The Comparison of Machine Learning Methods for Email Spam Detection

Gwonsik Kang[1], Kamronbek Yusupov[1], Md Rezanur Islam[1], Keunkyoung Kim[1], and Kangbin Yim[2(✉)]

[1] Department of Software Convergence, Soonchunhyang University, Asan, Korea
{johnny,yuskamron,arupreza,kkim}@sch.ac.kr
[2] Department of Information Security Engineering, Soonchunhyang University, Asan, Korea
yim@sch.ac.kr

Abstract. User privacy has become a prominent issue in the digital age, especially with the advent of the Internet and social media. Technologies have opened up new opportunities and different ways for us to communicate. At the same time, they have also brought other avenues and methods for cyberattacks. Email attacks such as the mass sending of malicious messages, links, and phishing dominate among them. Therefore, in our scientific article, we dealt with the most common type of cyberattacks that occur via e-mail. Machine learning methods (ML) have been actively involved in malicious email detection. To find out which algorithm is more effective, we tested different supervised ML algorithms such as Random Forest, Support Vector Machine (SVM), Decision Tree, Naïve Bayes, and K-Nearest Neighbors. And to work with real data, we used some datasets containing emails used in different phishing and bulk emails.

Keywords: Spam Filtering System · E-mail Filtering · Algorithm for ML · Random Forest · Support Vector Machine · Decision Tree · Naïve Bayes · K-Nearest Neighbors

1 Introduction

Email - is a powerful, convenient, and efficient way of communicating over the Internet [1]. In the age of electronic technology, this type of communication is widely used in many fields. From friends and families to work relationships. Despite the fact that new applications and ways of communication appear every day, email is not losing its popularity among users. According to Statista, the quantity of email users achieved 4 billion worldwide in 2020 [2]. This number is increasing year by year and will attain 4.6 billion by 2025. But many users use e-mail only to register with online stores or various websites. After registration, you are likely to receive annoying spam emails. However, in the worst cases, there is a possibility that a malicious link will be sent to your email. Spam is unsolicited messages that are sent in bulk. Annoying emails, link texts, and other messages can be referred to as SPAM. Bulk emails are often sent via email and text messages [3]. HAM - This term was coined in 2001 and is currently

L. Barolli (Ed.): IMIS 2023, LNDECT 177, pp. 86–95, 2023.
https://doi.org/10.1007/978-3-031-35836-4_10

defined as "a regular email that is not spam." HAM is the requested recipient and is one of the legitimate business communication messages. As a rule, such letters are sent by relatives, acquaintances, employers, and the sender of the letter known to the recipient. Business letters are usually written in a professional style. They respect the etiquette of communication and there are no spam emails and malicious data that can harm you.

In times like this, we need a spam filter and blocker that uses various methods to determine whether the message content is malicious or contains spam. With the help of machine learning, we can quickly and more accurately identify the content and compare it with spam messages. By detecting such emails, we can avoid damage and wasting time with unwanted messages.

The structure of sending and receiving letters by e-mail in the following order.

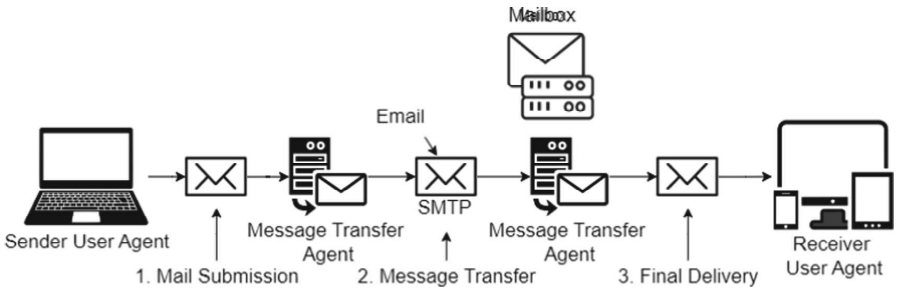


Fig. 1. Simple Mail Transfer Protocol [1]

Figure 1 Shows that this is a simplified common overview of the email architecture and that there may be additional components or options depending on the e-mail system used.

