Machine Learning Classifier Performance Comparison for Phishing Detection

Iqbal Hadi Azmi
Faculty Computer Science and Information Technology
Universiti Putra Malaysia
Malaysia
gs61356@student.upm.edu.my

Nor Fazlida Mohd Sani
Faculty Computer Science and Information Technology
Universiti Putra Malaysia
Malaysia

***Abstract*—Technology has advanced at a remarkable rate in recent decades, making communication simpler. Emails are the most effective method for both casual and formal conversations when compared to other forms of communication. Emails are a common form of communication for both business and personal purposes. Unfortunately, emails are also used to annoy internet users by sending viruses, spam, and ads. Spam emails are those sent by some undesirable users, also referred to as spammers. Some individuals misuse this kind of communication by sending spam emails that contain links to specific URLs for users to click on. Spam generates a number of issues, some of which may result in financial losses. Phishing is a method for attempting to obtain sensitive data through fraudulent email or website solicitation. This study compares the effectiveness of different machine learning algorithms in identifying spam or phishing emails. Different performance criteria were taken into account when evaluating the models, and the outcomes were compared.**

Keywords—Phishing Detection, Spam email, Machine Learning Techniques

I. INTRODUCTION

Spam is any irrelevant and undesired communication or unwanted email that is sent by the attacker via email or another information-sharing medium to a significant number of recipients[1]. Spam causes unnecessary use of server resources since there are so many unsolicited emails that need to be processed. Over 77% of all email traffic worldwide is spam, which poses an increasing threat on an annual basis. Users who get spam emails find it annoying. Internet scams and other dishonest tactics used by spammers to trick users into disclosing sensitive personal information have affected many users. According to statistics, spam emails made up 56.87% of all email traffic worldwide, with dating and healthcare spam being the most common types[2].

Phishing incidents are the most frequent attacks carried out by social engineers in recent years. Through phone calls or emails, they seek to deceptively get private and personal information from their intended targets. Attackers deceive victims in order to get private and sensitive information. They involve clicking on a link included in the emails, visiting fake websites, emails, PayPal websites, and so on. Phishing attacks are the most frequent attacks executed by social engineers [3-4]. They aim to deceitfully get private and secret information from intended targets via phone calls or emails [5]. Attackers trick their victims into giving them access to private and sensitive data, including credit card numbers and any other information that can be used to log into accounts with high security, including online banking or services [6].

Spam emails are now more prevalent for a variety of purposes, including advertising, multi-level marketing, chain letters, political correspondence, stock market advice, and so forth. Since machine learning algorithms can adapt to different situations, they have the potential to learn and recognize phishing messages. Based on what the machine has learned, new rules were developed and used during the spam filtering process.

The spam filtering decisions were updated in light of the contents using these dynamic approaches for recognizing the contents of the emails. Machine learning algorithms frequently utilize content-based filtering to provide automated filtering rules and categorize emails. In accordance with [7], the frequency and distribution of terms and phrases in email content were examined. The developed rules were used to filter incoming email spam. Utilizing an adaptive spam filtering technique, spam is identified and eliminated by grouping it into several categories, with a representative text in each one. Every email message is compared to each group and a percentage of similarity is created, determining the most likely category to which the message refers [8].

II. RELATED WORKS

A. Spam Filtering Techniques

Machine learning methods, such as the Bayesian Naive classification, K Nearest Neighbour, Neural Networks, and Support Vector Machine, frequently use content-based filtering to develop automatic filtering rules and categorise emails. This method often examines keywords, case, and distribution of words or phrases inside email text before using rules defined to filter received email spam [7]. This method uses pre-made rules or heuristics to test a large number of patterns against a selected message, which are often regular expressions. The rating of a message is improved by several comparable trends.

e-ISSN: 2550-1550

However, if some of the patterns do not fit, the score is reduced. While certain ranking criteria do not vary over time, others must be updated on a regular basis to deal with the threat of spammers consistently adding fresh spam emails that can be easily avoided without email filter detection. Spam Filtering Technique Based on Likeness is utilised to find emails that have been received and are concentrated on similarity samples for cases that have been preserved. This method combines methodologies based on examples with memory-based machine learning. The multi-dimensional vector of space, which used plot new file as points, was created using the email properties. The new file is then put into the same class as its K closest training sets [9].

