5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024)

**An Ensemble Classification Model for Phishing Mail Detection**

Anirudh Sa, P Radha Nishantb, Sanjay Baithac,K Dinesh Kumard,\*

*a ,b, c,d Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Benagaluru, India.*

**Abstract**

Phishing emails, which masquerade as legitimate correspondence, are a significant modern hazard that cybercriminals use to deceive people into disclosing personal information. This paper presents the complex tactics used by cyber adversaries in the always changing field of phishing detection by using a real-time dataset and sophisticated classification tools to thoroughly examine email properties. In addition to detection, this paper highlights the need for constant watchfulness, delves into the psychological and social engineering elements included in phishing emails, and promotes continuous user awareness. Furthermore, the research broadens its scope to encompass the wider implications for enterprises, clarifying possible financial and reputational fallout. By exploring the misleading strategies, it offers helpful advice on strengthening defences and acts as a crucial manual for protecting private and corporate data against the ever-changing and persistent danger landscape that phishing attempts provide. This research gives an exploration of new insights by cybercriminals used for insertion of malicious software using emails. By using attributes like email text, Email type, the research identifies the signs of fraud emails, which is critically needed for the users.

*Keywords:* Phishing emails; Malicious; Email security; Ensemble approach; Classification.

1. **Introduction**

One cannot stress the value of cybersecurity in the digital age particularly when it comes to email. One of the most popular tools for both personal and professional lives is email, but it also poses a lot of vulnerabilities. Sensitive information is frequently sent by email hence protecting it against breaches is necessary to maintain integrity and privacy. This reliance on email is a challenge as phishing attacks happen a lot [1]. The goal of these sophisticated cyberattacks is to steal sensitive information like logins and financial details by using deceptive emails that look like legitimate sources. Phishing techniques are constantly evolving which make it a serious threat [2]. In order to detect and prevent these attacks cybercriminals are constantly adapting their strategies to evade standard security protocols. Thus, a dynamic and robust approach is necessary to improve mail security.

In the research on phishing email detection, two key limitations are observed. Firstly, many models are trained on datasets that don't accurately represent real-world email traffic which reduces their effectiveness in actual situations. Secondly, there's a reliance on manual feature extraction in some studies, which may not be practical for large-scale or dynamic email systems. These findings highlight the necessity for models that use more representative data and automated processes for feature extraction to enhance their applicability in real-world scenarios.

The key contributions of this study in phishing email detection include:

* Employing a stacking ensemble approach that combines the strengths of SVM and XGBoost models, with Logistic Regression making the final classification. This method offers a sophisticated and effective solution for identifying phishing emails.
* The proposed methodology’s high accuracy, recall, and F1 Score showcase its precision and reliability in classifying both phishing and safe emails.
* This proposed methodology's strategic use of insights from different models results in a more precise and robust email classification system.

Section 2 presents the Related work. Section 3 presents the proposed framework. Section 4 discusses the experimental results. Finally, the conclusion is presented in Section 5.

1. **Literature Review**

In the previous research, conventional methods were used for identifying phishing emails, such as blacklists and whitelists, URL analysis, spam filters, and manual reviews that required human intervention or IP address scanning. However, as technology advanced, many new methods were found that allowed hackers to avoid detection, making them difficult to identify. By using machine learning algorithms, this problem can be solved. Phishing emails can be identified by machine learning with a higher accuracy and better performance.

Pankaj et al. [4] proposed how to combat email phishing using Naive Bayes and Support vector machine (SVM). This comparative analysis presented various models including Random Forest, Logistic regression, and Voted Perceptron algorithm. Weina et al. [5] presented detection of phishing emails using Cuckoo Search Support Vector Machine (CS-SVM). In this paper, 23 features were extracted and trained and compared with the baseline SVM model. Mustafa et al. [6] proposed framework which can be addressed the solution using classification to categorize emails as phished or not. 3 methods are used to check the efficiency Naive Bayes classifier, SVM, and Random Forests classifier. J. Ramprasath et al. [7] presented the identification of phishing email attacks. Training the model using Long Short-Term Memory (LSTM) cells and the comparative analysis is done using SVM and Constrained k-Nearest Neighbor (CkNN) classifier. Lingampally et al. [8] presented the detection of phished email using Linear Regression and logistic regression. In this paper, 11 features were extracted and took LA-BEL as a target attribute and removed its imbalance using tokenization, lemmatization and vectorization.

