International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

Develop a Hybrid Classification using an Ensemble Model for Phishing Website Detection

Subashini K[1], Narmatha V[2]
[1]Research Scholar, Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India
[2]Assistant Professor ,Department of Computer and Information Science, Annamalai University, Chidambaram, Tamil Nadu, India
subaphdscholar@gmail.com

Abstract - Solutions to threats posed by technical and social vulnerabilities must be found to secure the web interface. Social engineering attacks frequently use phishing as one of their vectors. The importance is promptly detecting phishing attacks has increased. The classifier model was constructed using publicly accessible data from trustworthy and phishing websites. A variety of methods were used to extract relevant features to build the model. Before a user experiences any harm, Machine Learning algorithms can reliably identify phishing attacks. To identify phishing attacks on the website, this study presents a novel ensemble model. In this paper, the Artificial Neural Network (ANN) and the Random Forest Classifier (RFC) are used in an ensemble method along with the Support Vector Machine (SVM). Compared to previous studies, this ensemble method more accurately and efficiently detects website phishing attacks. According to experimental findings, the proposed system detects phishing attacks 97.3% of the time.

Keywords - Phishing attacks; Feature extraction; machine learning; Ensemble methods; Accuracy.

I. INTRODUCTION

Phishing website fraud is a more recent form of online crime when compared to malware attacks. To attract users, phishers use social engineering approaches to send emails, online advertisements, or instant messages to trick individuals into disclosing sensitive information to phishing websites that impersonate trustworthy websites [1]. Blacklisting is a common filtering method used by security software products to block known websites to protect users from phishing websites. Website reporting and Blacklist updating always happen after a delay. Since phishing websites now only exist for a few hours instead of days, this approach may not work.

Much study has gone into creating clever methods for phishing website prevention and detection over the past few years. However, some issues persist. For example, studies that make use of domain name information, URL addresses, website rankings, etc.., because the website's characteristics always result in lower recognition rates; To detect phishing, ML and heuristics techniques, which utilize wordings and pictures content from webpages, have been developed, the majority of them, however, exhibit high levels of rates of false positives and complexity; A small experimental data set was used for the majority of the current studies, There is no guarantee that these algorithms will perform robustly and effectively on actual large-scale data sets; Additionally, since there are an increasing number of phishing websites, it is important to find a real-time way to distinguish them from legitimate websites[2].

Information security protocols must distinguish between phishing and legitimate web content, but there is still room for improvement in this area. In this study, we propose an ensemble model for detecting phishing websites using ML [3]. In terms of reliability and effectiveness, compare the performance of the suggested model with cutting-edge methods. Additionally, to significantly raise the ensemble model's rate of successfully detecting phishing attacks.

II. RELATED WORK

In comparison to malware, phishing is a more recent and serious security risk to the Internet because it is a semantic assault focused on the user as opposed to the computer. Heuristic techniques based on DNS and URL characteristics have proposed a replacement for the traditional signature-based detection methods to address this issue [4,5]. Finding the most pertinent characteristics by converting a sizable feature space into a new feature space with the fewest possible features constitutes feature selection and is a crucial step in the development of the model because the data are highly dimensional [6]. The process of developing the models has generally included feature selection. Web browsers may already have anti-phishing tools built in, or they may run separate programs. It has been tested against Firefox and Internet Explorer [7].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

Both non-linear and linear categorization issues can be resolved using the ML algorithm SVM. Vapnik developed the algorithm which is based on the idea of maximizing the separation between class variables and hyperplanes [8]. Support vectors are used to describe the separation's edges, while hyperplanes are used to describe the area in the middle. To maximize the functionality of an algorithm, several kernel functions including Radial Basis Function (RBF) kernels string kernels, and polynomial kernels have been defined [9]. SVM has received positive reviews in numerous studies as a method for anticipating phishing attacks. SVM outperformed other methods in several studies. An algorithm for supervised ML called ANN was developed based on how human brain cells function [10].

To improve the accuracy of individual classifier predictions, by combining the output from various classifiers, ensemble algorithms are created [11]. Less noise and bias are likely to be present in the combined result of the individual classifiers. Consequently, it would become a strong learner capable of handling enormous classification problems [12]. For identifying phishing attacks, there are numerous literature-available ensemble algorithms. Incorporating three different ensemble algorithm variants Bagging, Boosting Learners, and Random Forest a two-step process is used in the current study [13]. These students receive sequential training while absorbing knowledge from the whole data set [14]. This procedure is repeated until it can be said with certainty that strong learners will result from optimized performance [15]. The idea is developed to develop a learner to work well in an ensemble and anticipate the appearance of phishing websites [16]. To determine which ensemble learner model performs the best, the model that was thusly derived is evaluated against the others.