To reduce spam email, a number of methods and spam filtering systems have been created utilising various ideas and algorithms. A number of protocols are used by the method of filtering email spam to determine whether or not the message is spam. Through these, the spam filtering standard mechanism operates as protocol-set classifiers and adheres to a set of principles [8].

One of the popular machine learning techniques among various spam filtering techniques is case based or sample base filtering. Through the collection model, all emails, spam and genuine, were gathered from individual email users. The initial transformation then starts with the pre-processing phases through the user interface, including data collection and extraction, email classification, process evaluation, and data classification into two groups using vector language. The next step is to use machine learning algorithms to training sets and testing sets to determine if an email is spam or genuine. To assess if an email is spam or not, the classifier's results will be used [10].

B. Machine Learning Techniques

AI uses the machine learning technique to train computers to process data more efficiently. Machine learning is employed in this case since it may not be possible for a human to interpret the pattern or extract information from the data after looking at it. Machine learning aims to learn new things from data. Numerous studies have been conducted on the subject of teaching robots to learn on their own.

The type of Machine learning is illustrated in several forms in Figure 1.

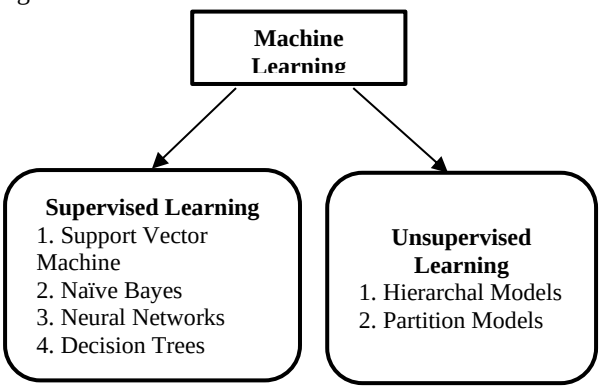

Fig. 1. Types of Machine Learning

In supervised learning, models that can predict new data are trained using tagged data. Numerous issues, including the popularity of adverts, the classification of spam, face recognition, and object classification, can be resolved with this kind of training. Unsupervised learning uses unlabeled data to group the data into clusters depending on its characteristics. This type of learning can be used to solve a variety of problems, such as aggregating user logs, identifying buying trends, and recommender systems. Machine learning methods are frequently used in phishing detection to differentiate between phishing and non-phishing. The detecting task is carried out by machine learning algorithms utilising an automated and adaptive method.

1) Decision Tree

The result of a decision tree classification is a binary tree structure known as a decision tree, where each branch node expresses an option among various options and each leaf node denotes a classification or decision-making. A decision tree model offers guidelines for forecasting the target variable. This technique scales effectively even though there are various numbers of training instances and several attributes in large databases. The drawback of this technique is that even a small change in the datasets can have a big impact on how the decision tree is built, leading to model instability. It fails because of the limited number of data characteristics [10,12].

2) Naïve Bayes

In a study on Naive Bayes email spam filtering that was done in [13], the performance metrics used were F-measure, recall, and accuracy. A dataset's frequency and value combinations were counted in terms of probability. Two datasets were used to assess the Naive Bayes algorithm's output. Similar to this, the metrics for assessment are precision rate, recall rate, and accuracy rate, where the precision is 99.87%, recall is 91.72%, and accuracy is 93.20%. It was suggested in [14] to use semantic approaches to filter emails effectively. The enormous number of textual features deleted was minimized using a variety of semantic-based methods employing ontology and similarity stages, which decreased the complexity of space and time.