Nishant et al. [9] proposed the detection of email phishing using CNN and LSTM models. Comparative analysis is done using various models such CNN, LSTM, Bi-LSTM, Naive Bayes, SVM, Decision Tree Classifier, and Logistic Regression. Sikha et al. [10] presented the classification of phishing email using CNN and Decision tree. Comparative analysis is done on various models like LSTM, CNN, and Word Embedding, and classifiers like Naïve Bayes, SVM, and Decision Trees. Md Fazale et al. [11] proposed vectorization of dataset and train using Methods like Logistic Regression (LR), AdaBoost (AB), K-Nearest Neighbors (KNN),), Multinomial Naive Bayes (MNB), Random Forest (RF), and Gradient Boosting (GB). The dataset is vectorized using Term-frequency inverse frequency document frequency.

Md Abdullah Al Ahasan et al. [12] advocated the use of Improved Random Forest (IRD) and Optimized Fuzzy Multi-Criteria Decision Making (OFMCDM) to identify phishing emails. OFMCDM model is used to extract the features and IRD for training the model. Comparative analysis is done with other models that include Naïve Bayes (NB), Logistic Regression (LR), KNN, and Decision Tree. Ishwarya et al. [13] proposed the detection of phishing emails using probabilistic classifiers like naïve bayes, KNN, SVM, and Random Forest. The features are extracted as input and output manually without using any algorithms. Sadia Parvin Ripa et al. [14] proposed a comparison of different algorithms like Naive bayes, Random Forest, Decision tree and others. Around 9499 entries constitute the dataset used for phishing emails. Manually, five features were extracted. XGBoost was used to train the dataset, and it was then compared to other models such as KNN, Random Forest, Decision Tree, and others.

Sami Smadi et al. [15] presented the 6 different types of classification algorithms such as C4.5, Naive Bayes, Linear Regression, SVM, and KNN. The model uses the J48 algorithm to express the structure of the data and ten-fold cross validation is used to test and train the model. The model used 23 hybrid features to demonstrate the results to classify the emails. Prasanta Kumar Sahoo et al. [16] present the six different classification algorithms—Naive Bayes, SVM, C4.5, Linear Regression, and KNN—that are used to identify emails as phishing or authentic. The structure of the data is expressed by the model using the J48 technique, and it is tested and trained using ten-fold cross validation. 23 hybrid traits were employed by the model to show how the emails were classified. Isredza Rahmi A Hamidet et al. [17] proposed the implementation of phishing email detection used a feature selection based on hybrid model, content-based, and behaviour-based. Seven features were extracted using hybrid feature selection and trained using data mining algorithms. Three different datasets were taken from various sources which was performed using WEKA (Waikato Environment for Knowledge Analysis) which consists of around 3000 phishing emails that were collected from November 2004 to November 2005.

Mahmoud Khonji et al. [18] explore a method for distinguishing phishing emails by analysing URLs leading to malicious websites. This approach incorporates lexical URL analysis, where the model segments feature into subsets before validation - an unconventional but effective method. The process involves developing unique and efficient features through regular algorithms. In this study 47 features were extracted, utilizing both wrapper and best-fit approaches. Two distinct datasets were used where one comprises 4,116 emails and another with 4,150 emails. The comparative analysis is done on the subsets that were created when features were taken. Gal Egozi et al. [19] discusses how to identify phishing emails with strong natural language processing algorithms. About 26 characteristics were retrieved. A comparison of about 17 models, out of those,14 produced results that were acceptable. The dataset was gathered from Wikileaks, SpamAssassin, Nazario Phishing corpora and other sources for the IWSPA competition, which included about 9000 submissions. Regardless of whether they were weighted or not, a single machine learning algorithm—Bernoulli's Naïve Bayes, Decision Tree, Linear Kernel SVM, and Gaussian Naïve Bayes were compared.

