable solution, should there be deficiencies in the decision tress the random forests would cover up the flaws of the decision trees and vice versa. An example is that of the random forests, the random forests model is not affected by overfitting, but the decision trees is affected to some extent by overfitting therefore that would be the sole reason of having a high classification capability as shown by the hybrid model. The random forests are also not affected by the dataset size, the number of dataset features, noisy data and therefore makes it a perfect fit to combine it with the decision trees. The hybrid voting ensemble model yields a precision score of 98% which is the highest among other 8 models. This is because of the power of the combined classification effect. A few major limitations of this study are that the reference model does not factor in the concept of adversarial machine learning attacks in detecting misclassified phishing domains and the phishing detection system proposed in this study is only limited to 13 important features.

Table 4.1 Model classification capability

Model Name	Accuracy	Precision	Sensitivity	F1-score	ROC Score	MCC
Logistic	0.80	0.84	0.73	0.78	0.80	0.61
KNN	0.91	0.92	0.90	0.90	0.91	0.82
SVN	0.89	0.91	0.84	0.88	0.88	0.77
Naïve Bayes	0.74	0.87	0.54	0.67	0.73	0.56
Decision Trees	0.95	0.95	0.95	0.95	0.95	0.91
Neural Networks	0.90	0.90	0.89	0.89	0.96	0.80
Random forests	0.96	0.97	0.95	0.96	0.96	0.92
Voting (DT + RF)	0.96	0.98	0.94	0.96	0.96	0.92

Table 4.2 has an explicit detail of all the formulae describing how each of the metric used in the evaluation of the models have been computed

Table 4.2 Model evaluation metrics

	Metric	Formula (Values extracted from the confusion matrix)
1	Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
2	Precision	$\frac{TP}{TP+FP}$
3	Recall/Sensitivity	$\frac{TP}{TP+FN}$
4	F1 Score	$2*\frac{Precision*Recall}{Precision+Recall}$
5	MCC	$\frac{TP\ X\ TN-FP\ X\ FN}{\sqrt{((TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))}}$

B. *Model ranking by accuracy*

In figure 4.1, the random forest model and the voting (combination of the decision trees and the random forests) models show the highest accuracy of 96% of all the 8 models in figure 4.3.2. The high accuracy of the voting model attributed by its ability to handle the problem of overfitting coupled with the computation of the best result through voting of the random forests and the decision trees which other models cannot handle and therefore having the ability to classify complex patterns [33]. The accuracy in the logistic regression is low because the high number of features in the dataset. The naïve bayes algorithm shows the lowest accuracy of 74% because of its inability to handle many features.

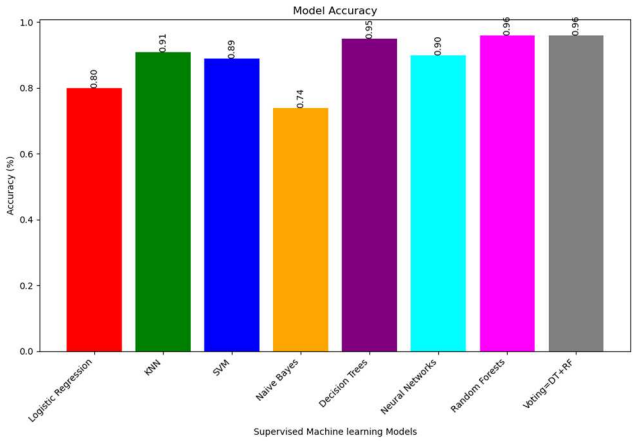

Figure 4.1 Accuracy

C. *Model ranking by Precision*

In figure 4.2, the naïve bayes has the least precision, and the voting model has the highest precision amongst all the models in the experiment. The hybrid model has a higher precision of 98%, this means that the voting model has a high accuracy in

predicting genuine emails (precision) more than the random forests level having 97% attributed but the voting algorithm.

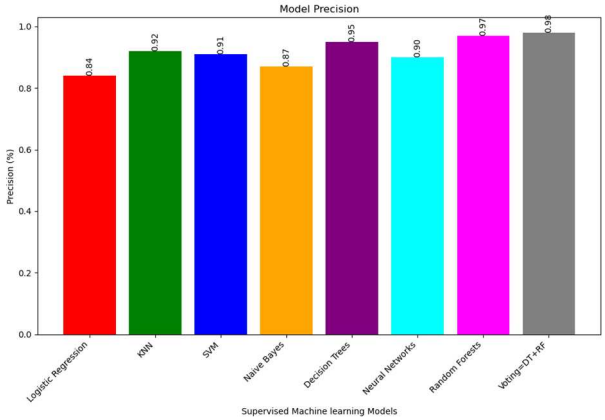

Figure 4.2 Precision

D. Model ranking by Recall/Sensitivity

In figure 4.3, the recall or sensitivity parameter has been described and it measures the fraction of detected phishing attacks. The decision tree and random forest both have the highest sensitivity of 0.95 and the voting algorithm with a value of 0.94. The naïve bayes preforms the lowest at 0.54.