The authors of [15] discussed how to classify emails as spam or not using the Naive Bayes method and Support Vector Machine algorithms. Calculations were made for the accuracy, recall, precision and F-measure. Spam Data and Spambase, two datasets for the Naive Bayes algorithm, were used. Utilising spam data increases precision and reliability. Spambase's accuracy, which is 88%, is higher than that of Spam Data, which is lower.

A simple probabilistic classifier, the Naive Bayes algorithm computes a series of probabilities by determining the frequency and combinations of values in a dataset Based on the Bayesian theorem, the Naive Bayes approach is particularly effective when the input dimensionality is high. Naive Bayes classifiers make the assumption that a variable's impact on a particular class is independent of the effects of other variables [13].

Naive Bayes classifiers currently seem to be particularly popular for both open-source and commercial spam filters. This may be attributed to their ease of implementation, linear computational complexity, and precision, which are equivalent to more advanced learning algorithms used in spam filtering [15].

By computing the conditional probability of the classes given the instance based on the precise existence of the probability model, the Naive-Bayes inducer determines the highest posterior class. One advantage of the Naive Bayes Classifier is that it is a very good classifier that has been used in many information processing applications. In a supervised learning environment, it is possible to effectively teach naive Bayes classifiers. When the information attributes in the training datasets are connected, the Naive Bayes classifier performs badly, which is its fundamental weakness. The data components need to be independent [10,12].

3) Neural Networks

A mathematical or computer structure called a "neural network" is based on the composition and/or operation of biological brain networks. The majority of the time, a neural network that is composed of an integrated artificial neuron community uses the connectionist method to process information. The main benefit is an adaptive mechanism that is dependent on knowledge that is circulated through the learning process, modifying its structure, either internally or externally. Non-linear statistical data can be modeled using it. Typically, they are employed to illustrate complex relationships or to look for patterns in the data between inputs and outputs. The drawback is that it can encourage over-fitting. The weights created for the training data may not be applicable to other datasets from the same populations [12,15]. From the viewpoint of numerous information networks, it is also possible to obtain different latent feature elements. As a second method of solving the same issue, the identification of spammers using a Deep Graph neural network is demonstrated 16].

4) Support Vector Machine

Support Vector Machines (SVM) are supervised or out-performed learning models that analyse data and find patterns through regression and classification using related learning algorithms and successful generalisation. SVM displays examples as points in space that are mapped so that the instances of the various categories are separated by a simple distance as wide as possible. By identifying two or more groups with a margin-maximizing hyperplane, SVM may solve the quadratic programming problem with inequality constraints and linear equality [17].

Based on the provided training data, SVM generates a hyperplane, a two-dimensional line that best delineates the categories [18]. The border for judgement is the name of this hyperplane. A number of features used in phishing detection specify input, such as the existence or absence of a specific term, and output, which is either 1 or -1, indicating whether the email is phished. Because of its quick processing time and effective text classification ability, SVM is a well-liked supervised method. SVM is one of the most effective classification algorithms, producing high classification accuracy. Training time increases in SVM with a high penalty parameter value (C). Testing and training error trade-offs are also made the option of the value of C [12]. In order to develop and assess a machine learning model for spam detection, SVM is utilised in [18]. 58 columns make up a dataset that is used as the spambase for the machine learning model. The dataset was cleaned and processed to make sure there were no null values.

The Support Vector Classifier produced an accurate result of roughly 89.21% after accuracy training and testing. The SVM algorithm was discovered to be the best choice since it accurately distinguishes between spam and non-spam emails [15].

5) Logistic regression

Using the binary logistic model, logistic regression is used to calculate the likelihood of a binary response based on one or more predictors (independent variables or features). It enables one to draw the conclusion that the presence of a risk factor raises the likelihood of a particular outcome by a specific percentage. Analysing is simpler and less challenging. Furthermore, a normal distribution need not contain identical variation in the independent variables. Since it is incorrect to assume that the relationship between the independent and dependent variables is one of linear effects, it can also take into consideration nonlinear effects. The average prediction accuracy of logistic regression is low, and it cannot handle several unacceptable characteristics well [18].

6) Random Forest

An ensemble bagging machine learning algorithm is Random Forest. The decision tree approach just creates one decision tree, whereas Random Forest creates multiple decision trees based on the various subsets of the replacement dataset. The votes of each individual tree decide the algorithm's outcome. Due to its random selection of many data sets that reduce the likelihood of over fitting, the Random Forest machine learning algorithm becomes resistant to variance in the collection of data. As a result, it produces the best classification results. In compared to the decision tree, it will perform well if the dataset changes

Prior to building a powerful classifier, speed training is sluggish and directly depends on the number of classifiers that must be taught. Due to the amount of decision trees that must be computed in order to develop an effective technique, this is not transparent as a decision tree. The results are therefore challenging to interpret [12]. Similar to how it is used in [19] to train and assess a machine learning model for spam detection, Random Forest is also employed in this case whereas the accuracy of the result from the Random Forest Classifier was about 91.36%.

Support Vector Machine and Random Forest Classifier were used in training and evaluating a machine learning model for detecting spam emails. A spambase dataset was used in training the machine model. The findings indicate that whereas the Random Forest Classifier produced 91.36%, the Support Vector Machine Classifier produced an accurate result of roughly 89.21% [20].

e-ISSN: 2550-1550

III. METHODS

Different kinds of machine learning algorithms are used in this study to categorise spam and non-spam. The phishing dataset is one of several datasets that academics used to assess the performance of their suggested filter available to the public. These datasets are typically used for classifying text. The process for establishing a phishing detection strategy, which serves as the core framework for classifying phishing and non-phishing, is illustrated in Figure 2.

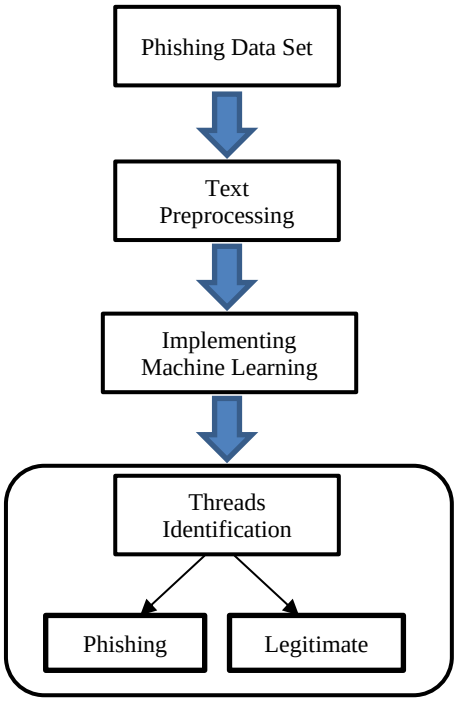

Fig. 2. Phishing Detection Stages

The datasets had a variety of features available. The dataset is evenly distributed and labelled for supervised learning. The labels for the normal and phishing emails are (1) and (0), respectively. The training and testing portions of the dataset are split in an 80:20 ratio. 8843 messages are in the training part, compared to 2211 in the training part. Each training dataset was used to train the models to create a classification classifier, and the testing dataset was used to test the model's output. Accuracy, recall, precision, and F-measure matrices were utilised to evaluate the performance.

To determine whether a classification is accurate, one can look at the numbers for correctly recognisable class examples (TP) as well as the numbers for correctly recognisable instances that do not belong to the class (TN). Cases that were either incorrectly allocated to the class or not detected as class examples are referred to as false positives (FP) and false negatives (FN) [21].

The ratio of the number of correct forecasts to the total number of predictions serves as a measure of accuracy, as shown below:

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

As demonstrated below, Recall is the ratio of the total number of accurately classified positive models to the total number of positive models, and it may be calculated as follows:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FN}}$$

Precision is estimated by dividing the total number of positive examples by the absolute numbers of predicted positive cases to estimate the precision.

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP + FP}}$$

Precision and recall are estimated for the F-measure. The following is the F-measure formula:

$$\boldsymbol{F - measure}\frac{\boldsymbol{2 * Recall * Precision}}{\boldsymbol{Recall + Precision}}$$

The classification of spam is a major issue in the modern electronic environment. Numerous classification techniques for spam and phishing are used. Spam detection algorithms let the mailbox distinguish between spam and non-spam emails. To detect spam, various methods are used. To implement a classification classifier, a model was trained from each training dataset, and the model's output was evaluated on the testing dataset. The classifiers were built using Decision Tree, Naïve Bayes, Neural Networks, Support Vector Machine, Logistic regression and Random Forest, and they were trained in Python.

IV. RESULTS

This section provides an explanation of the study's findings. To assess the effectiveness of the suggested spam comment identification, experiments are run. The features are initially extracted from the dataset to create the featured vector after being carefully chosen based on the fundamental behaviour of spam and ham remarks. Following feature extraction, a variety of classification algorithms are used to obtain performance accuracy. The outcomes of the suggested strategy on various machine learning algorithms are shown in Figure 3, Figure 4, Figure 5 and Figure 6. It illustrates how well the various machine learning techniques perform in terms of accuracy, recall, precision, and F-Measure, respectively.

The performance of the different machine learning techniques in terms of Accuracy, recall, precision and F-Measure are as shown in Figure 3, Figure 4, Figure 5 and Figure 6 respectively. Figure 3 illustrates Neural networks have the highest accuracy (97%). The accuracy rates for Random Forest and Support Vector Machine are 96.8% and 96.4%, respectively. However, the accuracy of the Naive Bayes Classifier is just 60.5%.

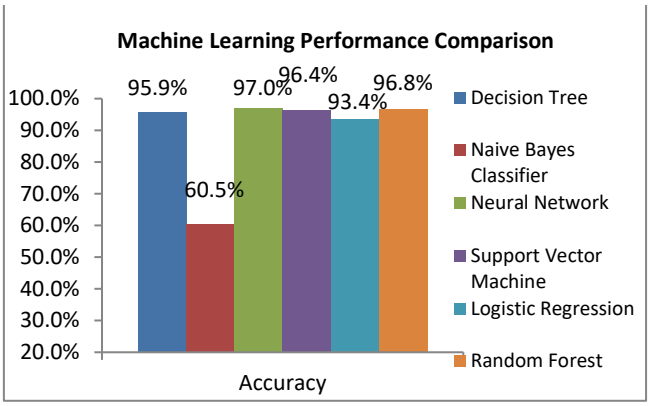

Fig. 3. Accuracy Performance Comparison

Neural networks and support vector machines both display very high recall rates of 98.9% and 98.0%, respectively. Naive Bayes Classifier, on the other hand, only achieves a very poor result of 29.4%. The outcomes are depicted in Figure 4.

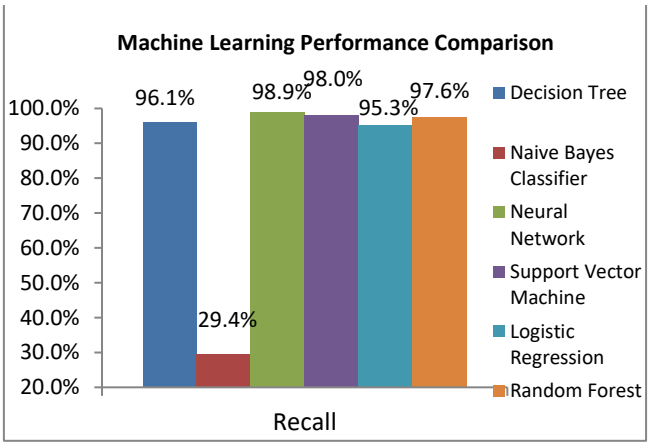

Fig. 4. Recall Performance Comparison

All machine learning techniques produced a precision performance of above 90%, as shown in Figure 5. The outcome indicates that Logistic Regression has the lowest precision (93.0%), while Naive Bayes Classifier has the highest precision (99.5%).

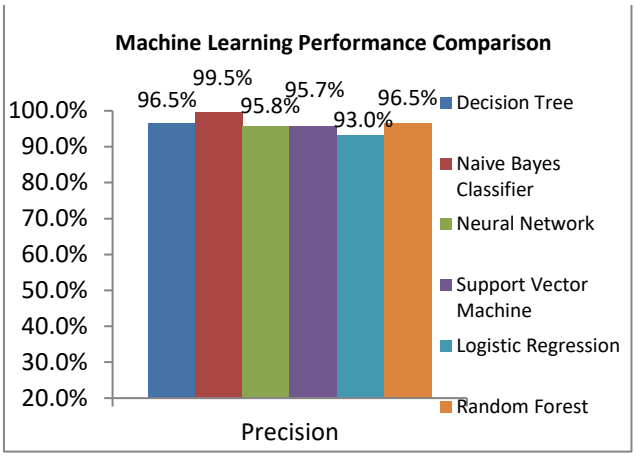

Fig. 5. Precision Performance Comparison

Figure 6 displays the F-measure findings for various machine learning techniques. The results demonstrate that Naive Bayes Classifier only managed to get 45.4% while Neural Network achieved the highest result with 97.4%.

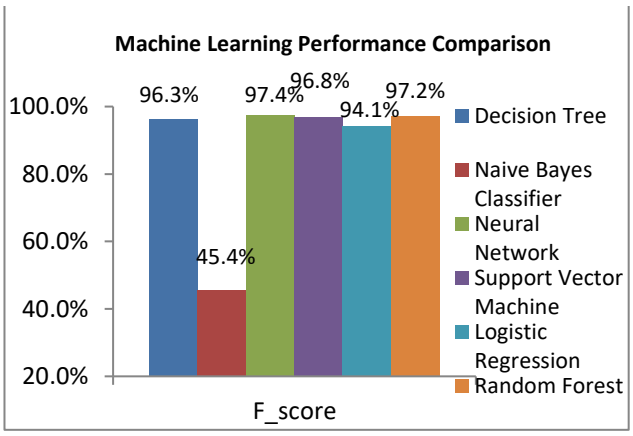

Fig. 6. F_Score Performance Comparison

V. Conclusion

Strong and dependable algorithms that may improve efficiency while also protecting the system against phishing attacks are more crucial as the volume of spam emails rises. A sizable research community became interested in spam identification and filtration. In numerous experiments, search has been utilised to distinguish between phishing and non-phishing messages. There has been a lot of research in this area because it has an expensive and significant impact in many circumstances. To identify and detect phishing emails, many machine learning techniques have been used. The algorithms used in these techniques have good classification accuracy. It may be said that supervised machine learning techniques constitute the basis of the majority of the suggested phishing detection approaches. The supervised model training process depends on a large and time-consuming labelled dataset.

In order to determine which machine learning algorithm performs better, this study compares this strategy to applying other machine learning algorithms to phishing attempts. There has been research done on several machine learning techniques. These algorithms' applications have been carefully documented, and accuracy, recall, and precision performance measures have been assessed. However, machine-learned heuristics in general and neural networks in particular have shown to be incredibly successful and dependable at properly recognising spam and minimising errors to an acceptable level. No single technique, however, can achieve 100% spam detection with zero false positives. In terms of spam detection, neural network algorithms outperform other models. The strategy is more sophisticated, mathematical, and quite possibly much more exact and dependable in carrying out this task.