To create twofold models, feed-forward neural networks are thought to be the best option because their combined predictive performance is superior to that of unique learners [17]. Notable considerations include the fact that these networks, which learn from previously generated models, are excellent ensemble learners. The three ensemble learners in the current study each had a feed-forward neural model are created while taking all these factors into account [18]. The terms "Random Forest Neural Network model," "Bagging Neural Network ensemble model," and "Boosting Neural Network ensemble model" refer to the neural network-based ensemble models built for the current study [19]. Neural networks and multiple learners are referred to as "heterogeneous ensemble learners."

To detect phishing attacks, the random-forest method is frequently used. Several researchers used RF, according to studies already conducted, and demonstrated promising accuracy. SVM, K Nearest Neighbor, Random Forest, and C4.5 classifiers were among the six used by the authors [20]. Their approach involved entering URLs, extracting 30 features from those URLs, and then using those features to anticipate phishing attacks. Random forests, decision trees, generalized linear models, gradient boosting machines, and principal component analysis were all employed [21]. The Random Forest algorithm was used as a binary classifier in the study proposed by the authors and 48 features were added later using the relief algorithm's forward selection method, which started with 10 features [22].

III. PROPOSED SYSTEM

A. Phishing data set

A fair set of benchmark data is a requirement for any machine learning-based model, and numerous repository sites offer data sets for various uses. The data set for this study includes 8266 instances of phishing emails from the anti-phishing website, 47 real features, and one class with two distinct class labels: ham and phishing.

B. *Data Preprocessing*

Data processing is required to eliminate redundant and outliers' tuples from the data set. The target variable is seen to not be distributed equally among the negative and positive classes in the data set from the phishing website. When referring to such data sets in classification problems, the term "unbalanced data sets" is frequently used. ML algorithms are not recommended for use on such data sets because they produce inaccurate classifications and skewed results. The literature contains several strategies for handling such unbalanced data sets. Cluster-based oversampling is one such technique. Data sets are clustered independently on negative and positive occurrences of a variable, where this method uses the k-means clustering algorithm. To create data groups with an equal distribution of negative and positive classes, additional oversampling is applied to each cluster that is obtained. As a result of oversampling used in k-means clustering, 988 data instances were identified as inappropriate entities.

C. Feature selection

Variable selection and Data reduction are the important steps that must be taken before creating the predictive model. When using the full set of data attributes to create a classification model, the results could vary. Numerous feature selection techniques exist to help identify significant attributes in this context. These methods fall under the categories of wrapper methods, filter methods, and embedded methods. In this study, relevant features are

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

chosen using both wrapper and techniques. First, correlational analysis is used to remove redundant features from the data set. A threshold value is used to determine whether to remove redundant features.

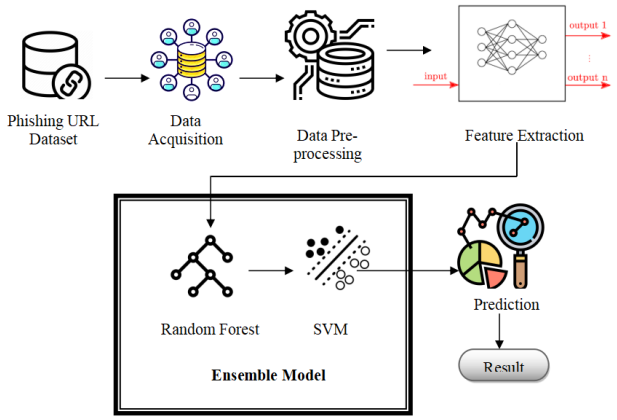

Figure 1. Proposed Architecture

The important measure for each of these features is computed using random forest classifiers that have been trained on the set of features. The more significant the attribute, the variable importance is the greater the higher its value. When the significant features are chosen and the unimportant features in the data set are disregarded, the algorithmic process comes to an end. The best combinations of data features are discovered using the wrapper selection method.