To identify the phishing emails, conventional deep learning models and baseline machine learning models are insufficient. These eloquently demonstrate the advantages of utilizing a hybrid model, which combines many baseline models applied to various dataset segments. In order to more accurately predict phishing emails, the proposed framework utilised an ensemble model. This ensemble model utilised SVM and XGBoost. Logistic regression was taken as a meta learner to make the final decision for the data that is generated after it has been trained in SVM and XGBoost.

1. **Proposed Methodology**

The proposed framework for enhanced classification of phishing emails employs a stacking ensemble approach. This approach integrates the strengths of SVM and XGBoost models as base learners, with Logistic Regression acting as the final decision-maker.

*3.1 Data Preprocessing*

In the data pre-processing step, at first data cleaning technique is used to involve in conversion to lowercase, removal of special characters and numbers, and elimination of stop words. This is followed by stemming, where words are reduced to their root forms. The processed text is transformed into numerical features using TF-IDF Vectorization. The dataset is also checked for missing values in crucial columns to ensure completeness. Finally, categorical labels are converted into numerical form through Label Encoding, preparing the data for subsequent analysis and modelling.

Figure 1 presents a logarithmic scale displaying email duration versus word count, which allows for the observation of patterns and relationships between the two metrics. The association between email verbosity and wordiness—which may be a sign of safe or phishing emails—is better understood with the use of this scatter plot.

