Comparative Analysis of Detection of Email Spam With the Aid of Machine Learning Approaches

Mangena Venu Madhavan¹, Sagar Pande²*, Pooja Umekar³, Tushar Mahore⁴, Dhiraj Kalyankar⁵

¹²School of Computer Science Engineering, Lovely Professional University, Punjab, India
³⁴⁵Computer Science and Engineering, DRGITR, Amaravati, Maharashtra, India.

¹madhavan.11601828@lpu.in, ²sagarpande30@gmail.com, ³poojaumekar123@gmail.com, ⁴mahoretushar@gmail.com, ⁵dhiraj.kalyankar50@gmail.com

Abstract. Over the past few decades, Technology has gained a rapid pace in its development making communication easier. Considering several modes of communication, E-mails(Electronic mails) are the best means for both informal and formal conversations. Some also use e-mails to store and share important information in the form of text, images, documents, etc. between people using electronic devices. Besides, some people improperly use this means of communication by sending useless or unwanted e-mails in bulk i.e., spammed emails which could result in disproportionate usage of memory in the mailbox. There are many suggested approaches in practice that could identify spam emails from the mailbox using machine learning methods. This paper mainly deals with the comparative analysis of detecting Spam Emails by various machine learning methodologies along with the proposed methodology. Considering various evaluation metrics such as Accuracy, Error, Evaluation time, Efficiency, and so on for the evaluation of models. This document draws the contrast on strengths, drawbacks, and limitations of some of the existing techniques that use the approaches of machine learning to detect spam emails. The machine learning method is further resourceful than the acquaintance approach of engineering which does not involve the specifications of any instructions. Considering various evaluation metrics such as Accuracy, Error, Evaluation time, Efficiency, and so on for the evaluation of models. The various accuracies obtained in this framework are KNN – 96.20%, Naïve Bayes – 99.46%, SVM – 96.90, Rough Sets Classifiers – 97.42%.

1. Introduction
E-mails transfer any form of information between user systems having proper internet connectivity. Unwanted emails in bulk, especially commercial emails affect the storage of the mailbox memory. It would be difficult for the user to delete each unwanted or unused emails manually. To handle this problem, with the increase in the problem of spam e-mails over the years numerous spam detection approaches have been developed. In general, all the e-mail messages are classified as “Ham” and “Spam”. Ham messages are the intended or safe legitimate messages in a mailbox; whereas Spam messages are the junk, unsolicited bulk or commercial messages in the mailbox. This filtering or classification of email messages into Ham and Spam helps in separating them, to delete the spam messages through automation. Usually, there are several parameters or components which help in
identifying spam e-mails. An e-mail could be considered as Spam e-mail when it is associated with Bad grammar, Distorted images, Distorted symbols or logos, Bad links, Tempting offers, and time-based subscriptions that forces the users to subscribe immediately. Phishing is also considered as one of the dangerous cyber-crime which targets the individuals and tricks them to click on links or subscribe to steal the individual’s data like login credentials of social accounts like Twitter, Facebook, or internet banking details in the worst-case scenario. Phishing e-mails are also considered as spam messages. This can also be manually prevented through unsubscribing e-mails, using safe e-mail readers/software like g-mail, yahoo, outlook, etc., installing security software, and keeping them updated all the time. But, it is not very easy to do as sometimes important or useful information might be deleted and would not be possible to recover. Spam e-mails also include Spamadvertised sites - e-mails that advertise products containing URLs that direct to other webpages, 419 Scams – spam e-mails where a small initial payment in a huge sum of money is offered to the users. Image spams – content present in an e-mail is displayed in the form of images. E-mail spam filtering is one of the frequently used processes that help in organizing all the e-mails based on specified criteria. This process comes under automation as it automatically organizes all the e-mails based on prerequisites once they reach the mailbox server. These techniques of approach to spam filters do not follow any set of rules and regulations. To improve it further, it can be trained which helps in learning from previously grouped or classified spam or ham messages. This improvement is termed as Classification which includes the processes of Training and Filtering for a given dataset of e-mails.