The set of data must be sectioned into various groups after significant features have been chosen to evaluate a classifier's performance over several iterations. The first data partition—which is frequently referred to as "training data"—is used to create the model. The remaining data are referred to as "test data," and they are used to evaluate the model's performance. Before the model development phase, different data partitions for training and testing are taken into account. It is further explained why there are multiple partitions. Any data set with only 1,000 or so instances compared to this data set samples from the phishing data set can be described as "hefty data samples". For sample sizes of this size, the distribution of the data split ratio may not always be equal. To reduce the likelihood of overfitting and bias, it is also advantageous to introduce randomness into the data-splitting process.

D. Anti-Phishing Strategy Model

Data acquisition is the process of the digitizing the data can be displayed, analyzed and stored in the database. Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. Automatic phishing classification has been accomplished in this paper using a hierarchical clustering approach, and to assemble all anticipated outcomes from diverse classifiers, an ensemble classification algorithm is presented. To categorize and identify phishing websites, a framework for an intelligent anti-phishing strategy was presented. The model is shown in Figure 1.

E. Performance Metrics

To assess a classifier's performance, it is crucial to make sure that the prediction of phishing websites is carried out correctly. These standards of evaluation gauge the accuracy with which a predictive model can classify websites according to their phishing risk [23]. There are many metrics from the literature used to estimate how well a classifier predicts outcomes. Precision, Accuracy, F-measure, and Recall are some of the common standards. The percentage of correctly classified data instances is determined by the accuracy metric. Equation 1 defines it as follows:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{(True\ Positive + False\ Positive + True\ Negative + False\ Negative)} \quad (1)$$

The ratio of phishing websites that are phishing websites is known as precision. Equation 2 defines it as follows:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (2)$$

Identifying phishing websites correctly can be measured by the recall. The recall is given by Equation 3.

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (3)$$

The harmonic average of recall and precision is known as the F-measure. Equation (4) provides a definition:

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \quad (4)$$

An analytical method for estimating classification errors or deviations is the mean squared error (MSE). Equation (5) gives the definition:

$$Mean\ Squared\ Error = \frac{1}{n}(Actual\ value - Predicted\ value\ from\ classifier)^2 \quad (5)$$

Consider the two-category classification problem as it pertains to our method here. The dataset's comparison property vectors are used as inputs, and the classifier converts them to binary outputs, 1 or 0. The classifier is trained on one set, and its performance is assessed using the other set, which serves as a testing set. SVM seeks to maximize the separation between the nearest points to determine the best-separating hyperplane between classes. Classifier performance is determined by the small number of support vectors generated following training, which protects

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

against dimension catastrophe and provides solid performance in robustness. An algorithm for supervision ML called ANN was developed based on how human brain cells function. Neurons, the interconnected nodes that make up an ANN model, are interconnected to transmit information across various layers. An ANN model typically has three separate layers: one or more hidden layers, the input layer, and the output layer. Based on a pre-determined threshold value, the learning process progression can be adjusted with greater precision, the weight function is changed among the various layers. Random Forest. Bagging is an algorithm for parallel ensemble learning, as opposed to boosting. It takes various sets from the practice set, and trains the fundamental learners using these various sample sets, the bagging technique is expanded through the creation of numerous decision trees for training. To achieve stronger generalizations, the random attribute is only introduced in certain contexts.

IV. EXPERIMENTAL RESULTS

Comparisons are made between the potency of the forecasting models developed for the three distinct phishing data sets. The cross-validation is carried out by randomly dividing the original data instances into 10 subsamples of equal size. Nine of the ten subsamples—or training data—are included in the analysis, while the remaining subsample is kept for testing. Tenfold cross-validation is used in the experiments for each classifier and a single predictive model as displayed in Table 1.

A. Performance of Ensemble Learners

The effectiveness of the initial group of students' learning was assessed using the data set from a phishing website, including bagging, random forest, and boosting, is assessed. It outperformed other individual variants, the feed-forward neural network is chosen for creating the twofold ensemble learners. Due to FF_ANN's capacity to learn from trained models, this choice is also made. The output of the feedforward neural network makes up the second layer, while the first input layer is made up of predictions made by the ensemble of learners. After several iterations, Additionally, ensemble learners can be divided into two types that must be distinguished, and this is crucial. Preliminary models from the random forest, boosting learners, and bagging, are included within the initial group of ensemble learners. The student t-test is employed in this situation. The t-test is applied across the board, across the entire phishing dataset instance. The following are the procedures for running a t-test on a group of students: Initially, 70% of the set of data is selected to serve as the learning of the designed architecture, and the other 30% is used to test the created predictive models. The test data set is used to compute the F1 measure for each model. Each category of ensemble learners goes through this process once.

