# A Comparitive Study of E-Mail Spam Detection using Various Machine Learning Techniques

Simarjit Kaur, Meenakshi Bansal*, Ashok Kumar Bathla

Computer Science and Engineering, Yadawindra College of Engineering Talwandi Sabo, India

*Corresponding author

doi: https://doi.org/10.21467/proceedings.114.56

**Abstract**

Due to the rise in the use of messaging and mailing services, spam detection tasks are of much greater importance than before. In such a set of communications, efficient classification is a comparatively onerous job. For an addressee or any email that the user does not want to have in his inbox, spam can be defined as redundant or trash email. After pre-processing and feature extraction, various machine learning algorithms were applied to a Spam base dataset from the UCI Machine Learning repository in order to classify incoming emails into two categories: spam and non-spam. The outcomes of various algorithms have been compared. This paper used random forest, naive bayes, support vector machine (SVM), logistic regression, and the k nearest (KNN) machine learning algorithm to successfully classify email spam messages. The main goal of this study is to improve the prediction accuracy of spam email filters.

**Keywords:** Email filtering, Spam email, support vector machine, Random Forest, naïve Bayes, k nearest, logistic regression.

## 1 Introduction

A well-structured and increasingly trendy contact medium is electronic mail. Like any strong medium, but mishandling is given to it. The blind distribution of unwanted email messages, also known as spam, to very large numbers of users, is one such case of exploitation. Spam on the internet is a bleak problem. The emails we get without our permission are spam emails. They are usually sent at a similar time to millions of users [15]. For an Addressee or any email that the user does not want to have in his inbox, spam can be described as redundant or trash email. Email addresses from unusual websites, chat rooms, and viruses [1] are collected by such an individual. Email classification is a crucial area of study for the regular classification of innovative emails from spam emails. For individuals and organizations, spam email is a fascinating issue because it is a level of mishandling [6]. Below are different filtering methods, such as:

(a) Content Based Filtering: This technique usually inspects phrases, the occurrence and distribution of words and expressions in the email body and is used to create regulations to sort spam from the inward email [2].

(b) Case Base Spam Filtering: To begin, all non-spam and spam emails are removed from each user's inbox via community representation. Pre-processing speed is approved later to convert communication using customer edge, feature origin, and assortment, a combination of electronic mail information, and process estimation [3].

c) Rule Based Spam Filtering: A pace beyond basic word-based filters, heuristic filters receive equipment. This form filters receive more expressions originating in a message into deliberation [3] slightly than overcrowding communication that includes a skeptical expression.

(d) Previous Likeness Based Spam Filtering: This approach uses memory-based, ML procedure to record inward e-mail based on its likeness to set aside case as e-mail preparation [3].



Proceedings DOI: 10.21467/proceedings.114; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-947843-8-8

(e) Adaptive spam filtering: Procedures define and filter spam by integrating it into a single module. It divides the body of an electronic message into various classes; each class has a representative text [3].

Spam E-dataset mail's with adequate data processing was included in the research. Various models were then trained and various classifiers were used: KNN, Random Forest, SVM, Naïve Bayes, Logistic Regression.

### 1.1 Email spam filtering process

i. Pre-processing: Pre-processing is a method for unrefined data to be prepared and made suitable for a model of machine learning. It is the initial step in the creation of a model for machine learning. [9].

ii. Tokenization: It is a technique for organising a correspondence that eliminates the use of words. Depending on the message, it divides it into a set of representative symbols known as tokens.

iii. Selection of features: Selection of features is the technique in which you repeatedly or physically pick those features that primarily contribute to the calculation performance you are concerned with. The selection of functionality includes processes such as stemming, noise reduction, and stop word removal steps [3].

### 1.2 Machine learning

The template is used to format your paper and style the text Machine Learning (ML) is essentially the field of computer science that can give good judgment to data in much the same way as human beings with the help of computer systems. ML is a sort of artificial intelligence that uses methods to derive patterns from unrefined data. The aim of ML's description is to allow computer systems to learn from experience without being specifically programmed. Machine learning requires a process that collects consistently dependent data during learning. A hottest algorithm from the event is a most excellent tool for innovation [10]. The trendy applications of AI are machine learning.