User Agent (UA) [4] is a software application used by an email user to compose, read, and send an email message. Message Transfer Agent (MTA) - The MTA is responsible for transporting email messages between mail servers when the user sends an email. The UA contacts the local MTA, which then sends the message to the recipient's MTA. Mail Delivery Agent (MDA) - The MDA is responsible for delivering email messages to the user's mailbox on the mail server. When the recipient's MTA receives the message, it forwards it to the MDA, which stores it in the appropriate mailbox. Mail Retrieval Agent (MRA) - the MRA is responsible for retrieving email messages from the user's mailbox on the mail server. When a user wants to read the email, the UA contacts the MRA to retrieve the message from the mailbox. Simple Mail Transfer Protocol (SMTP) is the standard protocol for sending e-mail messages between MTAs [4]. It defines the rules and procedures for exchanging e-mail messages. Post Office Protocol (POP) and Internet Message Access Protocol (IMAP) are standard protocols used to retrieve email messages from a user's mailbox on a mail server. POP downloads all messages to the user's device, while IMAP allows the user to view and manage messages on servers without downloading them.

2 Related Work

Spam emails are one of the most common communication problems. Being the most difficult problem in telecommunications, attackers exploiting this problem cause a lot of inconveniences and the number of affected users is increasing every day. Therefore, in our approach, we will study different ways to detect spam messages using ML with different algorithms and examine the efficiency, performance, and execution time of the different algorithms. The reason why we insist on using an automated method to detect and classify spam emails is that automatic classification is a more reliable way to determine the correct categories of spam emails and provides a more accurate result compared to other spam email detection methods. While we were conducting our experiment, we looked at the opinions of other people in this field. Some of them compared several algorithms and shared their conclusions.

Sharma et al. proposed Naive Bayes and J48 for categorization as an approach for spam filtering with high precision and performance [5]. The result showed that their method was effective and more accurate for the detection and classification of spam emails. Their method attained an accuracy of 83.5% Naive Bayes and for exactness by using the J48 algorithm was 91.5%.

Navaney et al. suggested a new procedure for email spam detection. That was using Supervised Machine learning based on Naive Bayes or Support Vector Machine [6]. They observed which algorithms they chose would be more accurate and perform for discerning normal or abnormal emails. The first method based on SVM showed high accuracy with 97.5% and Naive Bayes's result was 95% for finding spam emails.

Yeh et al. propounds a different Spam Classification based on Meta-Heuristics, [7] They chose two different classifiers SVM and Decision Tree to compare and pick one of them to detect Spam. Since SVM takes a long time for training, that is considered its weakness, in spite of this disadvantage it showed good performance. But the Decision Tree based on the classifier gave a more accurate and better result than SVM.

Taylor et al. proposes a method of Detection Spam email using supervised Machine Learning [8]. They also compared SVM and RF classifiers for the Detection system. Their investigation result demonstrated that Random Forest showed more accuracy at 91.36% than SVM at 89.21%.

Saab et al. recommends a new method based on Classification Algorithms for Email filtering by using SVM, Local Mixture SVM, Decision Tree, and ANN. They discussed some approaches for the detection of spam emails [9]. Their result demonstrated that the high accuracy was the SVM algorithm with 93.42%. Then, other algorithms were used LM-SVM, DT, and ANN which achieved accuracies of 91.12%, 91.51%, and 92.96%. The results suggest that SVM performed the best in this research work.

Generally, all these scientific studies show the effectiveness of both different machine learning methods and different approaches in the field of detecting spam with high accuracy. But despite this, each approach is unique and has its strengths and weaknesses, and the choice of algorithms depends on the characteristics of the data set. In our work, we compared the five most widely used algorithms and see their strengths, and weaknesses. See their performance and accuracy in our dataset from Kaggle. And at the end of our work, we will give our conclusion and recommendation which classifier determines spam with high accuracy for machine learning.

3 Machine Learning Algorithms

Unsolicited or meaningless emails that land in the user's inbox without their permission are called spam. They contain advertisements, offers to buy products or services, requests to participate in surveys or contests, links to phishing websites, etc. Spam emails, which are often sent in large quantities using special software or automated tools, can be very annoying for users and harm companies, organizations, and corporations. Spam filters, blocking senders, and creating lists of acceptable and unacceptable senders are just some of the strategies used to combat spam.

3.1 Random Forest (RF)

Random Forest is the most widely used machine learning algorithm invented by Leo Breiman and Adele Cutler in the last century [10]. Figure 2 Shows that this algorithm is considered a universal algorithm since it can be used for many problems. Random Forest is a set of decision trees consisting of an ensemble of decision trees. The algorithm is based on two ideas - the Breiman bagging method and the random subspace method. This algorithm is used for classification, regression, and clustering.