Some problems are associated with classification like Noise, Overfitting, Missing Values, Different forms of data. Noise is defined as the interference that occurs with reliability with which features are measured. Shadows, poor lighting conditions, images with blur, typing mistakes, or intended misspellings to hide the spam messages from filters are considered as Noise. Overfitting occurs when there are too many attributes and relatively fewer observations, which identifies trained values perfectly but faces a problem when classifying simple patterns of data, and hence resolving makes the classifier comparatively more complex. Missing values are those in which the dataset does not have information about all the features resulting in zero probability(Naïve Bayes Classifier) making it difficult to differentiate between the classes. Data may not always be in the same form. It may sometimes be the combination of images, text, videos, etc. that cannot be used directly for the classifier. All these problems that are associated with classification should be taken into consideration to define a classifier perfectly. Consumption space of recollection on servers which acquire added cost either to the user, provider or to the company although being of no usage altogether by the inception of Spam, considering a period and necessitating them to the acquisition of additional storage. Furthermore, The extent of this storage compounds exponentially as millions of operators consume the same e-mail client. It is very easy for the user to overlook or fortuitously delete emails which might be appropriate if regular emails are hustled along with spam. The reality of spam distresses an enterprise on all stages as critical communication on each level of an organization is reliant on e-mail. Spam filters can reduce the number of unwanted e-mails to the lowest possible limit. The filtering of emails is the collection of messages in compliance with such requirements to reorganize them. These filters are typically included in handling incoming mails, scanning, tracking, and deleting e-mails containing malicious files like viruses, Trojans, or ransomware. Any specific protocols, like SMTP, affect e-mail operations. Mutt, Elm, Eudora, Microsoft Outlook, Pine, Mozilla Thunderbird, IBM files, Kmail, and Balsa are among the most frequently encountered email server operators. They are web consumers who enable the customers to read and comprehend emails. Spam filtering can be found with both consumers and servers at important positions. Spam filtering is implemented by several ISPs on each network layer, in front of the mail server, or by mail while the firewall is present. The firewall is a network protection framework that controls and administers input and output network traffic based on default safety laws [1]. The email server is a built-in anti-spam and anti-virus device that provides robust email protection on the periphery of the network [2]. Filters can be introduced as external inputs in computers to intermediate between certain terminal machines. These filters can be used in clients[3]. Unwanted or questionable emails are blocked by filtering that compromise network
protection from accessing the operating system. Besides, the user may have a customizable filtering system on the e-mail level which prevents spam emails under certain special circumstances[4]. Various popular platforms exist to communicate between two individuals such as Outlook, Gmail, and Yahoo. These platforms also incorporated various forms of filters to filter the spam mails to provide legitimate emails to their consumers. On the contrary to this situation, these filters might also wrongly block the legitimate mails. It was estimated that approximately 20% of emails dependent on authorization normally failed to arrive at the recipient's mailbox. Email firms have built different frameworks for the utilization of spam filters. The threats posed to email clients by phishing, email-borne threats, and ransomware. The frameworks are used to assess the level of risk for each email received. Instances cover meeting spam restrictions, sender security mechanisms, blacklists and whitelists, and resources to validate receivers. Single or multiple clients may utilize these methods. If the spam content is low, more spam will be prevented and input into the recipient's mailboxes. With a very high threshold, certain big emails may be excluded unless they are redirected by the user.

This document presented in various sections such as section-1 deals with the introduction of the concept, section-2 deals with the related work in the form of a literature review, section-3 deals with the mentioning of the considered methodologies, Section-4 represents the results obtained as well as as the comparative study, and finally, section-5 deals with the conclusion of the document.

2. Literature Review

The World Research Community displays huge curiosity on e-mail spam filtering which gained a rapid upsurge these past days. In this section, the discussion of Similar reviews that are presented within the literature is done. Articulation of problems that are not yet addressed is surveyed to spotlight the conflicts within the review. Usage of e-mails on both the professional and private stages and that they could also be well-thought-out as official documents amongst individuals for communication. Email analysis and data processing are going to be directed for several purposes like subject classification, spam detection, and classification, etc. The revelation made clear that to filter the input file set by unsupervised filtering is utilized to overlook the utmost of prevailing researches. The maximum of prevailing practices that utilize additional features are limited to some substantial features of e-mails and might deliver significant results at most.

E. G Dada et al. in 2019 [5] discuss core principles, attempts, performance, and spam filtering study patterns. The latest study investigates the implementations of machine learning environments to the leading ISPs, including Gmail, Yahoo, and Outlook spam filters, to the spam processing e-mail process. There has been debate about the general approach of spam filtering and the efforts of different researchers to tackle spam using machine learning techniques. The study contrasts the advantages and disadvantages of the existing methodologies of machine learning and brings new problems with spam filter growth. The study suggested broad and strong opposing education as the strategies for managing spam e-mail risks to cope successfully with the potential. S. O. Olatunji in 2017 [6] there was a study to investigate how SVM and ELM contrasted the special and significant E-mail spam identification problem, which is a grading concern. No focus can be put on the significance of e-mail in this current economic scenario. Therefore, it is difficult to reiterate that unwanted mails must be identified and removed quickly and reliably via the spam detection technique. Experimental studies from quite common data sets have shown that both strategies outperformed the best previous studies strategies on the same famous data set used in this analysis. On a scale based on precision, however, SVM performed better than ELM. However, ELM has greatly improved SVM in terms of running speed. S. O. Olatunji in 2019 [7] proposed a model based on support vector machines that are suggested for spam identification when carefully searching for optimized parameters for better results. Experimental findings indicate that all earlier models on the same common dataset used in this work succeeded the model suggested. 95.87 and 94.06% accuracy for preparation is reached and collections of testing, respectively. The 94.06% accuracy of the test reflects a 3.11% increase from the latest studies. S. Muhammad Abdulhamid et al. in 2018 [8] studied the analysis based on the classification of algorithms and their efficiencies. For this study various methodologies considered and their
Efficiencies were measured in terms of basic metrics. Any function collection or efficiency improve approach was used to provide a holistic view of the efficiency of classification techniques. Study shows that there are a variety of classification techniques that are more reliable if better investigated by way of selecting features. Of all the various methodologies utilized, Rotation Forest is the most reliable classifier of 94.2 percent. While no algorithm was 100% specific in handling spam e-mails, Rotation Forest has proved to be among the most reliable product.

A. A. Alurkar et al. in 2017 [9] the prevalence of spam emails is one of the greatest issues facing global communication systems because emails can be reached by anybody with an Internet connexion. Numerous methods for blocking and covering spam involved the automated identification of such phrases and the blacklisting of such spam domains. These techniques do, however, have some weaknesses in the definition of spam or ham communications. This framework aims to use techniques for machine learning to identify a series of repeated keywords known as spam. The method also recommends the grouping of e-mails using a variety of other criteria, including Cc / Bec, domain, and header, in their form. Any parameter is seen as a characteristic. When it is related to the algorithm for machine learning, K. Agarwal and Tarun Kumar in 2018 [10] proposed a combined methodology of machine learning techniques such as the NB algorithm and optimization algorithm namely, the PSO algorithm for identification of spam e-mails. NB algorithm is mainly utilized for classification of the obtained emails into two categories such as spam or non-spam. PSO algorithm is utilized for the optimization parameters that are of the NB algorithm. The implementation of this algorithm was made with the aid of the popular dataset of Ling spam evaluated the efficiency based on the popular metrics. PSO outperforms relative to individual NB approaches based on the validated findings. M. Sahami et al. in 1998 [11] investigated the methodologies that automatically detect the spam emails. For deploying such a framework, the probabilistic methods to classify into spam and non-spam emails from the corpus mails. In a real-world implementation case, it was demonstrated the efficacy of such filters, claim that the technique is developed enough to be used. A. Bergholz et al. in 2014 [12] a variety of new features have been identified that are specifically useful for detecting phishing emails. In this context, we have mathematical models to define email topics with low dimensions, evaluate emails and external connexion sequentially, and identify embedded logos as well as secret salting indicators. The deliberate insertion or manipulation of material that the viewer can not detect is secret Salting. A broad practical corpus of e-mails premarked as spam, phishing, and ham (legitimate) is gathered for methodological assessment. The studies use the techniques to identify phishing e-mails other published approaches. The system addresses the effect of these effects on the process of incorporating this method into an email provider's system. Eventually, it outlines a plan for updating and adapting filters to detect phishing categories.

B. Issac et al. in 2009 [13] it introduces a proposal for Java spam identification technologies and addresses its application with its findings for two separate spam corpuses such as Ling and Enron datasets. The method uses Bayesian formulas for a variety of organized in accordance and keyword collections, together with phrase backgrounds, to enhance the identification of spam and to maintain proper precision. W. Feng et al. in 2017 [14] Suggested a Bayes – SVM-NB – processing framework supporting vector-based computer. The SVM-NB initially creates an ideal hyper-plane dividing sample into two groups during the preparation. For samples in the vicinity of the hyper-plane, one of them is excluded from the training range in multiple categories. This decreases the dependency among samples and simplifies the whole exercise room. The Naive Bayes algorithm is used for classifying e-mail in the test set with the shortened training set. The data set derived from DATAMALL is used to validate the SVM-NB method. The results of experiments show that SVM-NB can achieve better spam detection accuracy and speed. N. Pérez-Díaz et al. in 2012 [15] this paper analyses and unites past methods and innovative methods for applying the rough set (RS) principle to the spam filtering domain by identifying three separate rules execution systems: MFD, the most common decision-making system, LNO, and LTS. To better determine the feasibility of the suggested algorithms, major issues such as corpus size, pre-processing and conceptual concerns as well as various relevant benchmarking steps are explicitly discussed and evaluated for effective model validation. From the
studies that have been performed using a variety of implementation strategies to choose the related decisions produced by rough set sets, the suggested strategies which surpass other well-known anti-spam filtering techniques, such as SVM, Adaboost, and various forms of Bayes classifiers. M. Qi and R. Mousoli in 2010 [16] various methods for detecting spam emails have been used. It discusses the Bayesian algorithms and the SVM two key conceptual techniques. The paper incorporates more recent spam filters. They all use conceptualizing the information to determine if an email was spam or not. S. K. Tuteja and N. Bogiri in 2016 [17] The greatest issue factor was mass mailing or phishing emails in the last generation. In addition to the weariness of such unsolicited spam emails from several email consumers, it also adds a burden on organizations' IT networks and costs companies billions of dollars in missed productivity. The need to filter spam has become more and more critical. this way BPNN filter technique is used to clasp relevant e-mail from unsolicited emails.

It can be summarized the concept of email spam filtering essentiality for the consumers. Popularly, the emails can be classified into spam and non-spam emails. This concept is popularly implemented utilizing machine learning algorithms. But, this scenario is getting very important day by day needs to be updated with improving technologies. The automated framework is required for the filtering of email spams.

3. Methodologies

Recently, the detection of spam Emails typically manages machine-learning (ML) algorithms designed to separate spam from non-spam. This would be done by employing automated and adaptive techniques by machine learning algorithms. Methodologies of the ML framework are more likely to extract information from a collection of emails and to utilize the gathered information to identify new Emails it has just obtained rather than relying on hand-coded guidelines which are vulnerable to the continuously evolving features of spam emails. ML methodologies can best work depend on their practice [18]. In this portion, it will be analyzed some of the most common approaches for the learning of spam. The mentioned figure 1 represents the basic structure of the methodology for classifying the spam and non-spam emails from the Corpus.

![Figure 1. Basic Methodological Structure.](image)

All the messages in an Email are stored in the form of a dataset in the database is known as Corpus. The E-mail message that needs to be classified is initially pre-processed which includes the removal of null values, missing values, and duplicate values. The Data after preprocessing is split into two parts, Training and Testing. In the Training Phase, the algorithm modifies the parameters for the model. The parameters are passed to the model and based on the algorithm and process of the model, it evaluates the given parameters and output is generated. The output obtained from the classification model is then further classified as spam and non-spam. A new testing phase can be added to the model to check
the precision of the model. In this stage, based on predicted output and testing data, and accuracy score is generated to define the perfection of the model and compare it with other models.

3.1. Concepts used in various methodologies

Machine Learning is a field of the technical research of methodologies and mathematical or statistical models that a system utilizes to attain the capability of learning or to achieve a certain task without utilizing an unambiguous set of guidelines, trusting on patterns and interpretation as an alternative. It is a subset of a broad field of AI which allows machines and computers, act or perform certain activities as a human does. Machine Learning comes into application in many scenarios like Spam Detection, Speech and Image Recognition Systems, Medical Diagnosis, Prediction Systems, etc.. It helps in reducing human effort, hence making the tasks easy to be performed with the help of a machine. There are a lot of algorithms that could be used in e-mail filtering, which are broadly studied by the approach of Machine Learning. This includes the K-Nearest Neighbor (KNN) algorithm, Naïve-Bayes (NB) Algorithm, Support Vector Machines (SVM) Algorithm, and Rough Sets Classifiers. The broad division of Machine Learning is made into three major categories, depending on the nature of learning. They have Supervised Machine Learning, Unsupervised Machine Learning, Reinforcement Learning. Supervised Learning provides the system with certain inputs and corresponding outputs where a general rule is generated that maps input to its corresponding output(example: Spam detection, fraud detection, image recognition). Unsupervised Learning is where outputs are not defined, allowing the system to find a pattern from the given input(for example grouping fruits based on size, shape, or color). Whereas in Reinforcement Learning, A computer program interacts with an environment to reach a certain goal and it does not have any prior knowledge about the target(example: robotic systems, learning to drive a vehicle). Machine Learning includes a lot of pre-processing required for an algorithm to work more efficiently. Initially, Data(any unprocessed text, value, fact, sound, or a picture) is converted to Information(interpreted and manipulated data) and further made useful by providing it in the form of Knowledge(further inferred resulting in concept building). Data is split to perform several actions like Training, Testing, and Validation. Processing of Data is done through the steps of Collecting, Preparing, Input, Processing, Output, Storage. Data Processing, Data Cleaning takes place includes Exclusion of observations that are not required, Fixing Structural errors, Managing Unwanted outliers, Handling missing data. As Supervised Machine Learning models are used by us for e-mail spam detection, Classification is majorly used for spam detection, as the name implies, grouping or classifying a similar object based on the training dataset obtained. Classification can further be divided into two sub-categories i.e. Binary Classification – Categorizing data into two distinct classes, Multiple Classification - Categorizing data into multiple(more than 2) subclasses. Some of Supervised Machine Learning Techniques that are frequently used for e-mail spam detection are:

- K-Nearest Neighbour (KNN) Algorithm
- Naïve-Bayes (NB) Algorithm
- Support Vector Machine (SVM) Algorithm
- Rough Set Classifiers

3.1.1. K-Nearest neighbor (KNN) algorithm. The K-nearest neighbor (KNN) classifier in which usage of the training documents for comparison as an alternative of a particular category representation hence called an instance-based classifier taken into account, like the category profiles employed by other classifiers. There is no real process of training in KNN. The k most related documents and neighbors are identified where a substitution document has to be categorized and an outsized proportion of them is allocated to a certain category and the current documents to the current category are still classified, otherwise not. In comparison, the neighbors are also fixed using conventional indexing techniques. It looks at the following group of communications to determine if an email is a spam or a ham. A comparison between the vectors is always conceived in the nearest neighbor's algorithm as a real-time process. The assumption of this methodology deals with instances with
similar properties that exist close to each other in the provided dataset. During the training phase, split the training dataset and store it. For a given Email, determine k nearest neighbors for each attribute within the training dataset. Classify the spam messages among neighbors as spam, else classify them as ham. K-Nearest Neighbour being an example-based classifier consumes less computational time in training and more computational time in testing.

- **Algorithm:**
  
  1. **Step 1:** Load the Training Data
  2. **Step 2:** For each test instance, evaluate the Distance Metric (distance from each training instance used) by calculating the Euclidean Distance as mentioned in the following equation-1.
     
     \[ D(x, y) = \left( \sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2} \]  
     
  3. **Step 3:** Find the k-neighbors with the nearest (minimum) distance
  4. **Step 4:** Consider the label which has major votes among the given dataset labels to decide the label of a test instance.

The advantages of the KNN algorithm are the output obtained is of high accuracy for small datasets, and takes all the features present in the dataset into consideration. The disadvantages of the KNN algorithm are the Computes all the training instances per test instance during classification, resulting in high time complexity during the testing phase, further increasing the computational cost, and require a large amount of memory.

### 3.1.2. Naïve-Bayes (NB) Algorithm

The Naïve-Bayes (NB) algorithm is a machine learning methodology which was a statistical model that usually has strong independence properties, probability distribution, and skill to tackle huge datasets. In the NB algorithm, from the distribution of dataset probability distribution is evaluated. Bayes's decision rule is employed to designate a category in classification problems. Classes having the highest value of posterior probability are chosen by the classifier as defined by the Bayes decision rule. The posterior probabilities are often evaluated with the following mentioned in equation-2. Based on Bayes Theorem for Conditional Probability, the probability that a given set of features \(x_1, x_2, \ldots, x_n\) are enclosed in a vector \(L\) belonging to a category or a class \(M\) is given by the following equation-3.

\[
P(M|L) = \frac{P(M)P(L|M)}{P(L)} (2)
\]

\[
P(S|L) = \frac{P(S)P(L|S)}{P(S)P(L|S) + P(T)P(L|T)} (3)
\]

The assumptions of this algorithm are the values of a specific feature is independent of all the other features given in that class. During the training phase, parse each Email into its respective tokens, then a probability is generated for each token, and values of spam probability are stored. The filtering process deals in the categorization of each Email into spam and ham considering a threshold value to define spam content. The filtering technique followed popularly known as the Gaussian NB Filtering. The assumptions of this method are the continuous values to be considered which follow Gaussian Distribution. The training phase mainly deals with segmentation of the provided data by category, by computing the mean and the variance of all the values present in each class. Filtering deals with the instance categorization of the category represented by ‘M’ depending on the probability for each test instance with attribute value \(v\) equal to ‘\(l\)’. The following mentioned equation – 4 represents the gaussian NB filter.

\[
P\left(\frac{X = V}{C}\right) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} e^{-\frac{(v - \mu)^2}{2\sigma^2}} (4)
\]

The Advantages of this method are training speed is very fast that will help in the computation of the mean and variance of the training data, this approach based on statistical modeling and it is very easy to implement. The disadvantage of this method is not able to hold well when data is correlated or
the assumption of data independence fails and is affected by zero probabilities (occur when the product of individual probabilities = 0; due to missing values).

3.1.3. Support vector machine (SVM) algorithm. This algorithm is grounded on the notion of Structure Minimization of Risk which intends at identifying the hyper-plane which divides the mentioned two categories perfectly. Points lying on the hyper-plane are known as support vectors that are utilized in the decision-making function. The concept of decision planes that outline decision boundaries supports Support Vector Machines. A group of objects having non-identical class memberships is separated by a choice plane, and The SVM modeling algorithm determines an ideal hyperplane with the maximum margin of separation for two groups, which involves simplifying the subsequent optimization problem. Cross-Validation is a typical process that is conducted on the training dataset. Cross-validation also involved assessing the potential for generalization of new samples which are not included in the training data set. Cross-validation partitions the training data set arbitrarily into K subsets which are of almost equal, those partitions referred to as K-fold, in which one subset is left out, and a classifier is built on the samples remaining, then the efficiency of classification on the unused subset is measured. This procedure is recurred k times for every subset to obtain the cross-validation performance over the whole training dataset. A little subset is often used to minimize computing costs for cross-validation If the training dataset is large, the subsequent algorithm is often utilized in the classification process. During the training, From all the samples of the training set that require classification, find k nearest neighbors for them. Obtain the decision points and train the SVM model. During the filtering, all attribute points are classified from the obtained model on either side of the hyperplane and output the results.

The advantages of the discussing algorithm are: it is highly influential for high-dimensional spaces, and it is very efficient in managing the memory as its decision function utilizes the subset of training points. The disadvantages of this model are: if the number of features is relatively greater than the number of samples, it might not be efficient, and direct probability values are not available, hence cross-validation is required.

3.1.4. Rough set classifier. Rough Set Classifiers are very capable of computing the reduction of data systems. Attributes that are unrelated to the empirical definition (i.e. judgment attributes), and may have multiple redundant attributes in the data model. Reduction, a minimum subset of conditions attributes that correspond to decision attributes, is sufficient to attain basic usable knowledge of this method. The following mentioned way that the discussing algorithms work:

- Firstly, it will attempt on the incoming emails is picking the foremost appropriate attributes to be further utilized for classification. The input of the data collection is then processed into a system that separates further into datasets for training and research. The training data set generates a classifier to be used for the successful evaluation of the test data set. Step 2 and Step 3 are followed for the preparation of the dataset.
- Boolean reasoning needs to finish the discretization strategies as the decision system has real value attributes.
- For obtaining the decision rules, genetic algorithms should be utilized. Proceed with step-4 for the testing dataset.
- For employing equivalent cuts that are computed from step-2, discretize the testing dataset. Make sure, each new object in the testing dataset needs to match with the principles generated in step-3.

4. Results and Discussion
Machine Learning algorithms play a crucial role when it comes to spam classification. Four major machine learning models that are used in spam classification are discussed in this paper. E-mail messages consist of numerous parts: header, body, etc. The header contains the fields in the mail like ‘From’, ‘Subject’. The subject consists of most of the information which is generally used to classify
as spam or ham; whereas From is used to knowing about the sender and to mark the sender as spam if required so that all the e-mail messages from that sender can be directed to the spam folder without any further classification process required. The body is the main part of the e-mail message which defines the structure of the message for proceeding with steps of preprocessing. Several features in the Body are selected to define or categorize the words as spam which further defines the message as a spam message. While consideration of methods is done, they are chosen based on the features selected or how the message should be classified. Every classification algorithm has its advantages and disadvantages when parameters like computational time, computational cost, memory allocated, etc. We consider three parameters to define the performance of an algorithm,

- **Accuracy**: The e-mails that are properly classified and categorized per all e-mails considered based on the accuracy score. It defines how accurately the algorithm works.
- **Spam Recall**: The spam e-mails that are properly classified and categorized as spam per all spam e-mails considered is Spam Recall.
- **Spam Precision**: The Spam Precision defines the percentage of related spam e-mails identified among all the e-mails. Shows how many e-mails classified and categorized as spam are spam.

The results obtained in terms of the above-mentioned evaluation metrics obtained as mentioned in Table-1. The visual comparison of these metrics across the various Machine Learning algorithms implemented was represented as mentioned in Figure -2. From both of these representations, one can identify that the Naïve Bayes algorithms working much efficiently than any other Machine Learning algorithms. In the case of the Naïve Bayes algorithm, it is not only maintained the accuracy but also Spam Recall and Spam Precision which indicates better efficiency of the model.

| Algorithm Used         | Accuracy (%) | Spam Recall (%) | Spam Precision (%) |
|------------------------|--------------|-----------------|--------------------|
| K-Nearest Neighbor     | 96.20        | 97.14           | 87.00              |
| Naïve Bayes            | 99.46        | 95.00           | 99.66              |
| Support Vector Machine | 96.90        | 95.00           | 93.12              |
| Rough Set Classifier   | 97.42        | 92.26           | 98.70              |

**Table 1.** A slightly more complex table with a narrow caption.

![Comparison of Evaluation Metrics across the Machine Learning Techniques](image)

**Figure 2:** Comparison of Evaluation Metrics Across the Machine Learning Techniques.
5. Conclusion

Through this study in the paper, we learned about detecting the spam messages in e-mails through different approaches of classification algorithms by machine learning. This review justifies the working and functionality of the algorithms along with their advantages and disadvantages based on numerous considered parameters. To solve the problem of spam e-mails through machine learning classifiers, several attempts have been made by many researchers. This also became leverage in producing new loopholes for spam e-mail generation. Detection of spam e-mail messages has evolved from filtering to classification. Besides, there are numerous amount of algorithms from which some of the major algorithms are looked into. This paper presents the issues based on several challenges based on spam filtering and classification when a particular algorithm is considered in specific. Major studies and researches that are developed based on several challenges have been discussed. Some of the open research problems also include the usage of these algorithms that have been thoroughly identified and performance metrics for an algorithm is evaluated form accuracy, spam recall, and spam precision. In brief, this paper discusses how a spam detection and processes of filtering and classification works, current trends of spam, how the approach of machine learning field helps in the spam detection process, how a general machine learning classification algorithm works, how a specific algorithm classifies the e-mails into substituent spam and ham messages, the parameters in which a particular algorithm is efficient and in which it isn’t. Through this document, the selection of a particular algorithm can be made based on the features considered in detecting a spam e-mail. Also helps develop hybrid algorithms through a combination of algorithms as their peer review is made. As observed from all the models of classification in the field of machine learning, every method that is considered has its pros and cons. So, for an efficient algorithm to be developed that performs at best even when any parameters like evaluation time, acquaintance cost, the memory of allocation, etc. Therefore, Hybrid Algorithms seems to be the best and feasible solution for Spam detection in e-mails.

References
[1] Katakis, Tsoumakas G, Vlahivas I, Email mining: emerging techniques for email management 2007, Web Data Manag. Pract.: Emerg. Tech. and Tech., Idea Group Publishing, chapter 10.
[2] Teli S and Biradar S, Effective spam detection method for email 2014, International Conference Advanced Engineering Technologies, 2014, 68–72.
[3] Irwin B, Friedman B, Spam Construction Trends 2008, Proc. of the ISSA 2008 Innov. Minds Conf., 1–12.
[4] Christina V, Karpagavalli S, Suganya G, Email spam filtering using supervised machine learning techniques 2010, Intem. J. Comp. Sci. Eng., 02, 3126 – 29.
[5] Dada E G, Bassi J S, Chiroma H, Abdulhamid S M, Adetunmbi A O, and Ajibuwa O E, Machine learning for email spam filtering: review, approaches, and open research problems 2019, Heliyon, 5.
[6] Olatunji S O, Extreme Learning Machines and Support Vector Machines models for email spam detection 2017, Canad. Conf. on Elect. and Comp. Eng., 1 - 6.
[7] Olatunji S O, Improved email spam detection model based on support vector machines 2019, Neu. Comp. and App., 31, 691–99.
[8] Muhammad Abdulhamid S, Shuaib M, Osho O, Ismaila I, and Alhassan J K, Comparative Analysis of Classification Algorithms for Email Spam Detection 2018, Inter. J. Comp. Net. Inf. Sec., vol. 10, 60–67.
[9] Alurkar A A, Ranade S B, Joshi S V, Ranade S S, Sonewar P A, Mahalle P N, and Deshpande A V, A proposed data science approach for email spam classification using machine learning techniques 2017, 2017 Inter. of Things Bus. Mod., User., and Net., 2018, 1–5.
[10] Agarwal K and Tarun Kumar, Approach of Naïve Bayes and Particle Swarm Optimization 2018, 2018 Sec. Int. Conf. on Int. Comp. and Cont. Sys., pp. 685–90.
[11] Heckerman, David, and Horvitz, Eric and Sahami, Mehran and Dumais, Susan, A Bayesian Approach to Filtering Junk E-Mail 1998, AAAI Workshop on Learn. for Text Categ.
[12] Bergholz A, De Beer J, Glahn S, Moens M F, Paab G, and Strobel S, New filtering approaches for a phishing email 2010, J. Comp. Sec., 18, 7–35.
[13] Issac B, Jap W U, and Sutanto J H, Improved Bayesian anti-spam filter - Implementation and analysis on independent spam Corpus 2009, 2009 Inter. Conf. on Comp. Eng. and Tech., 2, 326–30.
[14] Feng W, Sun J, Zhang L, Cao C, and Yang Q, A support vector machine-based naive Bayes algorithm for spam 8.
[15] Pérez-Díaz N, Ruano-Ordás D, Méndez J R, Gálvez J F, and Fdez-Riverola F, Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification 2012, App. Soft Comp. J., 12, 3671–82.
[16] Qi M and Mousoli R, Semantic analysis for spam filtering 2010, 2010 Seventh Inter. Conf. on Fuzzy Syst. and Knowl. Disc., 6, 2914–17.
[17] Tuteja S K and Bogiri N, Email Spam filtering using BPNN classification algorithm 2017, 2016 Inter. Conf. on Aut. Cont. and Dyn. Opt. Tech., 915–19.
[18] Mitchell T M, Machine Learning 1997, first ed., McGraw-Hill, 1997.