### 1.3 Machine learning techniques

Following are the various Machine Learning techniques used in this paper:

- Supervised Machine Learning: In this, the preparation data provided to the machines work as the superintendent that teaches the machines to calculate the output acceptably. A supervised learning algorithm aims to detect a mapping task with the production variable to record the main variable(x) (y).
- Unsupervised Machine Learning: There is no certain object variable to quantify in unsupervised learning. There are no explanations in the dataset for any representation of participation information.
- Reinforcement Learning Machine Learning: Reinforcement Learning is described as a process of machine learning that wonders about how software agents can obtain procedures in a training. Reinforcement learning is a branch of the method of deep learning that allows you to capitalize on many parts of the increasing incentive [12].

## 2 Literature Review

The data on the methods presented is referred to as a literature review. The key goal of the literature review is to figure out how current fields relate to the expected mechanism [1]. Short Message Service (SMS) has developed into a multi-billion-dollar industry that uses deep learning classifiers to compute spam messages, according to this report. Since messaging services have become less expensive, there has been an increase

in the number of repetitive advertising ads (spam) sent to mobile phones at the same time. In this mission, a database of real SMS Spam from the UCI Machine Learning repository is used, and unusual machine learning techniques are used to pre-process and extract features from the database. Finally, the results are compared, and the most effective text message spam filtering algorithm is implemented. A machine learning model was used in this study [2] to determine whether or not an email was spam. A Spambase dataset with 58 columns is used in this representation. The data has been cleaned, and the procedure has confirmed that no void values exist.

Using min max scaler, the concepts of the dataset were well scaled for proper fit in the reproduction instruction using the two machine learning algorithms. The dataset was also at variance with the variables x and y. Where there are 58 columns in the x variable, the y variable contains the output. The variables x and y continued to x-train, x-test, y-train, y-test at odds. Two machine learning algorithms were used to suit this x-train, y-train, Help Vector Classifier and Random Forest Classifier. For precision, those two machines learning algorithms have been weathered. The Support Vector Classifier produced a standard result of about 89.21 percent when kernel = 1, while the Random Forest Classifier produced a true result of about 95.36 percent where estimator quantity = 2. Random forest Classifier had the upper limit of true result, which is 95.36 percent, after going over for accuracy. Then the Random Forest Classifier was saved and used for spam email inspection.

In [3] authors discussed that the rise in the magnitude of unwanted spam emails has created a significant need for the expansion of more consistent and influential anti-spam filters. Present machine learning processes are used to efficiently interpret and filter spam emails. This paper immediately contrasts the advantages and disadvantages of the provided machine learning approaches and the release explores problems in spam filtering with a coherent assessment of some of the normal machine learning-based electronic mail spam filtering strategies and study wrapping overview of the significant concepts, aptitude, and exploration style in spam filtering and analysis. In order to improve the skills of SVM authors in [4] implements a clustering-based SVM approach. Using clustering algorithms, the training data is pre-processed and then the SVM classifier is implemented on the processed dataset. This technique would improve efficiency by resolving the issue of erratic delivery of training knowledge. Compared to that of SVM, the experimental results indicate that the efficiency is improved. An enhanced CLSVM is a SVM classification algorithm that is a mixture of algorithms for clustering and classification is implemented. The new proposed CLSVM method uses the K Means clustering technique to clear the outliers in the dataset compared to the traditional SVM classification technique, and then the dataset is categorized according to the SVM classification algorithm. This improved technique has contributed to a substantial improvement in both the accuracy rate and the system execution time. SMS spam collection dataset is used in this paper and output is 14.43 s with 100 percent accuracy and time required. In [5] authors suggested that, the planned procedure is a well-organized method for categorizing electronic mail spam communication using Support Vector Machines (SVM) and using a Kernel function, an advanced machine learning technique in R, to improve the accuracy of the SVM model, we can provide a nonlinear data SVM usage separation. SVM is a type of superior supervised techniques for machine learning that focuses on information and classifies it into one of the two classes. To enhance the presentation of the categorization and calculation, it is performed using R. In [6] authors explained that the categorization of e-mail spam requires additional consideration to discern the primarily imperative terrorization and reduce the spammers' unnecessary details. Many researchers have continued to categorize spam filtering as the most effective classifier. In comparison with the entire decision tree classifiers and the implementation time, accuracy and small false positive rate were seen only in the RndTree classifier. The accuracy of RndTree is 99% with a standard 98%

Proceedings DOI: 10.21467/proceedings.114
ISBN: 978-81-947843-8-8

Series: AIJR Proceedings
ISSN: 2582-3922

and a false positive of 0.34% compared to the LMT classifier. In addition to additional decision tree classifiers, the RndTree Classifier primarily demonstrates excellent presentation.