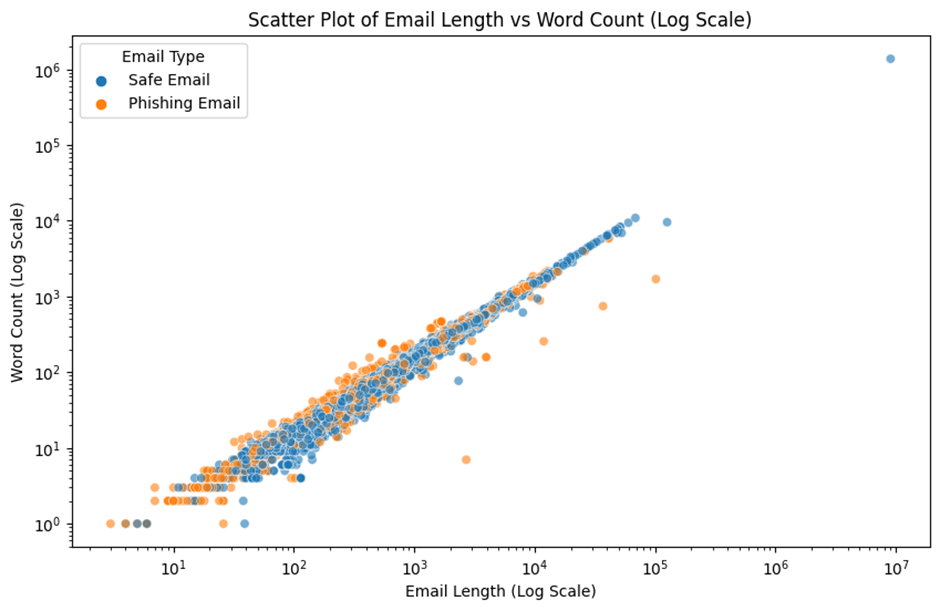


Fig. 1. Relation between Email length and word count

*3.2 Support Vector Machine*

Support Vector machine (SVM) [21] is a classification-based algorithm that are known for their effectiveness in classification tasks. SVM is employed in numerous applications ranging from image recognition to biometrics. SVM is based on the concept of finding the optimal hyperplane which is a boundary that maximizes the margin between two classes of data point. This hyperplane usually visualizes as a line in 2-dimensional space or a plane with higher dimensions which separates the two classes ensuring accurate classification. SVM can work really well with high dimensional data in the real world. Most data exist in multiple dimensions and SVM overcomes this by employing kernel functions that project the information into a higher-dimensional space, which makes linear separation feasible. SVM is also not influenced by outliers unlike other models as it considers the dataset closes to the hyperplane which results in more accurate and reliable prediction. Unfortunately, SVM can be technically expensive especially for complex datasets and require careful hyperparameter tuning to achieve optimal performance. It's vital to select the appropriate kernel function and tune its hyperparameters otherwise it would affect the model’s ability to handle complex datasets. Machine learning problems are nevertheless often solved using SVM, regardless of challenges faced.

*3.3 XGBoost*

A popular machine learning algorithm called XGBoost [22] is utilized for producing incredibly precise results because it creates a single, reliable model by merging many decision trees which makes this technique incredibly efficient. The total strength of XGBoost is increased as each tree in the sequence learns from the mistakes made by the trees before it. Additionally, a regularization approach is used to assist prevent overfitting and guarantee that the model performs well when applied to fresh, untested data. XGBoost is well known for its efficiency and speed and it performs exceptionally well with large and complex datasets because of these features, XGBoost is preferred among data scientists, especially for predictive modelling jobs like identifying emails as spam or not.

*3.4 Logistic regression*

Logistic regression is a statistical classifier and is used for binary classification. It operates probabilistically such that a given input belongs to a specific class with the help of logistic function. This function maps with any real valued input within a probability of 0 and 1.The model learns by multiplying input features with learned coefficients and adding a bias. It passes the result to the logistic function to predict a probability. In ensemble methods logistic regression can act as an optimal meta learner combining predictions from various models into a final decision by effectively weighing the reliability of different predictors to enhance overall predictive accuracy.

*3.5 Proposed Architecture*

The stacking ensemble method is an effective strategy for classifying data like emails. This approach utilizes various algorithms known as base learners, such as SVM and XGBoost (extreme Gradient Boosting), each making predictions independently. These learners analyse the data separately and give individual predictions. These predictions are then compiled in the combination layer. After this, a meta learner is trained on the combination of these predictions. It learns the best way to use these predictions by determining the right weights for them during the training. The formula for the meta learner is:

(1)

*P* is the final probability indicating whether an email is 'Phishing' or 'Safe’. 𝜎 is the logistic function, converting the sum of predictions into a probability between 0 and 1. *w*₁ and *w*₂ are the weights assigned to the predictions from SVM and XGBoost and Equation 1 is the predictions from the SVM and XGBoost models respectively. *P* would be compared to the threshold value which is set as 0.5 and if *P* is greater than the threshold value it would be classified as 1(safe) or else it would be 0(phishing). A subset of considered dataset which contains the email data labelled as ‘phishing’ or ‘safe’ is used to understand the composition of the input layer and combination layer. The input layer contains the emails and their types as presented in Table 1. Combination Layer consists of the predictions of the base learner and for the above dataset, the combination layer is presented as mentioned in Table 2.

Table 1. Sample Example

|  |  |  |
| --- | --- | --- |
| Index | Email Text | Email Type |
| 1 | I can't seem to build this package. It errors out because rpm found files not included in any of the package. | Safe Email |
| 2 | More Than $2500 in DEBT?We Can Help You PAY-OFF Your BILLS!! | Phishing Email |
| 3 | open season results attached are the results of the open season . | Safe Email |
| 4 | we owe you lots of money dear applicant, after further review upon receiving your application your ... | Phishing Email |

Table 2. Results of base learners

|  |  |  |
| --- | --- | --- |
| Index | SVM Prediction | XGBoost Prediction |
| 1 | Safe | Safe |
| 2 | Phishing | Phishing |
| 3 | Safe | Safe |
| 5 | Phishing | Phishing |

The stacking ensemble method is a useful tool, especially in identifying phishing emails. It combines different machine learning techniques to make a more accurate model.