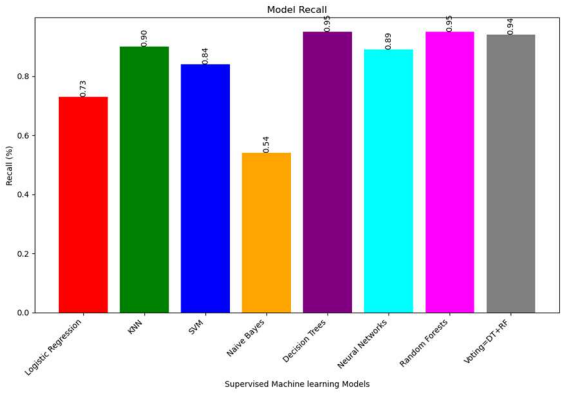

Figure 4.3 Recall

V. CONCLUSION

The best classifier from the study is the hybrid model which has exhibited the highest scores in all the six metrics used in table 4.2 to evaluate the model. The power of the hybrid model is that it is an ensemble of method which combines many decision trees and the random forests by using a voting principle to improve its performance in terms of classification per literature. The study suggests that there is a strong positive relationship between the classification capability of the model and the number of important features. The lesser the features the lesser the model performance. In-order to accurately detect phishing attacks, a more recent dataset with more data and more features is required to be able to accurately detect phishing attacks. The same applies to the system developed to detect phishing attacks, it would require more than thirteen features as input in-order to accurately predict whether a URL is malicious or not. The phishing detection system suggests an innovative approach which enhances the ability to monitor email traffic, detected malicious URL's can be identified for future model training and ultimately improve cybersecurity outcomes. The future work of this study would drive the development of phishing detection systems with advanced features that can respond to phishing attacks, promote phishing threat avoidance, trigger protection of critical assets and internet users.

REFERENCES

[1] D. Hillman, Y. Harel, and E. Toch, "Evaluating organizational phishing awareness training on an enterprise scale," *Comput Secur*, vol. 132, p. 103364, Sep. 2023, doi: 10.1016/J.COSE.2023.103364.

[2] G. Vrbančič, I. Fister, and V. Podgorelec, "Datasets for phishing websites detection," *Data Brief*, vol. 33, p. 106438, Dec. 2020, doi: 10.1016/j.dib.2020.106438.

[3] M. Khonji, Y. Iraqi, S. Member, and A. Jones, "Phishing Detection: A Literature Survey," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, vol. 15, no. 4, 2013, doi: 10.1109/SURV.2013.032213.00009.

[4] UK Government, "Prevalence, Impact, of data breaches and and cyber attacks." Accessed: Jan. 29, 2024. [Online]. Available: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2023/cyber-security-breaches-survey-2023#chapter-4-prevalence-and-impact-of-breaches-or-attacks

[5] A. Safi and S. Singh, "A systematic literature review on phishing website detection techniques," *Journal of King Saud University - Computer and Information Sciences*, 2023, doi: 10.1016/j.jksuci.2023.01.004.

[6] R. Zieni, L. Massari, and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites", doi: 10.1109/ACCESS.2023.3247135.

[7] A. K. Jain and B. B. Gupta, "A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP J Inf Secur*, vol. 2016, no. 1, p. 9, Dec. 2016, doi: 10.1186/s13635-016-0034-3.

[8] V. Shahrivari, M. M. Darabi, and M. Izadi, "Phishing Detection Using Machine Learning Techniques," 2020. Accessed: Feb. 12, 2024. [Online]. Available: https://arxiv.org/abs/2009.11116

[9] J. Cutler and M. Dickenson, "Introduction to Machine Learning with Python," in *Computational Frameworks for Political and Social Research with Python*, Cham: Springer International Publishing, 2020, pp. 129–142. doi: 10.1007/978-3-030-36826-5_10.

[10] K. Bresniker, A. Gavrilovska, J. Holt, D. Milojicic, and T. Tran, "Grand Challenge: Applying Artificial Intelligence and Machine Learning to Cybersecurity," *Computer (Long Beach Calif)*, vol. 52, no. 12, pp. 45–52, 2019, doi: 10.1109/mc.2019.2942584.

[11] C. Opara, Y. Chen, and B. Wei, "Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics," *Expert Syst Appl*, vol. 236, p. 121183, Feb. 2024, doi: 10.1016/J.ESWA.2023.121183.

