JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Machine Learning predictive model to detect spam email

[1]Utsab Ray, [2]Ipsita Sengupta,[3]Karabi Ganguly

[1]BTech, Department of Biomedical Engineering, [2]BTech, Department of Electrical Engineering
[3]Associate Professor and HoD, Department of Biomedical Engineering
[1]Department of Biomedical Engineering,
[1] JIS College of Engineering, Kalyani, India

Abstract : Email is growing as a well-known communication paradigm in many corporate operations as recent advancements in communication technologies transform the world. Email is an efficient, quick, and low-effort method of correspondence. Email spam is unsolicited information sent to E-letter drops. Spam might be a major issue for both customers and ISPs. According to research, clients now receive far more spam messages than non-spam emails. In some circumstances, spam messages can harm the credibility of a company process. It can be seen that the spam filters in most popular email systems are skewed for ad profit, i.e. they provide an exception for some organizations who pay for advertising. This is not the case. Many organizations and people have found that electronic mail has simplified communication procedures. Spammers use this strategy for dishonest gain by sending unsolicited emails. The goal of this work is to offer a method for detecting spam emails using machine learning algorithms optimized with bio-inspired methodologies. On seven separate email datasets, considerable research was conducted to apply machine learning models such as Nave Bayes, Vector Machine, Stochastic Forest, Decision Tree, and Multi-Layer Perceptron, as well as extraction and classification, and pre-processing. Thus, in the end, a predictive model is developed which will predict the status of any email easily.

Keywords **- Email, spam, machine learning, email datasets, predictive model.**

I. INTRODUCTION

There are numerous platforms accessible for individuals to share information from anywhere in the world. E - mail is really the cheapest, and fastest mode of information sharing available everywhere in the globe. Yet, because of their complexity, emails are open to various types of attacks, the most popular and destructive of which is spam [1]. Nobody desires to receive emails that aren't relevant to their interests because they waste the recipients' time and resources. Furthermore, dangerous content contained in the form of files or links in these emails may lead to security breaches in the host system [2]. Spam is any useless and unwelcome communication or letter sent by the hacker to a large number of recipients via email or any other information sharing channel [3]. Spam emails may contain viruses, rats, and Trojan horses [4]. Attackers typically employ this tactic to entice customers to use internet services. They could send spam emails with attachments with multiple-file extensions, packed URLs that direct the recipient to harmful and spamming websites, and end up with data or economic fraud and identity theft [5]. Many email service providers allow customers to create keyword-based basis policies that automatically filter email. Nevertheless, this strategy is ineffective since it is costly, and customers do not want to personalize their emails, which allows spammers to hit their email accounts. The with proliferation of the Internet enabling global communication, there has been a huge increase in spam emails [6].
Spam is generated from anywhere in the globe with the assistance of the Internet by concealing the attacker's identity. Despite the availability of numerous antispam tools and strategies, the spam rate remains quite high. Malicious emails including links to infected sites which can harm the victim's data are the most harmful kind of spam. Spam emails could also slow down data from the server by filling up server memory or capacity. Every business carefully assesses the available solutions to combat spam in the environment in order to correctly detect phishing emails and prevent the escalating email spam challenges. Whitelist/Blacklist [7], mail headers analysis, phrase checking, and other well-known procedures for identifying and analyzing email messages for spam detection are examples. According to social networking specialists, 40% of social media site accounts are exploited for spam [8]. Spammers utilize popular social networking applications to transmit hidden links in the text to obscene or other product sites aimed to sell something from false accounts. The toxic emails addressed to the same individuals or organizations have consistent highlights. By looking at these highlights, it is possible to increase the identification of these types or emails. We could classify emails into trash and non spam using artificial intelligence (AI) [9]. This approach is made possible by extracting features from the messages' headers, topic, and body. By collecting this information, we can categorise it as spam or bacon.

Learning-based algorithms [10] are widely employed for spam detection today. The detection procedure in learning-based categorization implies that junk mail have such a precise set of traits that distinguish them from legitimate ones [11]. Several factors contribute to the complexity of spam detection in learning-based models. Spam subjectivity, concept drift, linguistic issues, extra processing, and text delay are among these factors.

Extreme learning machine is one example of a learning-based model (ELM). This is a contemporary machine learning paradigm for feed forward neural networks with a single hidden layer [12]. When compared to standard neural networks, it reduces slow training time and over fitting issues. It only takes one cycles of iteration in ELM. This approach in particular is increasingly used in many disciplines due to its improved generalization potential, resilience, and controllability.

Models for machine learning have been used for a variety of applications in computer science, ranging from fixing a network traffic issue to detecting malware. Many people use email for communication and socialising on a regular basis. Security flaws damage consumer data, allowing'spammers' to fake a hacked email account and send illicit (spam)emails. This is also used to acquire unauthorised access to the user's device by fooling the customer into clicking on the spam link within the spam email, which is a phishing assault [13].