In [7] authors prepared a presentation assessment on different methods of categorization as well as: Bayesian Logistic Regression, Secret Naïve Bayes, Radial Base Function (RBF) Network, LMT and J48. In this paper, a presentation evaluation is prepared. The development of the methods was intended to use WEKA data mining apparatus under conditions of precision, correctness, recall, FMeasure, derivation meaning Squared miscalculation. Rotation Forest is the classifier that gives the exceptional accuracy of 94.2 to some outstanding algorithms that work reasonably well on our WEKA training and testing dataset on the Spambase dataset. J48 classification algorithms that record 0.923 precision and 0.885 for Naïve Bayes and 0.932 for Multilayer Perceptron.

In this paper [8] authors presented the use of random forest machine learning technique for specialist email spam correspondence categorization. The main concept is to create a filter for spam electronic mail with improved accuracy of measurement and a reduced number of features. With a consequential categorization accuracy of 99.92 percent, notably a slight false positive rate (0.01) and a very prominent correct constructive speed of 0.999, because the Enron public data set and the random forest method WEKA data mining and machine learning leisure education are running all the examinations. Useful and capable of email spam filtering, Random Forests algorithm and evaluated the performance of RFs algorithm on Enron spam datasets using accuracy, TPR, FPR, accuracy and F-measure to assess the algorithm's usefulness and effectiveness.

In this article [9], an email filtering method is predicted and analyzed using classification techniques. Two ways of expressing characteristics are proposed in advance. In the first, based on web document review techniques, features are extracted from body material. Second, to reduce the dimensionality of these extracted characteristics by only selecting significant terms from a constructed dictionary. Some classifiers have been used to perform experimental studies and use the same dataset compared with existing related work. The reported outcomes check the efficacy of the approach to proposal filtering. The filtering based on the dictionary had an adequate output with faster filtering execution.

In [10] authors analyzed some of the most stylish machine learning algorithms. Metaphors of the methods are given, and the relationship of their presentation on the Spam Assassin spam set is accessible and the study shows a very capable outcome on its own in the methods that are not fashionable in the saleable e-mail filtering association, although the proportion of spam reminiscent in the six methods has the lower rate between accuracy and ethics of accuracy, although hybrid systems seem to be the most accomplished strategy these days to construct an undefeated anti-spam filter.

In this paper [16], recent development in the relevance of machine learning techniques to spam filtering is evaluated in an inclusive manner. In its position, the best part of the consequence of considering the explicit uniqueness of the predicament in view of spam filtering as a customary categorization issue.

In [17] authors recommended successful anti-spam filtering methods based on Artificial Neural Networks (ANN) and Bayesian filters in this paper. For Bayesian categorization, three methods are involved: dual, probabilistic, and probabilistic groups are used. The methods are adaptive and have two equipment. Some strategies for anti-spam filtering in agglutinative languages were suggested in this paper. The roots were used by LM in the second. ANN and the Bayesian filter worked, achieving about 90 percent success. The experiments have shown that more than most Turkish words, certain non-Turkish words that usually occur in spam mails are recovered classifiers.

## 2.1 Research Gaps

After studying the relevant published literature, the subsequent research gaps are identified. This work defines various classification techniques and classifier that have emerged for spam email detection. The analysis shows that different technologies are used. More sophisticated machine learning approaches also are needed here for comparing the accuracy and performance of other machine learning classifiers. More datasets are needed for verify the model. Deep learning approaches also are needed here for comparing the accuracy and performance of other machine learning classifiers. In research work different classifier are used on same dataset and the performance of Naïve bayes, Support vector machine, and random forest, k-nearest neighbors and logistic regression have been compared.