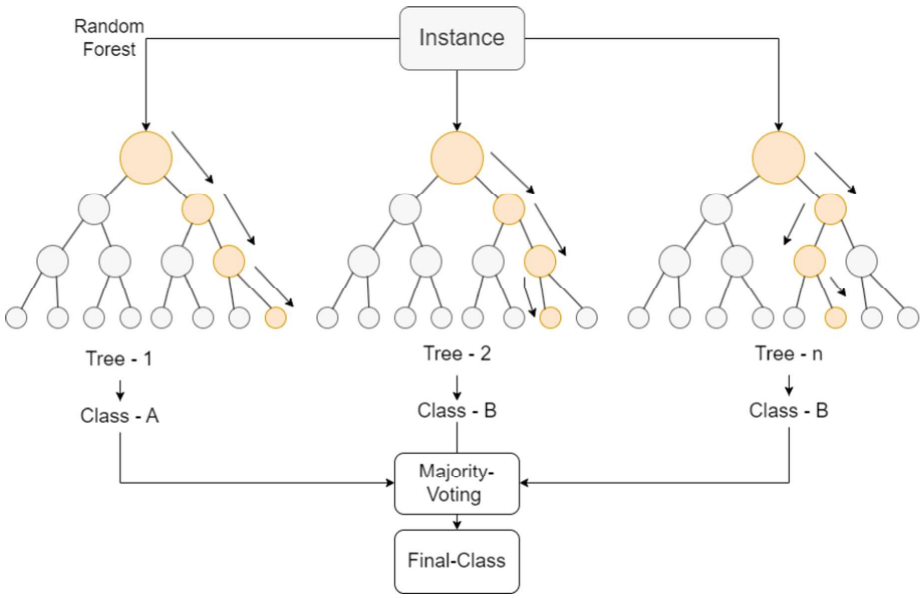


Fig. 2. Structure of Random Forest [11]

3.2 Support Vector Machine (SVM)

A support Vector Machine is an algorithm that learns by labeling objects. SVM is a powerful regression classification and outlier analysis method with a self-explanatory method [12]. Figure 3 shows that SVM learns to distinguish between malicious activities and fraudulent activities on the left map by examining many reports of malicious activities and non-malicious activities. This is the machine learning method that is becoming increasingly popular for analysis. Since SVM classification is simple and flexible, it provides a result for balanced prediction.

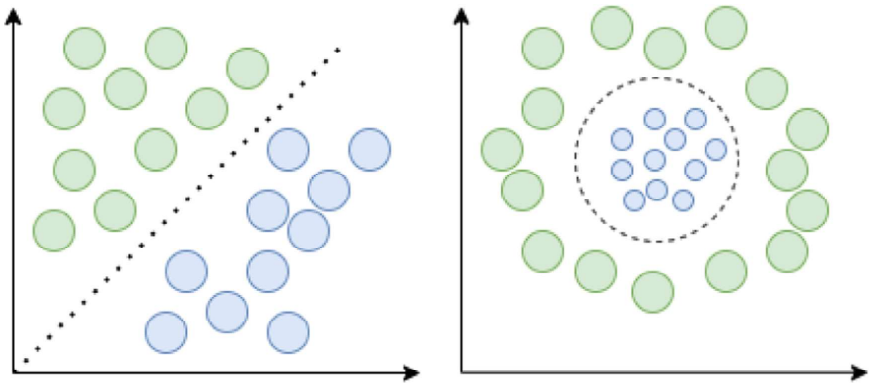


Fig. 3. Support Vector Machine [13]

3.3 Decision Tree

This is a machine learning algorithm used to predict the values of a target variable by constructing a decision tree based on feature values [14]. Figure 4 Shows that each node of the tree provides a test for the feature, the edges correspond to the possible values of the feature, and the leaves (those which do not have a child) of the tree provide predictions for the target variable. The tree is trained using the training dataset, selecting the best feature for the split at each node of the tree. A decision tree has the advantage of being flexible and easy to interpret. However, if the tree is too deep or does not fit the training data appropriately, it can lead to overfitting and poor generalization of new data.