[12] M. Tools, M. Nanda, and S. Goel, "URL based phishing attack detection using BiLSTM-gated highway attention block convolutional neural network," *Multimed Tools Appl*, 123AD, doi: 10.1007/s11042-023-17993-0.

[13] A. McCarthy, E. Ghadafi, P. Andriotis, and P. Legg, "Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification," *Journal of Information Security and Applications*, vol. 72, p. 103398, Feb. 2023, doi: 10.1016/J.JISA.2022.103398.

[14] R. J. van Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, "The applicability of a hybrid framework for automated phishing detection," *Comput Secur*, vol. 139, p. 103736, Apr. 2024, doi: 10.1016/J.COSE.2024.103736.

[15] R. Pelle, C. Alcântara, and V. P. Moreira, "A Classifier Ensemble for Offensive Text Detection," in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, New York, NY, USA: ACM, Oct. 2018, pp. 237–243. doi: 10.1145/3243082.3243111.

[16] S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, IEEE, Aug. 2010, pp. 91–94. doi: 10.1109/FSKD.2010.5569740.

[17] A. Subasi, E. Molah, F. Almkallawi, and T. J. Chaudhery, *Intelligent Phishing Website Detection using Random Forest Classifier*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites#

[18] S. Saleem, J. Shiney, B. Priestly Shan, and V. Kumar Mishra, "Face recognition using facial features," *Mater Today Proc*, vol. 80, pp. 3857–3862, 2023, doi: 10.1016/j.matpr.2021.07.402.

[19] L. Jiang, Z. Cai, D. Wang, and S. Jiang, "Survey of Improving K-Nearest-Neighbor for Classification."

[20] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer," *Behav Ther*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/J.BETH.2020.05.002.

[21] Y. Xu *et al.*, "A Phishing Website Detection and Recognition Method Based on Naive Bayes," in *2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC)*, IEEE, Mar. 2022, pp. 1557–1562. doi: 10.1109/ITOEC53115.2022.9734474.

[22] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve Bayes algorithm," *Knowl Based Syst*, vol. 192, p. 105361, Mar. 2020, doi: 10.1016/J.KNOSYS.2019.105361.

[23] A. Sharma and A. Suryawanshi, "A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure," *Int J Comput Appl*, vol. 136, no. 6, pp. 28–35, Feb. 2016, doi: 10.5120/ijca2016908471.

[24] S. Patil and U. Kulkarni, "Accuracy Prediction for Distributed Decision Tree using Machine Learning approach," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, IEEE, Apr. 2019, pp. 1365–1371. doi: 10.1109/ICOEI.2019.8862580.

[25] A. Singh, "Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)," 2016.

[26] D. Datta, M. Bhattacharya, S. S. Rajest, T. Shynu, R. Regin, and S. S. Priscila, "Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting," *Int J Bioinform Res Appl*, vol. 19, no. 3, pp. 151–169, 2023, doi: 10.1504/IJBRA.2023.133695.

[27] Maruf Tamal, "Phishing Detection Dataset," *Mendeley Data*, vol. 1, Jun. 2023, Accessed: Feb. 03, 2024. [Online]. Available: https://data.mendeley.com/datasets/6tm2d6sz7p/1

[28] S. R. Janani, R. Ashwin, S. Kumar, S. Dinesh, Siddharth, and Yashwanth, "Detection of Phishing Page Using Machine Learning and Response HTML," 2024, pp. 499–508. doi: 10.1007/978-981-99-7137-4_49.

[29] M. AlShaikh, W. Alsemaih, S. Alamri, and Q. Ramadan, "Using Supervised Learning to Detect Command and Control Attacks in IoT," *International Journal of Cloud Applications and Computing*, vol. 14, no. 1, pp. 1–19, Nov. 2023, doi: 10.4018/IJCAC.334214.

[30] M. H. Alkawaz, S. J. Steven, A. I. Hajamydeen, and R. Ramli, "A Comprehensive Survey on Identification and Analysis of Phishing Website based on Machine Learning Methods," in *2021 IEEE 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, Apr. 2021, pp. 82–87. doi: 10.1109/ISCAIE51753.2021.9431794.

[31] Simon Andrews and Laura Biggins, "Introduction to Machine Learning." Accessed: Apr. 03, 2024. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/training/MachineLearning/Introduction_to_Machine_Learning_slides.pdf

[32] Perceval Maturure, "Phishing detection code base repository." Accessed: Mar. 06, 2024. [Online]. Available: https://github.com/pmaturure3/msc_project

[33] M. A. Salam, A. Taher, M. Samy, and K. Mohamed, "The Effect of Different Dimensionality Reduction Techniques on Machine Learning Overfitting Problem," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, 2021, doi: 10.14569/IJACSA.2021.0120480.