- The data has been redistributed once more, with 30% used for testing and 70% for training.
- The training data used in the previous step is used to create the predictive models, and the test data is used to calculate the F1 measure.
- The procedures are carried out 30 times to get 30 F1 values as calculated for each ensemble learner.
- To calculate t-test results using 0.05 as the cutoff value for alpha to assess the statistical difference, the F1 measures between ensemble learners.

As shown in Table 2, there are statistical differences between the outcomes of the two groups' performance, where it is discovered that the p-values are lower than the alpha value

Table 1. The effectiveness of specific classifiers of ML

Algorithm	Variant	Accuracy	Precision	Recall	F-Measure
RFC	-	56.79	69.01	65.93	66.98
SVM	Polynomial (1 to 2)	70.02	75.85	73.38	74.58
ANN	FF (0-1)	72.47	80.15	77.45	78.76

Table 2. F1 measures are compared between classes using the t-test

Class 1	Class 2	Observations count	t-score	P-value
RF	RF_NN	33	22.42	0.03
Bagging	Bagging_NN	33	31.54	0.04
Boosting	Boosting_NN	33	34.92	0.05

B. Evaluation

We used several ML classifiers to test their effectiveness to gauge the efficacy of our approach. The authors evaluate the outcomes of our method using four metrics: precision, accuracy, F1 score, and recall. The percentage of web pages correctly identified as legitimate or phishing pages out of the total sampled web pages is the accuracy. Precision is the ratio of the number of websites correctly identified as phishing sites to the total number of websites detected. The recall is the ratio of the total number of phishing samples to the number of web pages that are correctly identified as phishing pages. Corresponding and Target page pairs, suspiciously similar pages make up positive samples shown in Table 3.

Table 3. Data set

Source	**Phish Tank**	
Dataset	Positive samples	Negative samples
Training set	3722	17928
Testing set	416	1993

C. Classifier effectiveness

Here, first assess the classifiers' performance considering various factors. we used all the 24059 effective

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

samples from the above experiments in this study to assess and disregard the imbalance between the negative and positive samples, which will be examined in the following experiment.SVM.

Four metrics related to the SVM algorithm's parameter gamma are tested, and SVM is used as the classifier in this case. Figure 2 representations of the experiment's findings, precision is around 96%, while the other three metrics are mostly above 80%. While recall decreases slightly and precision increases slightly as gamma rises, accuracy and F1 value almost remain constant. The four metrics perform at or near their peak levels when gamma is around 0.0002.

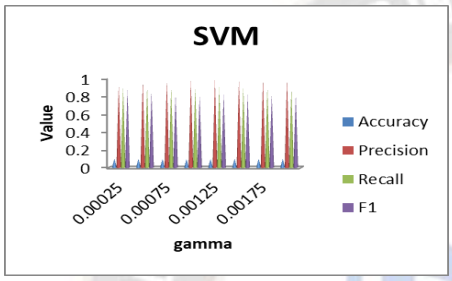

Figure 2. SVM performance measures

The ANN algorithm uses four metrics as a classifier and the n_estimators' test environment is used. Figure 3 displays four metrics that are all greater than 82.5% and as the number of estimators rises, so do their values. It is almost 94% accurate. In general, the system performs at its best when n_estimators is around 250.

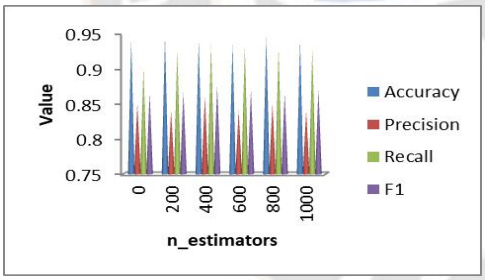

Figure 3. ANN performance measures

Four metrics about the parameter of the RF algorithm are tested using RF as the classifiern_estimators. Figure 4 displays the experiment's findings, and it is clear that the accuracy is greater than 96%, and the remaining three metrics all have a success rate of at least 90% and exhibit nearly constant values across a range of n-estimator values. When n_estimators is around 100, the system operates more efficiently.

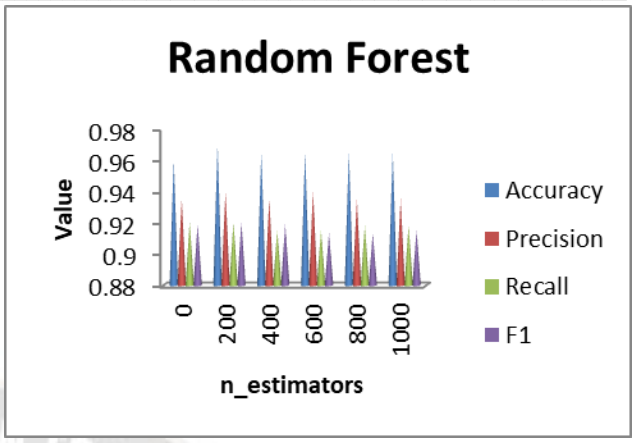

Figure 4. Random Forest performance measures

D. Negative and Positivistic sample distributions' efficiency

Investigate the impact of the negative/positive sample ratio in this case. The negative and positive ratios used to test the classifiers were as follows: 7.089, 3.499, 2.081, 1.375, 0.988, 0.727, 0.462, 0.346, 0.278, 0.139, 0.099, 0.075, and 0.058. Figures 5 – 7 display the findings. While all three other metrics rise along with the ratio, accuracy falls. The proposed ratio is 1 to 2.