# 3 Machine Learning Methods

Based upon the classification of machine learning techniques into supervised and unsupervised following are the various supervised techniques used in this paper

## 3.1 Support Vector Machine (SVM)

It's a classification and regression system based on supervised learning. It tries to construct an N-dimensional hyper plane that divides the data into two categories. It can also be used to divide a category into more than two categories. SVM is capable of successfully handling high-dimensional feature space. For specific applications of SVM-based classification, choosing the right kernel function is crucial. SVM learning helpful in choosing an appropriate kernel function. [4,5].

## 3.2 Random Forest

It is a familiar pattern of collection learning and degeneration method appropriate for solve the statistics cataloging troubles. Random forests have a number of reward such as: condensed cataloging miscalculation and enhanced f-scores when compared to various additional machine learning methods. It serves as an proficient algorithm for manipulative the conventional value of lost information and maintaining accurateness of the statistics in situation where a huge quantity of the data are lost [8].

## 3.3 Naive Bayes

The Naive Bayes algorithm is a Bayes theorem-based probabilistic process. A Bayesian classifier is used in this process, which calculates a set of probabilities by counting the combinations and frequency of terms in a training dataset. The Bayesian classifier assumes provisional self-determination between the attributes, which greatly decreases the number of parameters that must be calculated. Despite the fact that the categorization is naive, the algorithm performs well and learns rapidly in supervised classification problems [10]. Equation (1) is to compute the spam rating using Probability model as shown :

- Bayes Theorem:
- prob(B given A) = prob(A and B)/prob(A)
- Spam rating computed as:
  - $S(T) = \frac{C_{(spam)}(T)}{C_{(spam)}(T)+C_{(ham)}(T)}$ (1)

Proceedings DOI: 10.21467/proceedings.114
ISBN: 978-81-947843-8-8

Series: AIJR Proceedings
ISSN: 2582-3922

Where:

C_((spam) ) (T)= The no. of spam e-mail carried T,

C_((ham) ) (T)=The no. of ham e-mail carried T.

T= token

There will be no necessitate to combine the diverse token's smarminess to compute the general significance smarminess in classify to work out the possibility for a message M with tokens. Equation (2) is to calculate the result of particular token's smarminess and contrast it with the consequence of precise token's hominess is a simple way to construct categorization. It is representing as below:

$$\left(H[M] = \pi_{I=0}^{N} (1 - S[T_1])\right) \quad (2)$$

The communication is organize as spam if the entire spamminess result S[M] is better than the hominess result H [M] [3].

### 3.4 K-Nearest Neighbors

This classifier is measured an example-based classifier that revenue that the training documents are used for evaluation rather than an release combination illustration, such as the class profiles used by further methods. As such, there is no legitimate instruction segment. while a original article requests to be categorized, the k mainly related documents (neighbors) are initiate and if a great adequate ratio of them have been given to a definite category, the latest document is also give out to this cluster, or else [10].

### 3.5 Logistic Regression Classifier

Ordinary regression and logistic regression are seen to be related. It is used to investigate the relationship between a dependent variable and one or more independent variables, as well as to determine the model's suitability and the importance of the correlations (between the dependent and independent variables). [14].

## 4 Design Methodology

The proposed working model is based on a data mining method for classifying ham and spam emails separately so that content-based spam filters can be more efficient at the user level. The data mining method is divided into four sections: data collection, data pre-processing, data classification, and data evaluation. By detecting different features of spam emails, the efficacy of the proposed model is experimentally tested on simple text datasets of spam emails. Fig. 1 shows a flow chart of the proposed work. The following are some of the measures that are included in the proposed work:

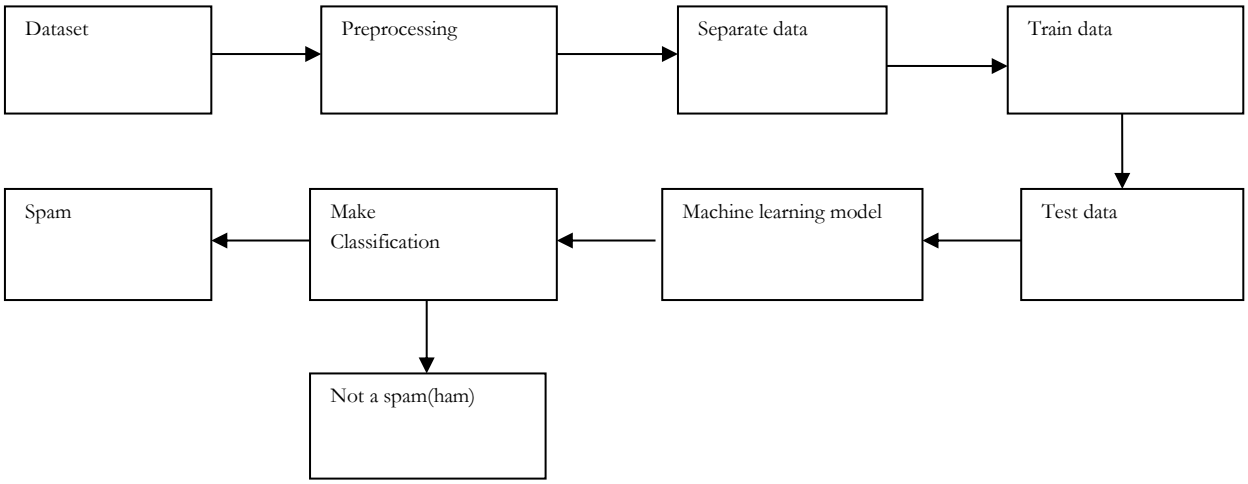


Fig 1. Flow chart of proposed work

- Dataset: Dataset is an assortment of data or interrelated information that is composed for separate elements. A collection of dataset for e-mail spam contains spam and non-spam messages.
- Preprocessing: It is used to transform the raw data in a useful and efficient manner. Preprocessing of emails in next step of training filter. Some words like conjunction words, articles are removed from email body because those are not valuable in classification [9].
- Separate data: It is the act of partitioning presented data into Train and Test data, where train data is used to develop the model and Test data is used to calculate the model 's performance.
- Machine learning model: It can be a mathematical representation of a real world process. The learning algorithm finds pattern in training data.
- Classification: It is a process of categorizing a given set of data into classes. It is identifying to which of a set of categories a new observation belongs, on the basis of training set of data [9].
- Spam and Ham email: Is also referred as rubbish electronic mail, is unwanted mail sent in vastness by electronic message. E-mail that is commonly preferred, is not allowing for as spam.

## 5 Implementation and Results

### 5.1 Description of Dataset

Data related to email spam detection is acquired from UCI machine learning repository dataset. The data was prepared in .CSV compatible format. Different classifier Naïve Bayes, KNN, Logistic regression, Random Forest and SVM are used to analyze the dataset. These classifier's used the Jupyter notebook for coding and Classification with Sklearn. The Python scientific computing library Numpy will be used along with the data analysis library pandas in order to convert these CSV files into pandas. This dataset contains 4601 instances and 58 attributes [13]. Data Set Description as shown in Table I

TABLE 1: DATA SET DESCRIPTION

| **No. of Instances:** | **No. of Attributes:** | ***Data set distinctiveness:*** | ***Attribute distinctiveness:*** | ***Related Task:*** |
|---|---|---|---|---|
| **4601** | **58** | **Multivariate** | **Integers, Real** | ***Classification*** |

### 5.2 Implementation Process and Results

The experimental results of Spam E-Mail Detection using different classifiers are discussed in this section. For classifying algorithms for the experiment, a Spambase dataset is used. The Naïve Bayes algorithm is a simple, easily understood and implemented supervised learning algorithm. Even with limited amounts of training data, the algorithm shows good results. But the algorithm operates with a dataset assumption with independent class characteristics. By testing the output parameters, the efficiency of spam detection is achieved. Parameters like accuracy, recall, precision, F-measure. A performance appraisal focused on the various indicators presented in Table II

Proceedings DOI: 10.21467/proceedings.114
ISBN: 978-81-947843-8-8

Series: AIJR Proceedings
ISSN: 2582-3922

TABLE II: PERFORMANCE EVALUATION MEASURES WITH DESCRIPTION AND FORMULA

| **Evaluation Measure** | **Description** | ***Formula:*** |
|---|---|---|
| Precision | It defines the effectiveness of classifier. | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$ |
| Recall (True Positive Rate) | Out of total class data, the positive labelled data returned by classifier. | $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ |
| Accuracy | Ratio of the positive predicted values to the total data. | $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$ |
| F-Measure | Overall performance by showing effective positive results by classifier. | $2\frac{precision.recall}{precision+recall}$ |

Various types of parameters can be taken which are as follows:

Precision: The ratio of the number of correctly identified instances of a class to the total number of instances classified as belonging to that class.

$$\text{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

Recall: The ratio of correctly categorised instances in a class to the total number of instances in that class.

$$\text{Recall} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

Accuracy : It is the Ratio of the positive predicted values to the total data.

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

F-Measure: It is the Overall performance by showing effective positive results by classifier.

$$\mathrm{F-Measure} = 2\frac{\mathrm{precision.\,recall}}{\mathrm{precision}+\mathrm{recall}}$$

e) True Positive (TP): Correctly detected ham and spam levels, i.e., real ham and spam messages.

f) False Positive (FP): Incorrectly detected ham and spam levels, i.e., not ham and spam messages.

g) False Negative (FN): No. of ham mails incorrectly identified as spam.

h) True Negative(TN): No. of spam mails correctly identified.

### 5.3 Comprehensive Results of SVM, Random Forest

Naïve bayes, Logistic Regression and KNN

Fig. 2. provides a graphical representation of the comparison of the outcomes obtained by individually classifying the respective algorithms

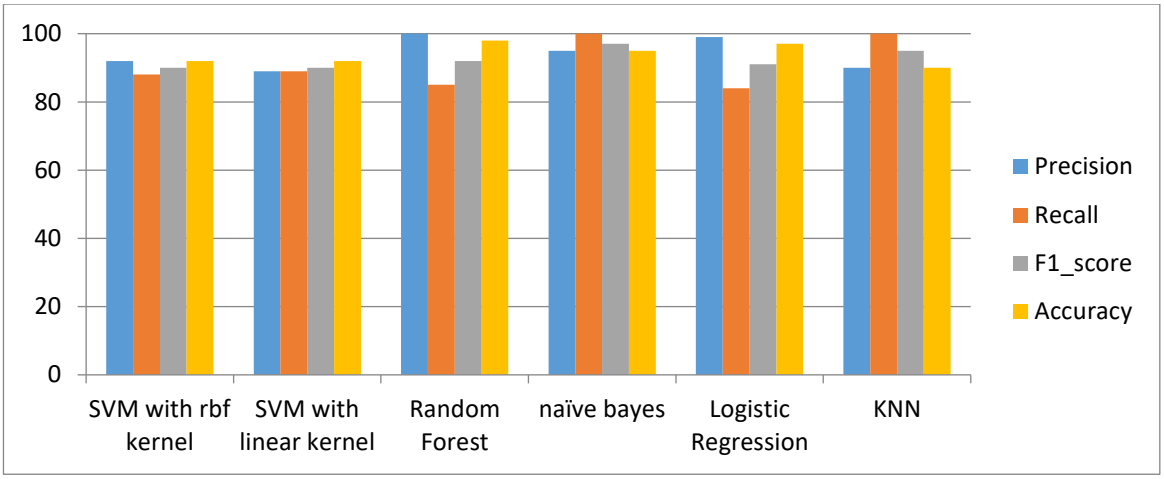


Fig. 2. Comprehensive results of SVM, Random forest, naïve bayes, Logistic Regression and KNN

### 5.4 Experimental Results

The findings of the comparative study as shown in table III. It clearly shows that, in terms of precision, recall, accuracy and F- measure, better results are obtained. However, in the case of Naïve Bayes, the percentage value of F-measure (97%) is higher than other F-measure values of the classifier and in the case of Random Forest, the percentage value of Accuracy (98%) is higher. In existing work [2], the authors have used only two classifiers i.e SVM and Random Forest and in their proposed work they have asked to implement all the other existing ML classifiers. In this paper we have implemented all the other classifiers and instead of SVM with kernel=1 we have used SVM (RBF Kernel) and SVM (Linear Kernel). In existing work [2], the accuracy of Random Forest was 95.36% but in this research the accuracy of Random forest improved to 98%.