Companies provide numerous tools and strategies for detecting spam emails in a network. Filtering methods have been set up by organizations to detect unsolicited emails by creating rules and configuring firewall settings. Google is one of the leading organizations that claim to detect such emails with a 99.9% success rate [14]. Spam filters can be deployed in a variety of locations, such as at the gate route (router), cloud-hosted applications, or the user's machine. Methods such as content-based filtering, rule-based filtering, and Bayesian filtering have been used to tackle the detection problem of spam emails. Unlike 'knowledge engineering,' where malware detection rules are created up and must be manually updated on a regular basis, ingesting time and resources, computer vision simplifies the process because it realizes to recognize unsolicited emails (spam) and legitimate emails (ham)automatically and then applies some these learned commands to unidentified incoming emails.

II. Related Work

Spam emails are becoming more widespread by the day, and have become a common nuisance over the previous decade. Spambots generally collect email addresses that receive spam . Machine learning applications have proved crucial in the identification of spam emails. It contains a number of models and strategies that researchers are employing to create unique spam recognition and filters models [15].

Kaur and Verma [16] report a survey on detecting email spam using a supervised technique with feature selection. They talk about the process of discovering knowledge for spam detection equipment. They also elaborate on various spam detection strategies and technologies. This poll also addresses the selection of characteristics based on N-Gram.

N-Gram [17] is an easily interpretable technique that predicts the likelihood of the forthcoming word occurrences after locating N 1 terms in a phrase or text corpus. N-Gram predicts the next word using probability-based approaches. For email spam detection, they examine various computer vision) and non machine learning techniques [18].

A survey into intelligent junk mail detection is presented by Saleh et al. [19]. They examine several email security issues, including spam emails, the breadth of spam analysis, and various evolutionary computation and non machine learning strategies for spam identification and filtering. They find that guided learning [20] algorithms are widely used for email spam detection. According to them, the great utilization of supervised methods is due to the consistency and precision of monitored procedures. They also examined multi algorithm frameworks and discovered that they were more efficient than a single method. They discovered that practically all research that uses email content to identify spam, especially phishing emails, relies on word-based categorization or clustering techniques.

Blanzieri and Bryl [21] present a collection of learning-based phishing mails filtering methods. They addressed spam issues and gave a study of learning-based phishing detection in their study. They describe several aspects of spam emails. The consequences of spam emails affecting various domains were explored in this study. This study also addresses several economic and ethical considerations with spam. The common antispam technique and learning-based filtering are well developed. The most prevalent filters are built on various classification algorithms applied to different components of email communications. This study reveals that the Nave Bayes classifier has a distinct advantage over other machine learning used for spam detection. It produces high accuracy outcomes at a superb pace and simplicity.

Bhuiyan et al. [22] provide an overview of current spam mails filtering techniques. By studying numerous processes, they aggregate multiple spam filtering algorithms and sum it up the accuracy on various characteristics of different proposed systems. They discuss how all of the existing approaches for filtering phishing emails are effective. Some have had positive outcomes, while others are experimenting with new methods to improve their accuracy performance. Despite their popularity, they all have concerns with spam filtering technologies, which is the key concern for researchers.

They are attempting to develop a next-generation sms spam method capable of comprehending vast amounts of multimedia information and screening spam emails. They conclude that the majority of spam mails filtering is done using Naive Bayes and the Classifier. Spam filtration can be taught on a variety of datasets, including the "ECML" and UCI databases [23].

Ferrag et al. [13] reviewed deep learning techniques used in systems that detect intrusions and spam review detection datasets. They discussed several detection techniques using deep learning models & assessed their performance. They divided 35 well-known cyber datasets into seven categories and studied them. Online traffic-based, networking traffic-based, Interanet traffic-based,

electricity network-based, virtual private infrastructure, andriod app-based, IoT traffic-based, and Internet linked device-based datasets are among these groups. They come to the conclusion that deep learning models outperform classical machine learning and lexical models for incursion and spam detection.

Vyas et al. [24] provide a review of supervised machine learning algorithms for spam email filtering. They determined that the Naive Bayes method outperforms all other strategies presented in terms of speed and precision (excluding SVM and ID3). SVM and ID3 provide more precision than Naive Bayes but require a considerably longer time to build a system. There's a trade-off among precision and timing. They find that the choice of learning algorithm is highly influenced by the situation as well as the necessary accuracy and time. They suggest that in the future, all components of a email should be analyzed to construct a more powerful spam screening framework.

III. PROPOSED WORK

The work has been conceptualized in figure(i).