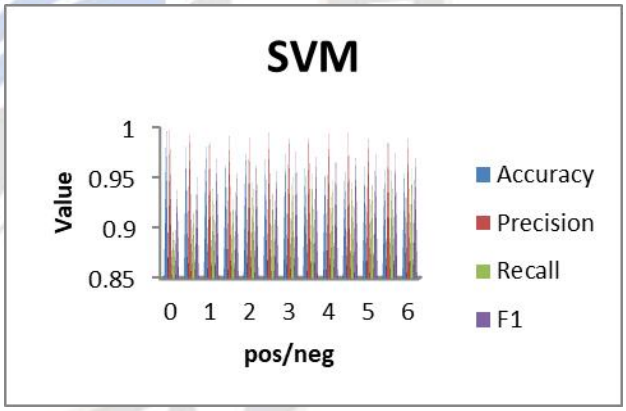

Figure 5. Result of SVM with different ratios

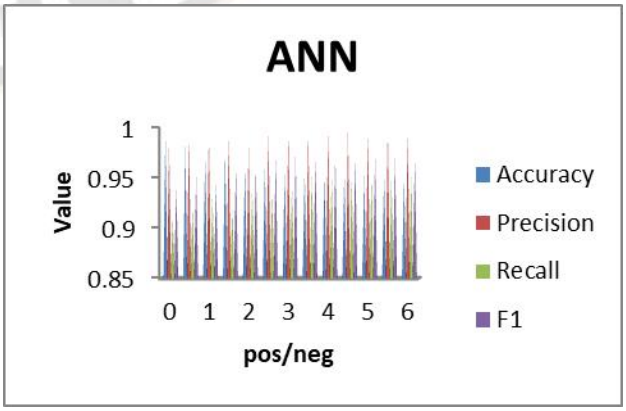

Figure 6. Result of ANN with different ratios

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

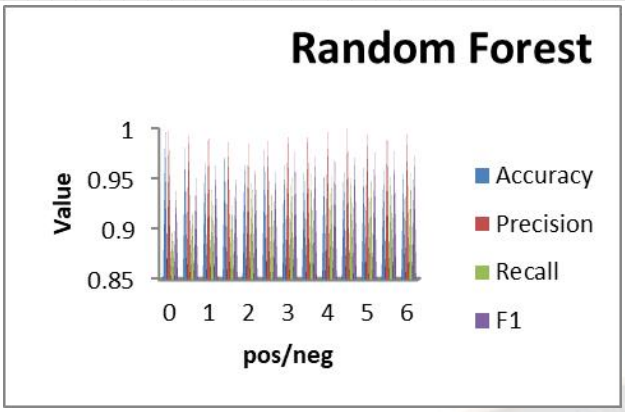

Figure 7. Result of RF with different ratios

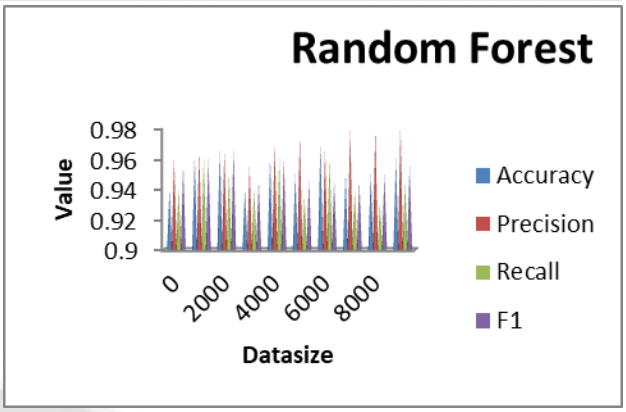

Figure 10. Results of RF using various training set sizes

E. Training data set performance measures

Lastly, consider how the results of detection are impacted by the size of the training set. Use the following subset sizes to test classifiers in this case: 803, 1622, 2357, 3201, 4043, 4909, 5628, 6425, 7230, and 8155, in which the positive to negative ratio is nearly 1. The outcomes are displayed in Figures 8 - 10. The precision is over 95%. Random Forest and SVM perform better as the size of the data increases, While the testing sample distributions of AdaBoost and Decision Tree show implicit tendencies.

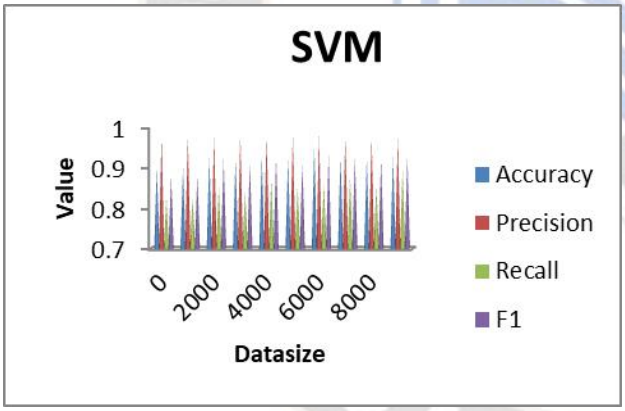

Figure 8. Results of SVM using various training set sizes

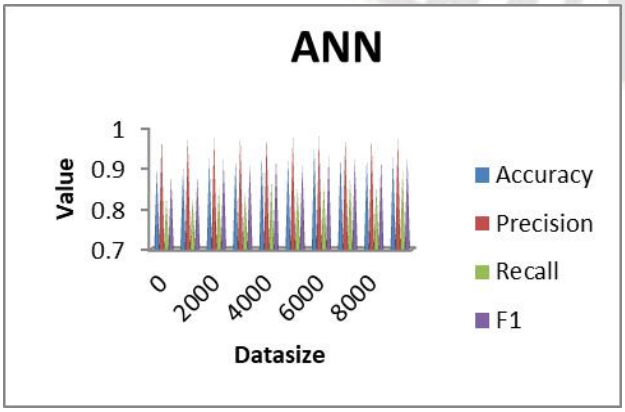

Figure 9. Results of ANN using various training set sizes

Based on the outcomes of the experiment, the best performance values for each classifier are displayed in Table 4. Random Forest outperforms the other three classifiers when all four metrics are considered. More than 84% F1 and over 93% accuracy are displayed by all classifiers, which proves our method can make a reliable phishing website detection.

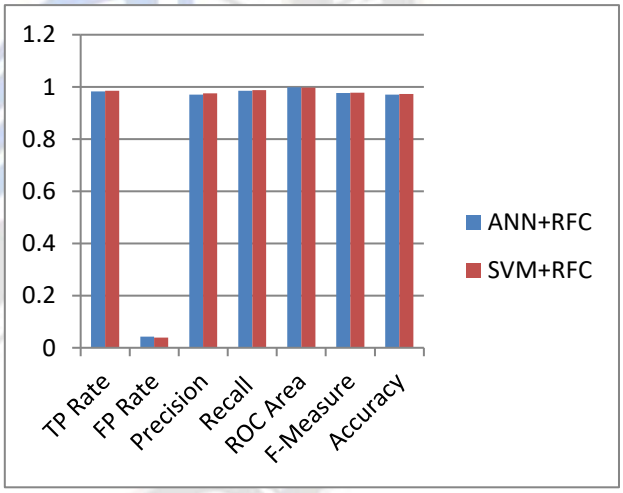

Figure 10. Performance measures

Table 4. Test results of the two classifiers

	ANN+RFC	SVM+RFC
TP Rate	0.983	0.985
FP Rate	0.043	0.039
Precision	0.971	0.975
Recall	0.985	0.988
ROC Area	0.999	0.998
F-Measure	0.977	0.978
Accuracy	0.97	0.973