Training Dataset

Combinational Layer





Phishing or Safe

Fig. 2. System Architecture

1. **Experimental Analysis and Result Discussion**

This experimental dataset [23] has 18,650 emails, each tagged according to one of two categories in the "Email Type" column. For emails that are harmless it's ‘Safe Email’, and for emails that deceive people it's ‘Phishing Email’. 11,322 emails marked safe mean they're not malicious. There are also some missing values in this dataset, and the remaining 7,328 are labelled 'Phishing Email' which means they're dangerous. In 16 emails, there is no 'Email Text' column, which is needed to understand the content. To evaluate the efficiency of the proposed methodology, considered four evaluation metrices as presented in Equations (2-5). SVM, Random Forest, Multinomial-Naïve bayes and XGBoost algorithms were used to compare the results of proposed method.

(2)

(3)

(4)

(5)

Multinomial Naive Bayes (MNB) obtained accuracy of 96.22 which is slightly lower than the proposed method. It received a precision score of 96.36, which is also worse than any other method, identifying phishing in the email. It also received a recall score of 97.36 which shows that the false negative affects the MNB a lot less compared to the proposed method, and an overall F1-score of 96.96. It classified 1624 non-phishing emails and 2730 phishing emails correctly. It classified 97 emails as non-phishing emails and 74 phishing emails incorrectly. In comparison with the proposed model, SVM obtained an accuracy of 96.74% which is slightly low. It received a precision score of 98.05%, which is still very good for correctly identifying phishing in the email. It also received a recall score of 96.52%, which shows that the false negative effect the model a lot more and an overall F1-score of 97.28%. It classified 1790 non-phishing emails and 2721 phishing emails correctly. It classified 54 emails as non-phishing emails and 98 phishing emails incorrectly.

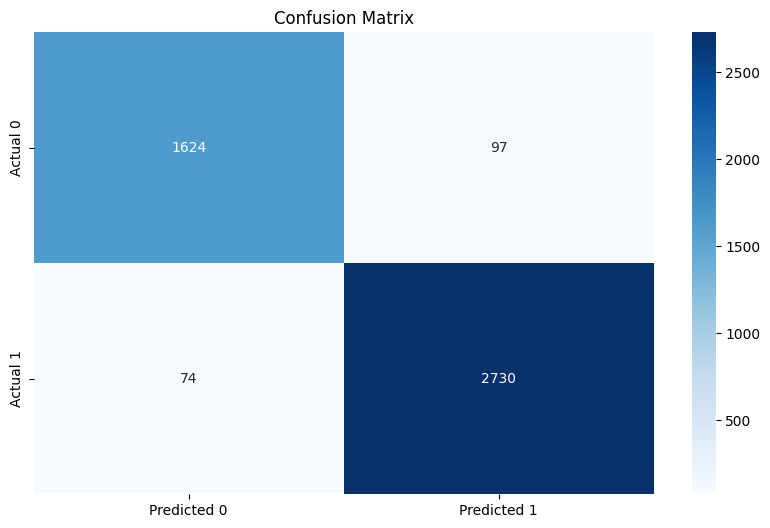
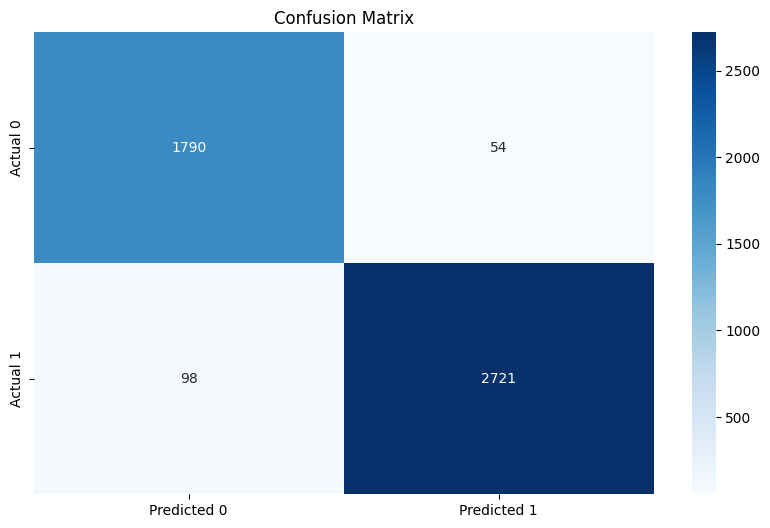
 

