| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

Spam Email Classification Using SVM: A Machine Learning Algorithm

Shubham Marathe

Academic Researcher, Department of Information Technology, B.K. Birla College of Arts, Science and Commerce (Autonomous), Kalyan, Maharashtra, India

ABSTRACT: Nowadays Email is used as an important form of the digital communication system. Lots of emails are exchanged every day on the internet. According to Statista in 2019 293.6 billion emails are exchanged worldwide. From them, 55% are spam mails. The main reason for mails in large amounts is due to their interoperability and ease to use. Spam emails contain malicious links, files and when the user clicks on these links then the victim can get access to sensitive information of the user. The victim can impersonate himself as a trustworthy entity to gain information from the user. For that, it is important to identify emails and their attachments. Machine learning involves training a machine with some algorithms. It will be beneficial as the machine can recognize whether there is harmful data present in email or not.

KEYWORDS: Machine learning, Spam email, Python, SVM, Natural Language Processing

I. INTRODUCTION

In the past few years, spam emails are increased in a tremendous amount which becomes big trouble for the internet. Intruder sends spam mail to genuine users to get sensitive information about the user, his personal information and tries to violate his information. Recently many peoples use email as a form of communication. To protect the user from this threat emails should classify as spam (Malicious) or ham (Good) mail. For that proper classification of mails is required. In the past few years, the application of machine learning in different fields is increased because of the capability of handling a large amount of data and the availability of necessary tools. Machine learning is an application of artificial intelligence that helps systems to learn automatically from experiences and use it without specific programming. Machine learning extract features from the data to generate the model, hence helping computers to make educated guesses about unseen data with a significant amount of accuracy. Using machine learning we can predict the outcome of an application or software before explicitly programmed. Hence the machine learning approach is very useful in email classification. Machine learning contains different types of methods but all of them are categorized into two major parts as supervised learning and non-supervised learning. Support vector machine (SVM) is a supervised learning algorithm that analyzes data using classification and regression analysis. In this paper, a model is proposed where a datasheet is taken as a source to get data of mails which contained both ham and spam. After that, this dataset is processed using the SVM algorithm. The classification accuracy of SVM is tested using different parameters to check how much accuracy can be achieved using it.

Support Vector Machine is a supervised machine learning algorithm which is used for classification and regression. Mostly it is used for classification. SVM mainly separate two classes using hyperplane. Hyperplane is a plane which separate two classes in two different planes. This hyperplane also plots two parallel planes which are marginal planes. These marginal hyperplanes are passed from one of the nearest points. These points and nearest to this margin plane are called support vectors. Figure 1 denotes the support vectors of each classes[3]. Aim of SVM is to maximize marginal distance. There are two types of hyperplanes as linear separable and non-linear separable. For that according to data we use different kernels. Kernel is used when we try to move data to higher dimensional plane. Kernels are used in non-linear SVM. Kernels are used to plot hyperplane in higher dimensional plane. There are four types of kernel SVM used which are linear, rbf (radial basis function), poly and sigmoid.