V. CONCLUSION

In this study, a technique for quickly identifying a phishing attack on a website was covered. It is essential to detect phishing attacks. An ensemble of ML classifiers was

proposed to enhance the accuracy of attack detection in this study. RFC with two classifiers: ANN, and SVM. A combination of SVM and ANN improves the performance of the RFC. Based on CSS layout features, we train classifiers automatically to determine the similarity between web pages without requiring human expertise. Many phishing web pages were used to prototype our evaluation and approach its effectiveness. Our approach to determining similarity from page layouts was shown to be effective and accurate in the experiment. By enhancing existing antiphishing mechanisms, we can effectively improve their performance.

References

[1] Kalabarige, L. R., Rao, R. S., Abraham, A., &Gabralla, L. A. (2022). Multilayer stacked ensemble learning model to detect phishing websites. IEEE Access, 10, 79543-79552.DOI: 10.1109/ACCESS.2022.3194672

[2] Alsariera, Y. A., Balogun, A. O., Adeyemo, V. E., Tarawneh, O. H., &Mojeed, H. A. (2022). Intelligent tree-based ensemble approaches for phishing website detection. J. Eng. Sci. Technol, 17, 563-582. https://doi.org/10.3390/math9212799

[3] Atre, M., Jha, B., & Rao, A. (2022). Detecting Cloud-Based Phishing Attacks by Combining Deep Learning Models. arXiv preprint arXiv:2204.02446.https://doi.org/10.48550/arXiv.2204.02446

[4] Puri, N., Saggar, P., Kaur, A., & Garg, P. (2022, July). Application of ensemble Machine Learning models for phishing detection on web networks. In 2022 Fifth International Conference on Computational Intelligence and Communication Technologies (CCICT) (pp. 296-303). IEEE.DOI:10.1109/INMIC50486.2020.9318210

[5] Hota, H. S., Shrivas, A. K., &Hota, R. (2018). An ensemble model for detecting phishing attacks with proposed remove-replace feature selection technique. Procedia computer science, https://doi.org/10.1016/j.procs.2018.05.103

[6] Bidabadi, F. S., & Wang, S. (2022). A new weighted ensemble model for phishing detection based on feature selection. arXiv preprint arhttps://doi.org/10.48550/arXiv.2212.11125Xiv:2212.11125.

[7] Balamurugan, K., Latchoumi, T. P., &Ezhilarasi, T. P. (2022). Wearables to Improve Efficiency, Productivity, and Safety of Operations. In Smart Manufacturing Technologies for Industry 4.0 (pp. 75-90). CRC PressDOI:10.1201/9781003186670-9

[8] Mohan, A., Prabha, G. ., & V., A. . (2023). Multi Sensor System and Automatic Shutters for Bridge- An Approach. International Journal of Intelligent Systems and Applications in Engineering, 11(4s), 278–281. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2665

[9] Hossain, F., Islam, L., & Uddin, M. N. (2022, September). PhishRescue: A Stacked Ensemble Model to Identify Phishing Website Using Lexical Features. In 2022 5th International Conference of Computer and Informatics Engineering (IC2IE) (pp. 342-347). DOI: 10.1109/IC2IE56416.2022.9970179

[10] Garikapati, P., Balamurugan, K., & Latchoumi, T. P. (2022). K-means partitioning approach to predict the error observations in small datasets. International Journal of Computer Aided Engineering and Technology, 17(4), 412-430.https://doi.org/10.1504/IJCAET.2022.126601

[11] Livara, A., & Hernandez, R. (2022, January). An Empirical Analysis of Machine Learning Techniques in Phishing E-mail detection. In 2022 International Conference for Advancement in Technology (ICONAT) (pp. 1-6). IEEE.DOI: 10.1109/ICONAT53423.2022.9725434

[12] Ansari, M. F., Panigrahi, A., Jakka, G., Pati, A., & Bhattacharya, K. (2022, November). Prevention of Phishing attacks using AI Algorithm. In 2022 2nd Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON) (pp. 1-5). IEEE.DOI: 10.1109/ODICON54453.2022.10010185