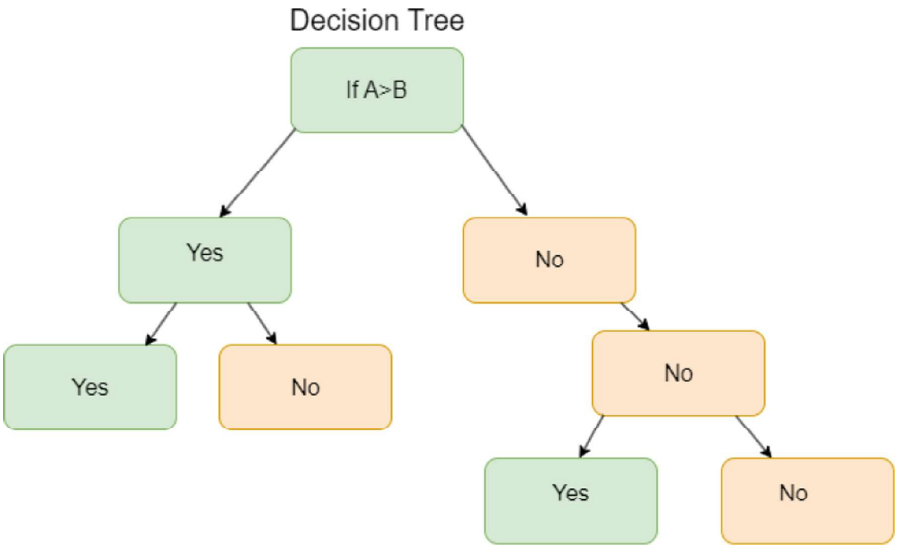


Fig. 4. Decision Tree [15]

3.4 Naïve Bayes (NB)

Naïve Bayes - is used to classifying task instances represented as feature vectors. [16] The classes to which instances may belong are selected from a finite set. There are several ways to train Naïve Bayes classifiers, but they are all based on a common principle: they assume that the value of any feature is independent of the values of other features, given a class label. Figure 5 shows that Naïve Bayes classifiers assume that each feature makes an independent contribution to the probability of belonging to a particular class, regardless of possible relationships between features such as color, shape, and size.

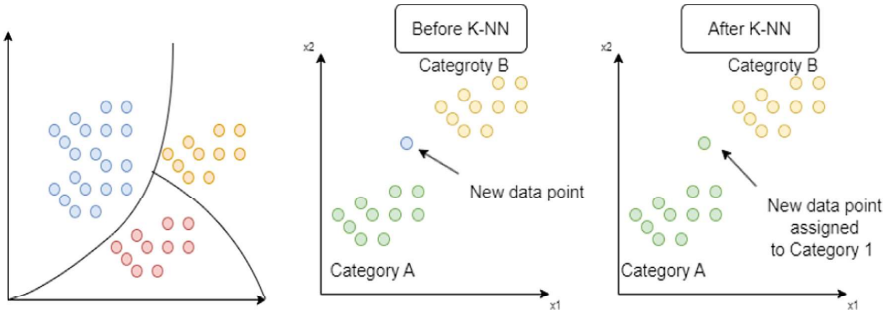


Fig. 5. Naive Bayes [17] and K-Nearest Neighbors [18]

3.5 K-Nearest Neighbors (K-NN)

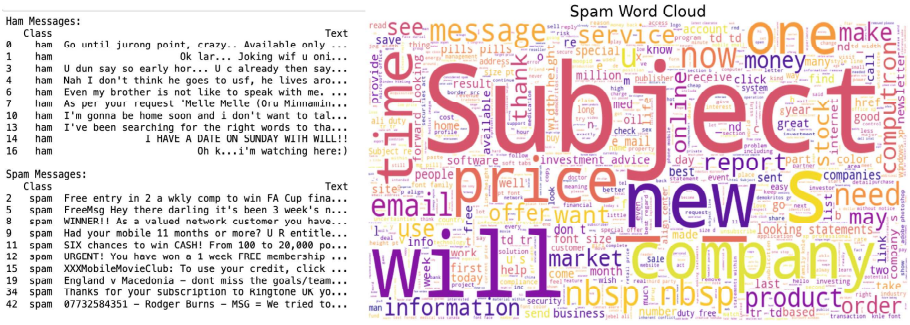
KNN is a machine learning algorithm that uses proximity between data instances to predict class labels or target variable values for new data instances. As a basis for prediction, the algorithm finds the k closest data instances from the training set using the distance metric [19]. In k-NN classification, the class label for a new instance is determined by choosing the most frequent class among the k-closest data instances. In contrast, k-NN regression uses the mean or median of the target variable for the k-closest instances. Despite its simplicity, K-NN can require a lot of computing power to process large amounts of data (Table 1).