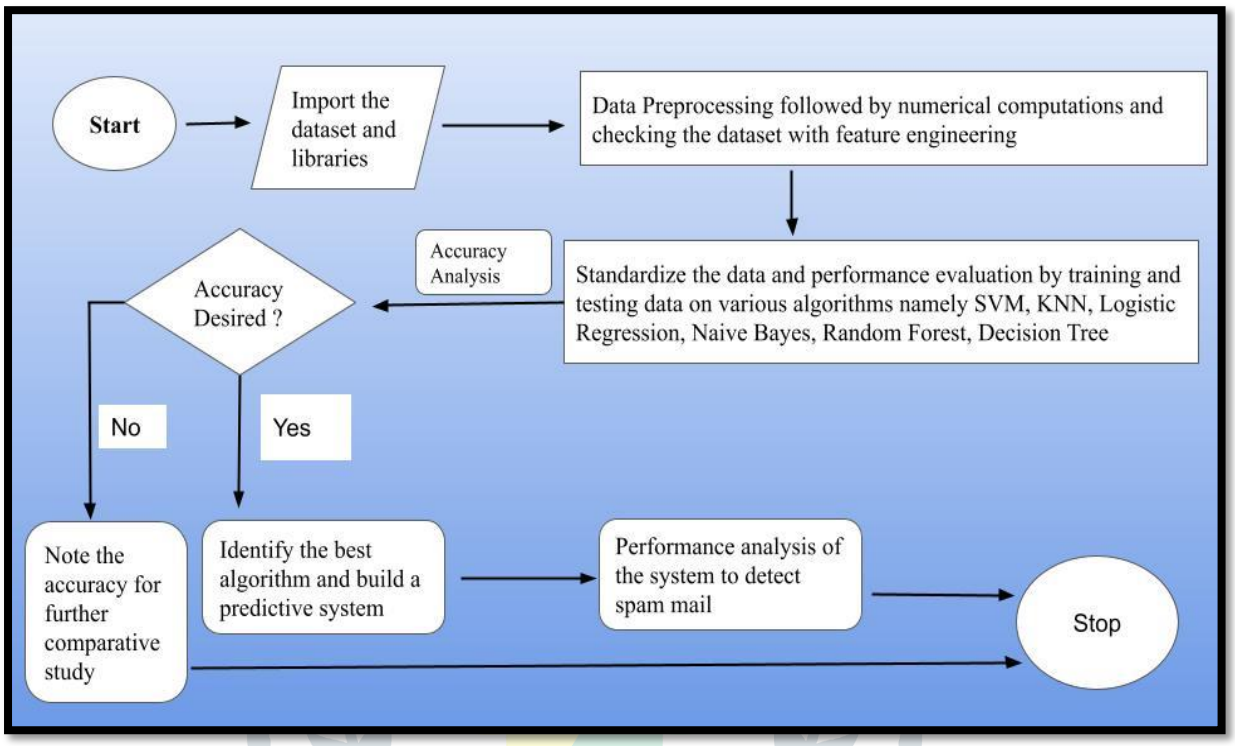

Figure (i) - Proposed Methodology

The dataset was initially collected, and the machine learning pipeline was started. Machine Learning is primarily based on the premise of training data, followed by testing the model, checking for predictions, and further deployment. The graphic (ii) depicts the machine learning process pipeline.