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

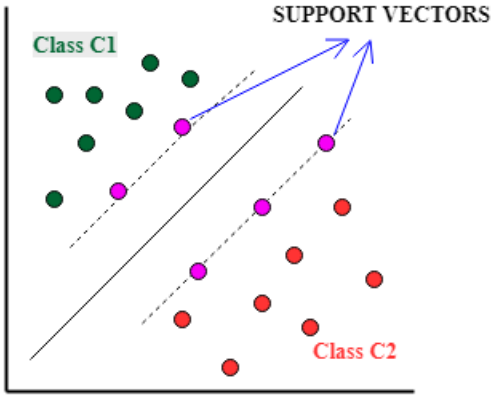

Figure 1: Support Vector Machine

II. RELATED WORK

The Support Vector Machine (SVM) model was first proposed by Vapnik andhas become one of thepopularly used model in machine learning. According to therecent studies SVM generally is known to give better accuracy opposed to otherdata classification model[1]. In spam detection, we give input features and the output gets as spam or ham. Equation y=f(x) where x is the features and y give decision output as spam or ham [2].Kumar, A et al. proposed a system where they used a new hybrid algorithm which is a combination of SVM, Neural Network, and NLP. The result of this proposed algorithm is compared with different classifiers and neural network based on different parameters. After results evaluation, they concluded that SVM and Bayes classify the data set with the highest accuracy [19].Roy S.S et al. experimented where they compare the classification of SVM, Neural Network, and deep SVM. Result analysis concluded that the classification of Deep SVM performed better than the other two models. The accuracy, precision. Recall. Accuracy, F1 score of deep SVM model is better than SVM and Neural Network [1]. V Vishagini et al. performed where they proposed a model in which they use three algorithm SVM, WSVM(Weighted SVM), and improvised SVM where they compare performance based on accuracy, precision, recall, and misclassification rate. Researchers concluded that improvised SVM has the least misclassification rate followed by WSVM and SVM [3]. Researchers H. He et al. performed four experiments where they used two datasets one the spam dataset and the other one is SMS dataset. For classification, they used the RBF kernel of SVM where each time they increased the input attribute space, and results are measured on basis of accuracy, TPR, and TNR [2].Rithesh, R. N. et al. studied first SVM, KNN and after that, they proposed a new algorithm called SVM-KNN algorithm where they used 4 datasets for the experiment in which they find out and compare the accuracy of SVM, KNN, and SVM-KNN algorithm where they concluded that the accuracy of the new algorithm is greater than individual algorithms [8]. W. Niu, et al. proposed a classification model where they use Cuckoo Search with Support Vector machine where the first email get in the first stage which contain pre-processing with feature extraction in which 40 features are extracted. After that it goes in the second stage where classifiers work on this email and give results as phishing or non-phishing [6].

Kumaresan, T. et al. developed a framework combined of S-Cuckoo and hybrid kernel SVM (MKSVM) where they use 2 datasets one for text based and other of image based. S-Cuckoo is used to select specific features from original features and multi kernel SVM for classification. Term Frequency is used for text feature extraction. For image feature extraction color correlogram and Wavelet Moment are used. Feature selection is performed using S-Cuckoo search[7]. Rithesh, R. N. proposed a new algorithm called SVM-KNN algorithm where they used 4 datasets for experiment in which they find out and compare the accuracy of SVM, KNN and SVM-KNN algorithm where they found that the accuracy of new algorithm is greater than individual algorithms[8]. I. Ahmad, et al. performed experiment for detection of violation in system where they take a datasheet and applied SVM, RF, ELF and compare performance on the basis of precision, recall and accuracy [9].

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

III. METHODOLOGY

In this experiment, SVM is used which is a machine learning classifier to classify emails as ham or spam. Also, for processing the raw email Natural Language Processing is used in the experiment.