Table 1. Advantages and Disadvantages of Algorithms

Algorithms	Strengths	Weaknesses
Support Vector Machine	Effective in high-dimensional spaces Good for small to medium-sized datasets Works well when there is a clear separation boundary between classes	Long training time for large datasets Work poorly when the dataset is noisy Prone to overfitting if not tuned properly
Random Forest	Reduces overfitting and improves accuracy. It is flexible for classification and regression, works well with categorical and continuous data, manages missing values automatically, and does not require normalization of data	Requires a large number of computational resources and time for training, is limited in its interpretability, and cannot determine the significance of individual variables due to the combination of the results of many decision trees
Decision Tree	It does not require normalization or scaling of the data, and missing values are not important for model building. In addition, the solution-decision model is easy for stakeholders to understand and express	Decision trees can be unstable due to slight changes in data. Model training can be expensive and time-consuming. The algorithm is not suitable for regression and prediction of continuous values
K-Nearest Neighbor	Simple and easy to implement Can manage non-linear relationships Can be effective with small datasets	Unsatisfactory performance when working with copious amounts of data and many features. Require feature scaling and account for noise, missing values, and outliers in the data
Naïve Bayes	Easy and fast to train, use and save time Can handle high-dimensional data Suitable for both categorical and numerical data	It is unlikely that his assumption of independence is consistent with the actual data. If the test data contain a feature that was not present in the training data, the naive Bayes procedure assigns a probability of 0 to that feature, resulting in inaccurate predictions

4 Email Dataset

We found two datasets for filtering and training on the Kaggle website [20, 21]. Using these datasets, we were able to assess and determine the accuracy of more than five algorithms. Figure 6 Shows that one dataset contains three columns (Num, Text, 0/1). There are more than 5171 rows, of which 3672 71% are hams. Of all other 1499 29% are considered spam emails.



Ham Messages:

	Class	Text
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
10	ham	I'm gonna be home soon and i don't want to tal...
13	ham	I've been searching for the right words to tha...
14	ham	i HAVE A DATE ON SUNDAY WITH WILL!!
16	ham	Oh k...i'm watching here:)

Spam Messages:

	Class	Text
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...
12	spam	URGENT! You have won a 1 week FREE membership ...
15	spam	XXXMobileMovieClub: To use your credit, click ...
19	spam	England v Macedonia - dont miss the goals/team...
34	spam	Thanks for your subscription to Ringtone UK yo...
42	spam	07732584351 - Rodger Burns - MSG = We tried to...

Fig. 6. Content inside the Dataset

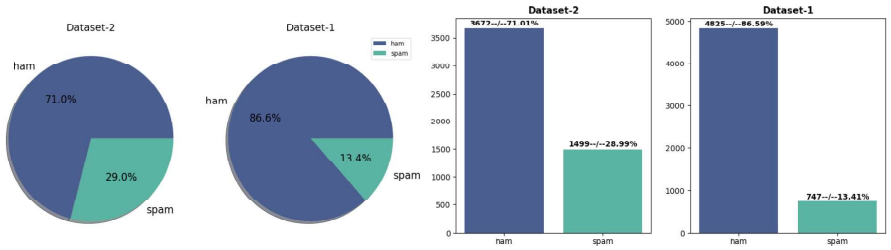


Fig. 7. Data distribution between normal email and spam [20, 21]

Figure 7 Shows that the second dataset, named spam.csv, contains a sizable percentage of normal e-mails, 86.6% - 4825 Ham messages. The spam messages contained 747 emails (13.4%). There are 5572 letters in total, but before we could use them, we had to empty the dataset. We left only two columns (class and text) and removed the rest to get an effective result for our experiment.