REFERENCES

[1] Naser. Email Classification Using Artificial Neural Network. International Journal of Engineering, 2, 8-14. 2018.

[2] F. Salahdine and N.Kaabouch, "Social Engineering Attacks: A Survey", School of, Electrical Engineering and Computer Science, University of North Dakota Grand Forks, ND 58202, USA, April 2019.

[3] Abdalla M. Abass, "Social Engineering Threat and Defense: A Literature Survey", Computer Science Journal of Information Security 2018.

[4] A. Alzahrani, "Coronavirus Social Engineering Attacks: Issues and Recommendations", Computer Science International Journal of Advanced Computer Science and Applications, 2020.

[5] P. Bolkas, "Survey on phishing attack and defence techniques", Alexander Technological Educational Institute of Thessaloniki, March 2018.

[6] Jagsir Singh and Jaswinder Singh,"A survey on machine learning-based malware detection in executable files", Department of Computer Science and Engineering, Punjabi University Patiala, India, 2020.

[7] M. Sharma and S.Sharma, "A Survey of Email Spam Filtering Methods", Control Theory and Informatics ISSN 2225-0492 Vol.7, 2018.

[8] H. Bhuiyan, Ashiquzzaman, A., Juthi, T.I., Biswas, S., & Ara, J. A Survey of Existing E-mail Spam Filtering Methods Considering Machine Learning Techniques. Global journal of computer science and technology, 2018.

[9] E.Dada, J. Bassi , H. Chiroma, S.Abdulhamid, A. Adetunmbi and O.Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems", Department of Computer Engineering, University of Maiduguri, Maiduguri, Nigeria, 2019.

[10] J.Singh and Jaswinder Singh,"A survey on machine learning-based malware detection in executable files", Department of Computer Science and Engineering, Punjabi University Patiala, India, 2020.

[11] I. Abdalla M. Abass, "Social Engineering Threat and Defense: A Literature Survey", Computer Science Journal of Information Security 2018.

[12] V. Christina, S. Karpagavalli and G. Suganya, "Email Spam Filtering using Supervised Machine Learning Techniques", International Journal on Computer Science and Engineering Vol. 02, No. 09, 2010.

[13] N. Rusland, N. Wahid, S. Kasim and H.Hafit, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets", Department of Web Technology, Faculty of Computer Science and Information Technology, UTHOM, August 2017.

[14] E. Bahgat S. Rady, W. Gad and I. Moawad, Efficient email classification approach based on semantic methods, Ain Shams Engineering Journal Volume 9, Issue 4, December 2018.

[15] Z. Mohammed, M.Farhaz, M. Irshad, M. Basthikodi and A. Rimaz, "A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm", May 2019, Volume 6, Issue 5.

[16] M. Deepika, S. Rani, "performance of machine learning techniques for email spam filtering", 2018.

[17] D Guo, Z., Tang, L., Guo, T., Yu, K., Alazab, M., & Shalaginov, A. (2021). Deep graph neural network-based spammer detection under the perspective of heterogeneous cyberspace. Future generation computer systems, 117, 205-218

[18] Y. Zamil, S. Ali and M. Naser, "Spam image email filtering using K-NN and SVM", International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 1, ISSN: 2088-8708, February 2019.

[19] S. Rawal, B. Rawal, A. Shaheen and S. Malik, "Phishing Detection in E-mails using Machine Learning", International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 12 – No. 7, October 2017

[20] O.Taylor and P. Ezekiel, "A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier", International Journal of Computer Science and Mathematical Theory E-ISSN 2545-5699 P-ISSN 2695-1924, Vol 6. No. 1 2020.

[21] A. Santra, and C. Christy. "Genetic algorithm and confusion matrix for document clustering." International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012.