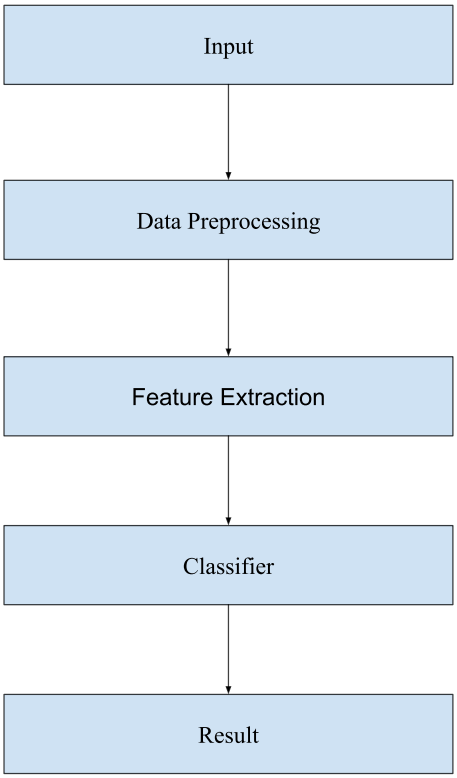

Figure 2: Methodology

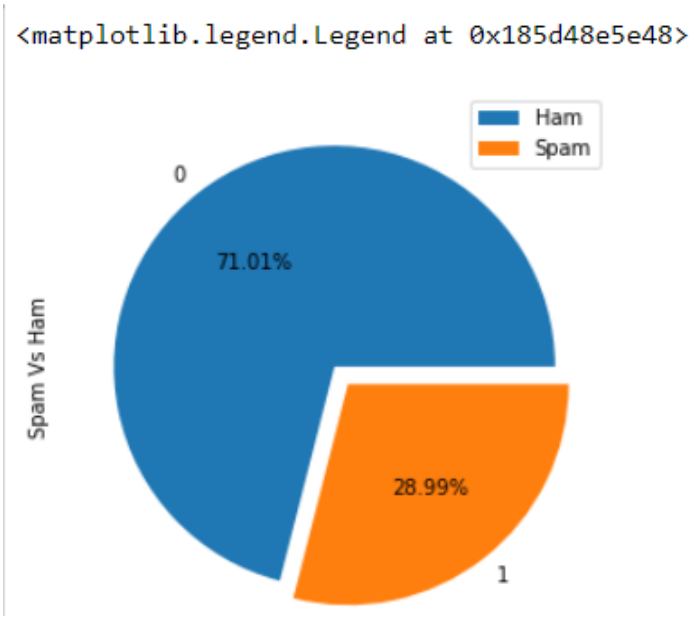

Figure 3: Pie chart for original dataset

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

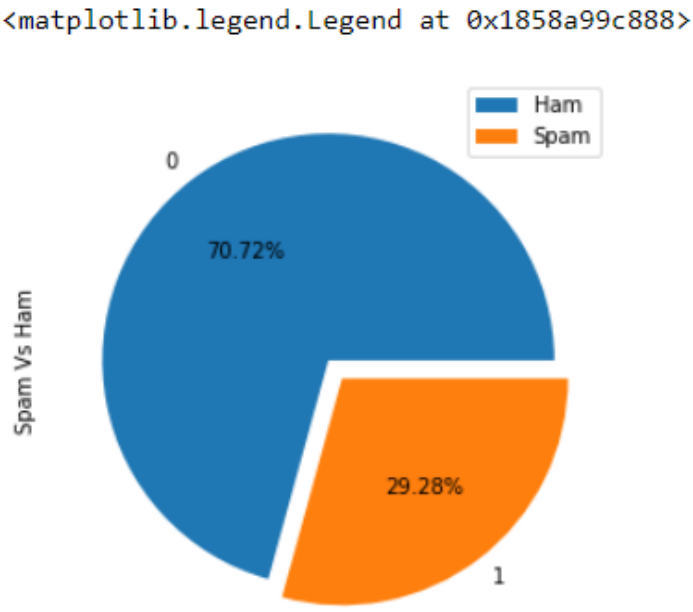

Figure 4: Pie chart after duplicates removed

Dataset-
Dataset is taken from Kaggle which contains four columns. Dataset is in CSV file format. It contains a total of 5771 values out of which 4993 are unique values. 71% of data in this dataset is spam and 29% of data is ham. In fig 3 there is a Pie chart of spam against ham. Figure 5 shows the dataset content.

:

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...	...	...
5166	1518	ham	Subject: put the 10 on the ft\r\nthe transport...	0
5167	404	ham	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	2933	ham	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	1409	ham	Subject: industrial worksheets for august 2000...	0
5170	4807	spam	Subject: important online banking alert\r\ndea...	1

5171 rows × 4 columns

Figure 5: Original dataset

Remove unwanted columns and duplicate values-
First from the dataset unwanted columns and duplicate values are removed. The dataset contains 4 columns and 2 columns are selected. Total 5171 values are present in the dataset from which 4993 are selected and duplicate values are removed.Figure 4 shows a pie chart of the dataset after removing duplicates. Figure 6 shows the dataset after duplicate values are removed.