Fig. 3. Confusion Matrix for (a).Multinomial Naive Bayes and; (b). SVM

XGBoost obtained an accuracy of 96.05% which is slightly lower than any other model. It received a precision score of 97.75% which is still good for correctly identifying phishing in the email. It also received a recall score of 95.67% which shows that the false negative affected the model a lot more than any other model, and an overall F1-score of 96.70%. It classified 1782 non-phishing emails and 2697 phishing emails correctly. It classified 122 phishing emails and 62 emails as non-phishing emails incorrectly. Random Forest obtained an accuracy of 97.06% which is slightly low. It received a precision score of 97.92% which is still good for correctly identifying phishing in the email. It also received a recall score of 97.33% which shows that the false negative affected the model a lot more than the proposed method, and an overall F1-score of 97.62%. It classified 1663 non-phishing emails and 2729 phishing emails correctly. It classified 58 emails as non-phishing emails and 75 phishing emails incorrectly.

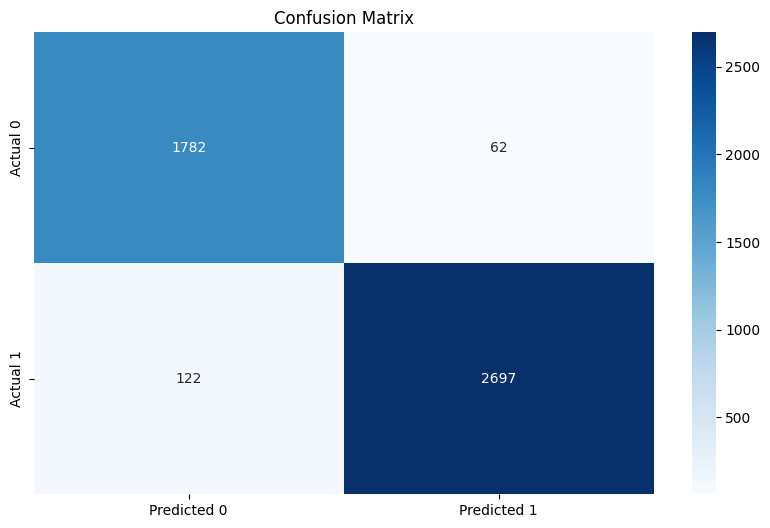
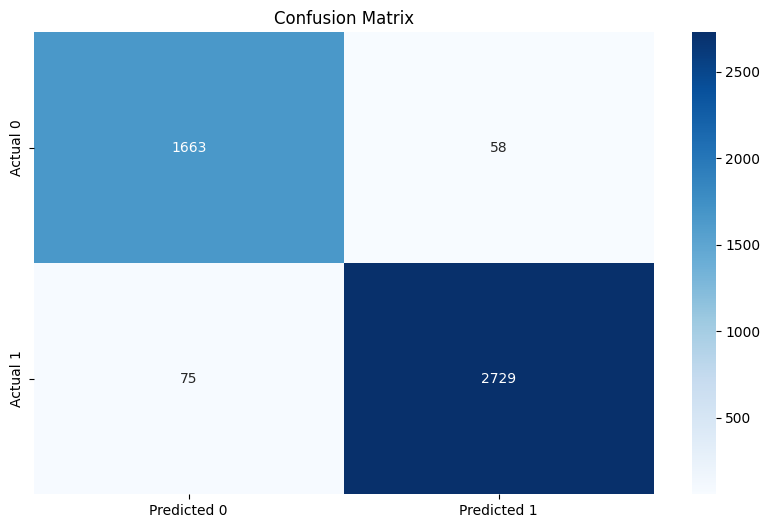
 

Fig. 4. Confusion Matrix for (a). XGBoost and; (b). Random Forest

The proposed method's accuracy is 97.92% higher than that of other models. When compared with other models, its precision score of 98.43% is marginally higher. Additionally, the suggested model has a recall of 98.22%, which is much higher than any previous models. The suggested approach has a 98.32% overall F1-score. The proposed method yielded 1677 emails successfully identified as safe emails, 2754 emails were classified as a phishing emails, 44 correctly identified as non-phishing emails, and 50 incorrectly identified as phishing emails.

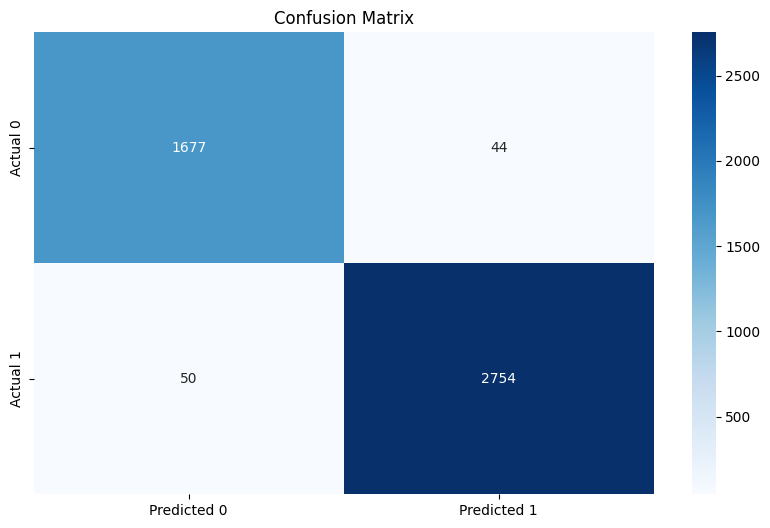


Fig. 5. Confusion matrix for proposed ensemble method

**5. Conclusion**

This research concludes by highlighting the effectiveness of ensemble learning to classify the phishing emails and emphasizing its superiority over conventional methods. The model provided in this study shows a significant boost in accuracy when identifying phishing attempts by smoothly combining various machine learning techniques. Combining different algorithms strengthens the system's capacity to identify minute details and patterns typical of phishing techniques, strengthening email security more thoroughly than traditional methods. This cybersecurity technology breakthrough is a significant development, demonstrating the flexibility and effectiveness of contemporary machine learning in tackling the complexities of dynamic threats. The ensemble approach to learning not only raises the standards for email security, but it also emphasizes the importance of continuous innovation in order to stay ahead of cyber adversaries in the ongoing cat-and-mouse game. This study highlights the transformative potential of leveraging various machine learning techniques and serves as a testament to the ever-changing landscape of cybersecurity, where modern solutions are indispensable in navigating the complexities of emerging threats. This breakthrough in cybersecurity technology using ensemble learning accentuates the adaptability and the efficiency of contemporary machine learning as a remedy for challenges that were created by dynamic threats.

**Reference**

[1] Hiransha M, Unnithan NA, Vinayakumar R, Soman K, Verma AD (2018) “Deep learning-based phishing e-mail detection” *1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA),* 1-5).

[2] Harikrishnan NB, Vinayakumar R, Soman KP (2018) “A machine learning approach towards phishing email detection” *In Proceedings of the Anti-Phishing Pilot at ACM International Workshop on Security and Privacy Analytics (IWSPA AP),* 455-468

[3] Unnithan NA, Harikrishnan NB, Vinayakumar R, Soman KP, Sundarakrishna S (2018) “Detecting phishing E-mail using machine learning techniques.” *In Proc. 1st anti-phishing shared task pilot 4th acm iwspa co-located 8th acm conf. data appl. secur. privacy (codaspy),* 51-54.

[4] Saraswat P, Solanki MS. (2022) “Phishing Detection in E-mails using Machine Learning”. *In 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS),* 420-424*.*

[5] Niu W, Zhang X, Yang G, Ma Z, Zhuo Z (2017). “Phishing emails detection using CS-SVM”. *In 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC),* 1054-1059.

[6] Al Fayoumi M, Odeh A, Keshta I, Aboshgifa A, AlHajahjeh T, Abdulraheem R (2022). “Email phishing detection based on naïve Bayes, Random Forests, and SVM classifications: A comparative study*”.  In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC),* 0007-0011.

[7] Ramprasath J, Priyanka S, Manudev R, Gokul M (2023). “Identification and mitigation of phishing email attacks using deep learning”. *In* *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE),* 466-470.

[8] Shalini L, Manvi SS, Gowda NC, Manasa KN (2022). “Detection of Phishing Emails using Machine Learning and Deep Learning”. *In 2022 7th International Conference on Communication and Electronics Systems (ICCES),* 1237-1243.

[9] Paradkar NS (2023). “Phishing Email’s Detection Using Machine Learning and Deep Learning”. *In 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS),*  160-162.

[10] Bagui S, Nandi D, Bagui S, White RJ (2019) “Classifying phishing email using machine learning and deep learning”. *In 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security),* pp. 1-2.

[11] Rabbi MF, Champa AI, Zibran MF (2023) “Phishy? Detecting Phishing Emails Using ML and NLP”. In *2023 IEEE/ACIS 21st International Conference on Software Engineering Research, Management and Applications (SERA)*, 77-83.

[12] Almomani A, Gupta BB, Atawneh S, Meulenberg A, Almomani E (2013) “A survey of phishing email filtering techniques”. *IEEE communications surveys & tutorials***15(4)**:2070-2090.

[13] Ishwarya R, Muthumani S, PG SS, Suriya S (2023) “Seperation of Phishing Emails Using Probabilistic Classifiers”. *Advanced Computing and Communication Systems 1,* 1676-1679.

[14] Ripa SP, Islam F, Arifuzzaman M (2021) “The emergence threat of phishing attack and the detection techniques using machine learning models*”. In 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI),* 1-6.

[15] Chiew KL, Tan CL, Wong K, Yong KS, Tiong WK (2019) “A new hybrid ensemble feature selection framework for machine learning-based phishing detection system”.  *Information Sciences* *484*, 153-166.

[16] Dewis M, Viana T (2022) “Phish responder: A Hybrid machine learning approach to detect phishing and spam emails”. *Applied System Innovation*,***5*(4)**: 73.

[17] Al Ahasan MA, Hu M, Shahriar N (2023) “OFMCDM/IRF: A Phishing Website Detection Model based on Optimized Fuzzy Multi-Criteria Decision-Making and Improved Random Forest”. *In 2023 Silicon Valley Cybersecurity Conference (SVCC),* 1-8*.*

[18] Hamid IR, Abawajy J, Kim TH (2013) “Using feature selection and classification scheme for automating phishing email detection”.

[19] Khonji M, Iraqi Y, Jones A (2011) “Lexical URL analysis for discriminating phishing and legitimate e-mail messages”. *In 2011 International Conference for Internet Technology and Secured Transactions 2011 Dec 11*, 422-427.

[20] Vishagini V, Rajan AK (2018) “An improved spam detection method with weighted support vector machine”. *In 2018 International Conference on Data Science and Engineering (ICDSE),* 1-5.

[21] Laxmikant K, Bhuvaneswari R, Natarajan B (2023) “An Efficient Approach to Detect Diabetes using XGBoost Classifier”. *In 2023 Winter Summit on Smart Computing and Networks (WiSSCoN),* 1-8.

[22] Ra V, HBa BG, Ma AK, KPa S, Poornachandran P, Verma A (2018) “DeepAnti-PhishNet: Applying deep neural networks for phishing email detection”.  *In Proc. 1st AntiPhishing Shared Pilot 4th ACM Int. Workshop Secur. Privacy Anal.(IWSPA),* 1-11.

[23] Subhadeep Chakraborty. (2023). Phishing Email Detection [Data set]. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/6090437> (Accessed on Sep 2023)