5 A Method for Email Spam Detection

In our investigation, we first needed to find a dataset that contains a small set of two categories, normal and abnormal. Since we have already found two data sets, we now need to do some preprocessing of the data, i.e., cleaning and processing the texts and removing words that contain stop words. After preparing a dataset to work with the model, we need to choose a model, i.e. a classification algorithm that is more suitable

and shows a more accurate result. In our case, we took 5 classification models, i.e. algorithms (SVM, DT, k-NN, NB, RF). Since we want to know which algorithm is the best for spam email classification, we need to consider some facts about algorithms, namely the strengths and weaknesses of the algorithms, in which cases we can improve the performance, and in which situations our model may not perform well. For example, Support Vector Machine (SVM) is very good at detecting spam, but when working with large datasets, SVM is not effective and takes a lot of time to train. When the dataset is noisy, performance and accuracy decrease because this is the weak side of SVM. After selecting an algorithm, we need to train the model on our dataset. Since the dataset is not large, training should not take much time. After training, we need to expose our model so that it works as a filter for spam. After deployment, we need a temporary system to test our model so that it can analyze and detect spam emails in real-time based on the dataset we gave it for training. After the tests, it will be clear to us which model gives a more accurate result in which situations. For further investigation, we would like to take the highest and lowest result and add n-gram there to find out how n-gram affects the results. For the lowest result, we want to find out if adding n-gram can improve performance and accuracy.

6 Conclusion

In this study, we wanted to find out which effective supervised machine learning algorithms provide an accurate result in spam detection. These include random forest, support vector machine (SVM), decision tree, Naive Bayes, and K-nearest neighbors. We prepared two different datasets for our tests. In this research, we compared the models and found out which models are more accurate in spam detection. The investigation showed the strengths and weaknesses of the algorithms and in which situations the accuracy of our models decreases. In addition, we could also consider using Deep Learning to detect faster and improve the accuracy of our spam email detection models. Our research concludes that machine learning is considered more effective for spam email detection. However, as spam evolves every day, we need to modify and improve our model to achieve even more accurate spam results.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2021R1A4A2001810) and the Institute for Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2019-0-01343, Regional strategic industry convergence security core talent training business).

References

1. Jameel, N.G.M., Mohammed, E.Z., George, L.E.: An online content based email attachments retrieval system. Kurdistan J. Appl. Res. **2**(1), 68–73 (2017)
2. Statista (2022). published by L. Ceci. https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/
3. Bassiouni, M., Ali, M., El-Dahshan, E.A.: Ham and spam e-mails classification using machine learning techniques. J. Appl. Secur. Res. **13**(3), 315–331 (2018)
4. Riabov, V.V.: SMTP (Simple Mail Transfer Protocol), River College (2005)
5. Sharma, P., Bhardwaj, U.: Machine learning based spam e-mail detection. Int. J. Intell. Eng. Syst. **11**(3), 1–10 (2018)
6. Navaney, P., Dubey, G., Rana, A.: SMS spam filtering using supervised machine learning algorithms. In: 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 43–48. IEEE (2018)
7. Yeh, C.Y., Wu, C.H., Doong, S.H.: Effective spam classification based on meta-heuristics. In: 2005 IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3872–3877. IEEE (2005)
8. Taylor, O.E., Ezekiel, P.S.: A model to detect spam email using support vector classifier and random forest classifier. Int. J. Comput. Sci. Math. Theory **6**(1), 1–11 (2020)
9. Saab, S.A., Mitri, N., Awad, M.: Ham or spam? A comparative study for some content-based classification algorithms for email filtering. In: MELECON 2014–2014 17th IEEE Mediterranean Electrotechnical Conference, pp. 339–343. IEEE (2014)
10. Resende, P.A.A., Drummond, A.C.: A survey of random forest based methods for intrusion detection systems. ACM Comput. Surv. (CSUR) **51**(3), 1–36 (2018)
11. Alghamdi, B., Alharby, F.: An intelligent model for online recruitment fraud detection. J. Inf. Secur. **10**(3), 155–176 (2019)
12. Pradhan, A.: Support vector machine-a survey. Int. J. Emerg. Technol. Adv. Eng. **2**(8), 82–85 (2012)
13. SVM. https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm/
14. Song, Y.Y., Ying, L.U.: Decision tree methods: applications for classification and prediction. Shanghai Archives Psychiatry **27**(2), 130 (2015)
15. DT. https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147
16. Rusland, N.F., et al.: Analysis of Naïve Bayes algorithm for email spam filtering across multiple datasets. In: IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, p. 012091. IOP Publishing (2017)
17. NB. https://dzone.com/articles/what-is-text-classification?fromrel=true
18. KNN. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
19. Sun, J., Du, W., Shi, N.: A survey of kNN algorithm. Inf. Eng. Appl. Comput. **1**(1) (2018)
20. First Dataset. https://www.kaggle.com/datasets/venky73/spam-mails-dataset
21. Second Dataset. https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset