INTRODUCTION

Email is the short form of electronic mail and it is defined as the exchange of information through communication channel. Typically, emails come from different email addresses rather than being entered from the key board or electronic files stored on the disk. Most mainframes, minicomputers, and the emailing system are applied on the computer network. The term electronic mail can also be written as Email or e-mail. Email address is required to send and receive email messages. The majority of internet service providers provide a free email account to customers. Email has been tested to be one of the Internet's preferred services; it is used for international communications. But, it is criticized for its insecurity, spam, as well as viruses and malware being unfold through email attachments. E-mail offers a way for web users to simply transfer information globally. E-mail presents a super way to send millions of commercials free of charge for the sender, but the bad thing is that these days’ emails are appreciably exploited. Generally, receiving an e-mail from an unknown supply comprises contents that are of no importance to the user. As a result, due to these e -mails, many people are getting cluttered with all unsolicited bulk e-mails also referred to as “spam” or “unsolicited mails” (Vinod et al., 2013). Spam often causes unwanted information or bulk information to get transmitted to email accounts. Spam mail could be a set of electronic spam involving nearly identical messages sent to numerous recipients. Spam emails conjointly embrace malware as scripts or alternative executable file attachments. Spam is waste of time, storage space and communication bandwidth. If spam continues to increase, it will be unmanageable in the near future to handle such huge spam.

Automatic e-mail filtering looks like the foremost effective methodology to counter spam at the moment and has a good competition between spammers and spam-filtering ways. In the past, most of the spams were treated by the interference of e-mails coming from sure addresses or filtering out of messages with sure subject lines. Spammers began to use many difficult ways to beat the filtering ways like victimization of random senders’ addresses and/or appending of random characters to the start or the tip of the message subject line. Spam emails vicinity unit is used to spread a virus or malicious code for fraud in banking, publishing, advertising and much more (https://pdfs.semanticscholar.org/c2ea/4bf0282b9b39a6ba773581332bb0587ec4ab.pdf). (Nilam et al.,2017) So to avoid this kind of bulk email it is essential to use spam filtering technique which is a machine learning algorithm. In this study, we cover the performance of three widely used supervised machine learning method for data classification and identify the best classifier algorithm. Those supervised machine learning algorithms are K-nearest neighbor (KNN), Support vector machines (SVM), and Naive Bayesian (NB). Supervised learning is one of the methods associated with machine learning which involves allocating labeled data so that a certain pattern or function can be deduced from that data. It is worth noting that supervised learning involves allocating an input object, a vector, while at the same time anticipating the most desired output value, which is mostly referred to as the supervisory signal. The bottom line property of supervised learning is that the input data are known and labeled appropriately. In a study by Binh et al. (2018), four Bayesian machine learning algorithms (NB, NBT, BN and DTNB) were selected and compared with one of the benchmark landslide models of the SVM for landslide susceptibility assessment at Pauri Garhwal district, Uttarakhand State, India. Results show that the SVM model was highly reliable followed by the NBT, DTNB, BN and NB. This is in accordance with the results of statistical index based methods and the ROC curve.

In this work, supervised machine learning is used rather than unsupervised machine learning because it is worth noting that both methods of machine learning require data to analyze to produce certain functions or data groups. However, the input data used in supervised learning are well known and labeled. This means that the machine is only tasked with the role of determining the hidden patterns from already labeled data. However, the data used in unsupervised learning are not known nor labeled. It is the work of the machine to categorize and label the raw data before determining the hidden patterns and functions of the input data. It does not only input data accurately but also gives accurate and reliable results. To review the performance outcome of the three machine learning strategies six terms were used: true positive, false positive, recall, precision, f measure and accuracy. These are called imagining the algorithms. The entire machine learning algorithms give different results on the same dataset. This paper focuses on effective and efficient email classification techniques based on data filtering method used for the training model and accuracy of the algorithm before and after filtering the classification method. It also compares the accuracy of algorithms before and after boosting.