TABLE III. EXPERIMENTAL RESULTS

| **Evaluation Measures** | **( SVM with rbf kernel )** | **SVM (linear kernel)** | **Random Forest** | **Naïve Bayes** | **KNN** | **Logistc regression** |
|---|---|---|---|---|---|---|
| Precision (%) | 92 | 91 | 1.00 | 95 | 90 | 99 |
| Recall (%) | 88 | 89 | 85 | 1.00 | 1.00 | 84 |
| Accuracy (%) | 92 | 92 | 98 | 95 | 90 | 97 |
| F-Measure (%) | 90 | 90 | 92 | 97 | 95 | 91 |

## 6 Conclusion and Future Work

This paper introduces machine learning algorithms for detecting spam emails that are used in training and evaluating a machine learning model. Naïve Bayes, support vector machine, k-nearest neighbors, logistic regression and random forest have been identified. It is possible to implement various supervised and unsupervised machine learning algorithms for machine learning. Each technique has benefits and demerits of its own. For the training of the system model, a dataset containing 58 columns was used. Comparing the accuracy and efficiency of other machine learning classifiers, such as Naïve Bayes K-Nearest Neighbor, Logistic Regression, Random Forest, Support vector machine, will further expand this article. It can be further expanded in the future by using Keras andTensor flow in the training of the network

## References

[1] D. Rao and T. Sangeetha, “A Competent Spam Prediction Technique by Supervised Deep Learning Classifiers.” 2020.

[2] O. E. Taylor and P. S. Ezekiel, "A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier," *Int. J. Comput. Sci. Math. Theory*, vol. 6, pp. 1–11, 2020.

[3] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," Heliyon, vol. 5, no. 6, p. e01802, 2019.

[4] D. Pandya, "Spam detection using clustering-based SVM," in Proceedings of the 2019 2nd International Conference on Machine Learning and Machine Intelligence, 2019.

[5] D. Mallampati, K. Chandra Shekar, and K. Ravikanth, "Supervised machine learning classifier for email spam filtering," in Innovations in Computer Science and Engineering, Singapore: Springer Singapore, 2019, pp. 357–363.

[6] C. Balakumar and D. Ganeshkumar, "A data mining approach on various classifiers in email spam filtering," Int. J. Res. Appl. Sci. Eng. Technol, vol. 3, no. 1, pp. 8–14, 2015.

[7] S. Muhammad Abdulhamid, Department of Cyber Security, Federal University of Technology, Minna, Nigeria, M. Shuaib, O. Osho, I. Ismaila, and J. K. Alhassan, "Comparative analysis of classification algorithms for email spam detection," Int. j. comput. netw. inf. secur., vol. 10, no. 1, pp. 60–67, 2018.

[8] E. G. Dada and S. B. Joseph, "Random Forests Machine Learning Technique for Email Spam Filtering." 2018.

[9] B. E. M., R. S., and W. Gad, "An e-mail filtering approach using classification techniques," 2016, pp. 28–30.

[10] W. A. Awad, "Machine learning methods for spam E-mail classification," Int. j. comput. sci. inf. technol., vol. 3, no. 1, pp. 173–184, 2011.

[11] "UCI Machine Learning Repository: Data Set," Uci.edu. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Spambase. [Accessed: 06-Mar-2021].

[12] P. Sharma, Maharshi Dayanand University, U. Bhardwaj, and Maharshi Dayanand University, "Machine Learning based Spam E-Mail Detection," Int. j. intell. eng. syst., vol. 11, no. 3, pp. 1–10, 2018.

[13] M. A. Shafi'I et al., "A review on mobile SMS spam filtering techniques," IEEE Access, vol. 5, pp. 15650–15666, 2017.

[14] K. Kaur, D. M. Kumar, and UIET,Chandigarh, "Spam detection using KNN, back propagation and recurrent neural network," Int. J. Eng. Res. Technol. (Ahmedabad), vol. V4, no. 09, 2015.

[15] Y. Kesharwani and S. Lade, "Spam Mail Filtering Through Data Mining Approach–A Comparative Performance Analysis," International Journal of Engineering, 2013.

[16] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to Spam filtering," Expert Syst. Appl., vol. 36, no. 7, pp. 10206–10222, 2009.

[17] L. Özgür, T. Güngör, and F. Gürgen, "Spam mail detection using artificial neural network and Bayesian filter," in Lecture Notes in Computer Science, Berlin, Heidelberg: Springer Berlin Heidelberg 2004, pp. 505–510.