[13] Pandey, M. K., Singh, M. K., Pal, S., & Tiwari, B. B. (2022). Prediction of Phishing Websites Using Stacked Ensemble Method and Hybrid Features Selection Method. SN Computer Science, 3(6), 488.DOI:10.1007/s42979-022-01387-4

[14] Prof. Barry Wiling. (2017). Monitoring of Sona Massori Paddy Crop and its Pests Using Image Processing. International Journal of New Practices in Management and Engineering, 6(02), 01 - 06. https://doi.org/10.17762/ijnpme.v6i02.54

[15] Latchoumi, T. P., Swathi, R., Vidyasri, P., & Balamurugan, K. (2022, March). Develop New Algorithm To Improve Safety On WMSN In Health Disease Monitoring. In 2022 International Mobile and Embedded Technology Conference (MECON) (pp. 357-362). IEEE.DOI: 10.1109/MECON53876.2022.9752178

[16] ShiraniBidabadi, F., & Wang, S. (2022). A new weighted ensemble model for phishing detection based on feature selection. arXiv e-prints, arXiv-2212.https://doi.org/10.48550/arXiv.2212.11125

[17] Bhowmik, P., & Bhowmik, P. C. (2022, October). A Machine Learning Approach for Phishing Websites Prediction with Novel Feature Selection Framework. In Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021 (pp. 357-370). Singapore: Springer Nature Singapore.DOI:10.1007/978-981-19-2445-3_24

[18] Shaaban, M. A., Hassan, Y. F., &Guirguis, S. K. (2022). Deep convolutional forest: a dynamic deep ensemble approach for spam detection in text. Complex & Intelligent Systems, 1-13.https://doi.org/10.1007/s40747-022-00741-6

[19] Thomas Wilson, Andrew Evans, Alejandro Perez, Luis Pérez, Juan Martinez. Machine Learning for Anomaly Detection and Outlier Analysis in Decision Science. Kuwait Journal of Machine Learning, 2(3). Retrieved from http://kuwaitjournals.com/index.php/kjml/article/view/207

[20] Catal, C., Giray, G., Tekinerdogan, B., Kumar, S., & Shukla, S. (2022). Applications of deep learning for phishing detection: a systematic literature review. Knowledge and Information Systems, 1-44.https://doi.org/10.1007/s10115-022-01672-x

[21] Balamurugan, K., Latchoumi, T. P., &Ezhilarasi, T. P. (2022). Wearables to Improve Efficiency, Productivity, and Safety of Operations. In Smart Manufacturing Technologies for Industry 4.0 (pp. 75-90). CRC Press.DOI:10.1201/9781003186670-9

[22] Mithra Raj, M., & Arul Jothi, J. A. (2022, October). Website Phishing Detection Using Machine Learning Classification

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11s
DOI: https://doi.org/10.17762/ijritcc.v11i11s.8182
***Article Received:** 24 June 2023 **Revised:** 27 August 2023 **Accepted:** 10 September 2023*

Algorithms. In Applied Informatics: 5th International Conference, ICAI 2022, Arequipa, Peru, October 27–29, 2022, Proceedings (pp. 219-233). Cham: Springer International Publishing.DOI:10.1007/978-3-031-19647-8_16

[23] Shmalko, M., Abuadbba, A., Gaire, R., Wu, T., Paik, H. Y., & Nepal, S. (2022). Profiler: Profile-Based Model to Detect Phishing Emails. arXiv preprint arXiv:2208.08745.https://doi.org/10.48550/arXiv.2208.08745

[24] Rao, R. S., Umarekar, A., &Pais, A. R. (2022). Application of word embedding and machine learning in detecting phishing websites. Telecommunication Systems, 79(1), 33-45.https://doi.org/10.1007/s11235-021-00850-6

[25] Almousa, M., Furst, R., & Anwar, M. (2022, September). Characterizing Coding Style of Phishing Websites Using Machine Learning Techniques. In 2022 Fourth International Conference on Transdisciplinary AI (TransAI) (pp. 101-105). IEEE.DOI:10.1109/TransAI54797.2022.00025

[26] Rathee, D., & Mann, S. (2022). Detection of E-mail phishing attacks–using machine learning and deep learning. International Journal of Computer Applications, 183(1), 7
DOI: 10.1109/ICCES54183.2022.9835846