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

	text	label_num
0	Subject: enron methanol ; meter # : 988291\r\n...	0
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	Subject: photoshop , windows , office . cheap ...	1
4	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...
5165	Subject: fw : crosstex energy , driscoll ranch...	0
5166	Subject: put the 10 on the ft\r\nthe transport...	0
5167	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5169	Subject: industrial worksheets for august 2000...	0
5170	Subject: important online banking alert\r\ndea...	1

4993 rows × 2 columns

Figure 6: Dataset after duplicates removed

Remove stopwords-
After removing unwanted columns and duplicate values Natural Language Processing is used to remove stopwords. Stopwords are those words that did not contribute to classification so they are removed. Stopwords are like ‘is’, ‘the’ etc. removal of these words helps to clean the data.
Regular expression, convert uppercase to lowercase, split the sentences into words-
Regular expression is used to select only letters and removed numbers, special characters because these numbers and special characters are not useful for classification. After that all words are converted into lowercase and split sentences into words. Machines can only understand lower case words so all words are converted into lowercase using lower() method. Sentences are split into words using .split() method.

Apply the stemmer technique-
PorterStemmer technique is used for stemming. The purpose of stemmer is to convert derivational words into their base form by cutting ends of words.For example, if there are words 'hard', 'harder', 'hardest' then after applying stemmer only the 'hard' word will consider.

Splitting data-
After performing Natural Language Processing data is divided into train and test. I divide the whole dataset into 80:20 ratio where 80% of data is trained with SVM where 20% of data is used to test the algorithm.

Feature extraction-
After splitting datafeatures are extracted using the TfidfVectorizer method. TfidfVectorizer is Term frequency inverse document frequency method which works on words frequency. It is used to convert sentences into vectors. Term Frequency is ratio of no. of repetition of words in sentence and no. of words in sentence. Inverse Document Frequency is used to remove more frequent words and use less frequent words. Inverse Document Frequency is log of ratio of no. of sentences and no. of sentences containing words.Figure 7 shows the dataset which is after data cleaning. Now this dataset is ready for feature engineering and Support Vector Machine algorithm.

IV.EXPERIMENTAL RESULTS

After data cleaning, feature extraction, and Support Vector Machine results are calculated based on the following parameters.
True Positive Rate-
It is the rate of how many positives are correctly recognized as legitimate or ham. It is also known as recall. 99.86% recall is obtained.

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

```
tpr=tp/(tp+fn)
tpr=tpr*100
print('True Positive Rate:',tpr)
```

```
True Positive Rate: 99.86431478968792
```

Figure 9: True positive Rate

```
Recall: 99.86431478968792
```

Figure 10: Recall

True Negative Rate-
It is the rate of how many actual negatives are recognized as spam.96.94% TPR rate is obtained.

```
tnr=tn/(tn+fp)
tnr=tnr*100
print('True Negative Rate:',tnr)
```

```
True Negative Rate: 96.94656488549617
```

Figure 11: True Negative Rate

False Positive Rate-
It is the rate of how many negatives are recognized as ham. 3.05% FPR rate is obtained.

```
fpr=fp/(fp+tn)
fpr=fpr*100
print('False Positive Rate:',fpr)
```

```
False Positive Rate: 3.0534351145038165
```

Figure 12: False Positive Rate

False Negative Rate-
It is the rate of how many positives are recognized as spam. 0.13% FNR rate is obtained.

```
fnr=fn/(fn+tp)
fnr=fnr*100
print('False Negative Rate:',fnr)
```

```
False Negative Rate: 0.13568521031207598
```

Figure 13: False Negative Rate

Precision-
It is a ratio of actual positives with total positives. 98.92% precision is obtained.

```
precision=tp/(fp+tp)
precision=precision*100
print('Precision:',precision)
```

```
Precision: 98.9247311827957
```

Figure 14: Precision

Accuracy-
It is how many positives are correctly recognized from all predictions. 99.09% accuracy is obtained.

```
accuracy=(tp+tn)/(tp+tn+fp+fn)
accuracy=accuracy*100
print('Accuracy:',accuracy)
```

```
Accuracy: 99.09909909909909
```

Figure 15: Recall

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

For better understanding of results Precision-recall curve and ROC curve are plotted to see how classifier performed. Figure 16 shows precision-recall curve and figure 17 shows ROC curve.

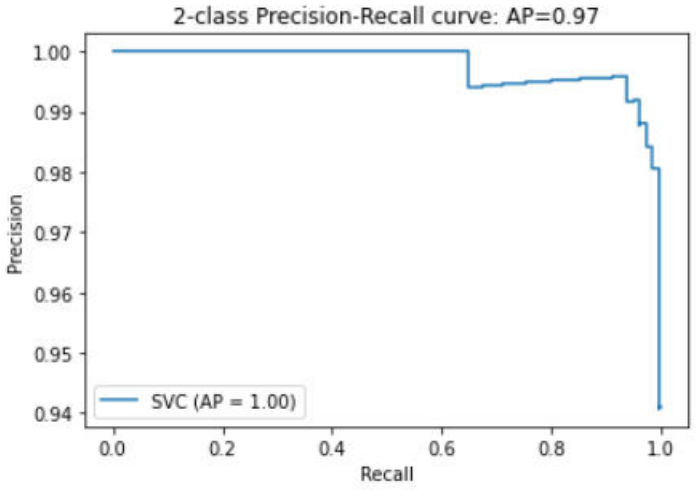

Figure 16: Precision Recall curve

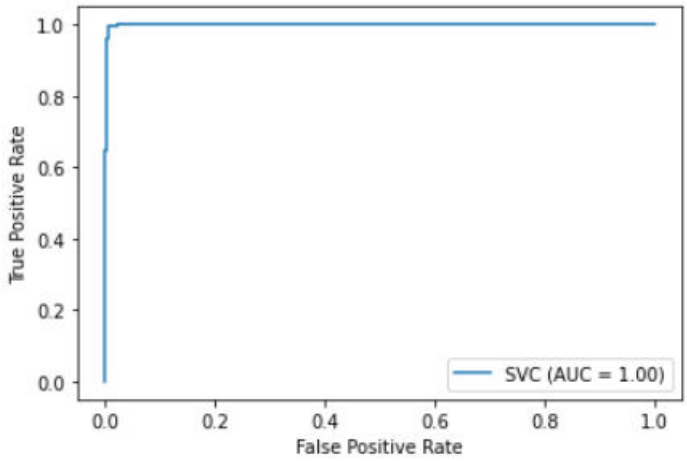

Figure 17: ROC curve

IV. CONCLUSION

In this paper I tried to see how well Support Vector Machine performs for email classification. It performs well for classification as it gives 99% accuracy which is very good. Figure 17 shows that the model have high performance and there is very little spam are classified as ham compared to actual ham one. This shows that model is working well. Again, the further improvement is possible here by using neural networks, the combination of other machine learning classifier with Support Vector Machine to increase accuracy even more.

V. ACKNOWLEDGEMENT

I would like to give Special thanks to Professor Swapna Augustine Nikale for helping me with this topic for research.

REFERENCES

[1] Roy S.S., Sinha A., Roy R., Barna C., Samui P. (2018) Spam Email Detection Using Deep Support Vector Machine, Support Vector Machine and Artificial Neural Network. In: Balas V., Jain L., Balas M. (eds) Soft

| ISSN: 2395-7852 | www.ijarasem.com | Bimonthly, Peer Reviewed & Referred Journal|

| Volume 7, Issue 1, January 2020 |

Computing Applications. SOFA 2016. Advances in Intelligent Systems and Computing, vol 634. Springer, Cham. https://doi.org/10.1007/978-3-319-62524-9_13

[2] H. He et al., "Incremental information gain analysis of input attribute impact on RBF-kernel SVM spam detection," 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, 2016, pp. 1022-1029, doi: 10.1109/CEC.2016.7743901.

[3] V. Vishagini and A. K. Rajan, "An Improved Spam Detection Method with Weighted Support Vector Machine," 2018 International Conference on Data Science and Engineering (ICDSE), Kochi, 2018, pp. 1-5, doi: 10.1109/ICDSE.2018.8527737.

[4] A. A. Alurkar et al., "A proposed data science approach for email spam classification using machine learning techniques," 2017 Internet of Things Business Models, Users, and Networks, Copenhagen, 2017, pp. 1-5, doi: 10.1109/CTTE.2017.8260935.

[5] P. Patil, R. Rane and M. Bhalekar, "Detecting spam and phishing mails using SVM and obfuscation URL detection algorithm," 2017 International Conference on Inventive Systems and Control (ICISC), Coimbatore, 2017, pp. 1-4, doi: 10.1109/ICISC.2017.8068633.

[6] W. Niu, X. Zhang, G. Yang, Z. Ma and Z. Zhuo, "Phishing Emails Detection Using CS-SVM," 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC), Guangzhou, 2017, pp. 1054-1059, doi: 10.1109/ISPA/IUCC.2017.00160.

[7] Kumaresan, T., Saravanakumar, S. & Balamurugan, R. Visual and textual features based email spam classification using S-Cuckoo search and hybrid kernel support vector machine. Cluster Comput 22, 33–46 (2019). https://doi.org/10.1007/s10586-017-1615-8

[8] Rithesh, R. N. (2017). SVM-KNN: A Novel Approach to Classification Based on SVM and KNN. *International Research Journal of Computer Science*, *4*(8), 43–49. https://doi.org/10.26562/irjcs.2017.aucs10088

[9] I. Ahmad, M. Basheri, M. J. Iqbal and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," in IEEE Access, vol. 6, pp. 33789-33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[10] M. S. Swetha and G. Sarraf, "Spam Email and Malware Elimination employing various Classification Techniques," 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), Bangalore, India, 2019, pp. 140-145, doi: 10.1109/RTEICT46194.2019.9016964.

[11] Dada, E. G., Bassi, J. S., Chiroma, H., Abdulhamid, S. M., Adetunmbi, A. O., & Ajibuwa, O. E. (2019). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon, 5(6), e01802. https://doi.org/10.1016/j.heliyon.2019.e01802

[12] S. Mishra and D. Malathi, "Behaviour analysis of SVM based spam filtering using various parameter values and accuracy comparison," 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, 2017, pp. 27-31, doi: 10.1109/ICCMC.2017.8282698.

[13] Deshmukh S., Dhavale S. (2020) Automated Real-Time Email Classification System Based on Machine Learning. In: Bhalla S., Kwan P., Bedekar M., Phalnikar R., Sirsikar S. (eds) Proceeding of International Conference on Computational Science and Applications. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-0790-8_36

[14] Zamir, A., Khan, H. U., Mehmood, W., Iqbal, T., & Akram, A. U. (2020). A feature-centric spam email detection model using diverse supervised machine learning algorithms. The Electronic Library, 38(3), 633–657. https://doi.org/10.1108/el-07-2019-0181

[15] Gangavarapu, T., Jaidhar, C.D. & Chanduka, B. Applicability of machine learning in spam and phishing email filtering: review and approaches. *Artif Intell Rev* (2020). https://doi.org/10.1007/s10462-020-09814-9

[16] S. Nandhini and J. Marseline K.S., "Performance Evaluation of Machine Learning Algorithms for Email Spam Detection," *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Vellore, India, 2020, pp. 1-4. doi: 10.1109/ic-ETITE47903.2020.312

[17] Jawale, D. S., Mahajan, A. G., Shinkar, K. R., & Katdare, V. V. (2020). Spam Detection Using Machine Learning. *Computer Engineering and Intelligent Systems*, *11*(3), 2222–2863. https://doi.org/10.7176/ceis/11-3-04

[18] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118.