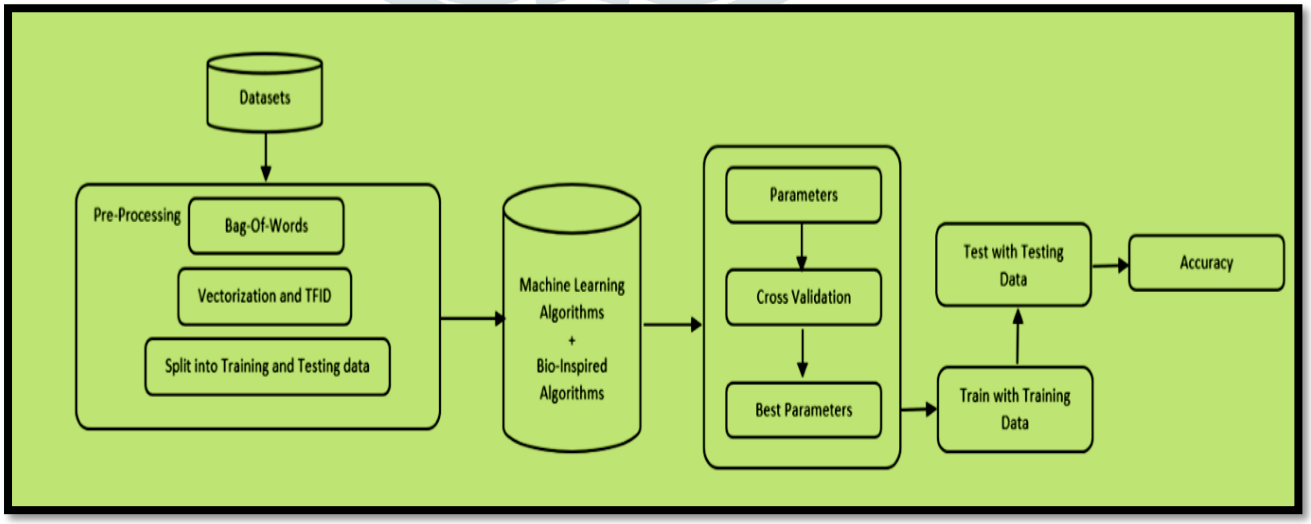

Figure (ii): Machine Learning workflow

Following this pipeline, the algorithms Random Forest, KNN, Decision Tree, Nave Bayes, and Logistic Regression are distributed on just this special dataset to calculate credit card frauds, and it is visualised that Logistic Regression performed better on testing data. It is then necessary to develop a predictive system that can easily determine which fraud claim pertaining to any type of card.

IV. RESULTS AND DISCUSSIONS

```
[ ] print(raw_mail_data)

     Category                                            Message
0         ham  Go until jurong point, crazy.. Available only ...
1         ham                      Ok lar... Joking wif u oni...
2        spam  Free entry in 2 a wkly comp to win FA Cup fina...
3         ham  U dun say so early hor... U c already then say...
4         ham  Nah I don't think he goes to usf, he lives aro...
...       ...                                                ...
5567     spam  This is the 2nd time we have tried 2 contact u...
5568      ham              Will ü b going to esplanade fr home?
5569      ham  Pity, * was in mood for that. So...any other s...
5570      ham  The guy did some bitching but I acted like i'd...
5571      ham                        Rofl. Its true to its name

[5572 rows x 2 columns]
```

Figure (iii): Classify the dataset

Figure (iii) depicts dataset classification, which is an important criterion for doing any computational examination.

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Figure (iv): Categorization of dataset

Label Encoding

```
[ ] # label spam mail as 0;  ham mail as 1;

    mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
    mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1
```

spam - 0

ham - 1

Figure (v): Label Encoding

Label Encoding is a well-known encoding technique for dealing with categorical information. Based on alphabetical ordering, each label is issued a unique integer in this technique as shown in figure (v).

Splitting the data into training data & test data

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)

[ ] print(X.shape)
    print(X_train.shape)
    print(X_test.shape)

    (5572,)
    (4457,)
    (1115,)
```

Figure (vi): Splitting into training data and testing data

Feature Extraction

```
# transform the text data to feature vectors that can be used as input to the Logistic regression

feature_extraction = TfidfVectorizer(min_df = 1, stop_words='english', lowercase='True')

X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)

# convert Y_train and Y_test values as integers

Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

```
[ ] print(X_train)
```

```
[ ] print(X_train_features)
```

Figure (vii): Feature Extraction

The process of translating raw data into a format features that may be processed while keeping the knowledge in the underlying data set is referred to as feature extraction. It produces better outcomes than merely applying machine learning to raw data as shown in figure (vii).

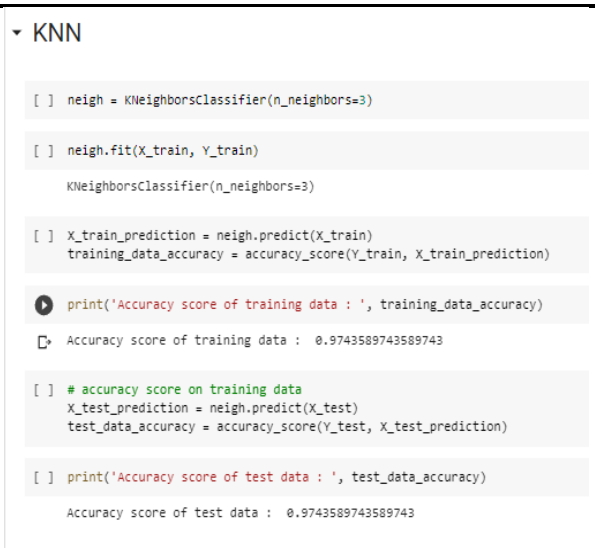

Figure (viii):- Accuracy analysis of KNN

Figure (ix):- Accuracy of Naïve Bayes

Figure (x):- Accuracy of Random Forest

Decision Tree

```
[ ] clf = tree.DecisionTreeClassifier()

[ ] clf.fit(X_train, Y_train)

    DecisionTreeClassifier()

    X_train_prediction = clf.predict(X_train)
    training_data_accuracy = accuracy_score(Y_train, X_train_prediction)

[ ] print('Accuracy score of training data : ', training_data_accuracy)

    Accuracy score of training data :  1.0

[ ] # accuracy score on training data
    X_test_prediction = gnb.predict(X_test)
    test_data_accuracy = accuracy_score(Y_test, X_test_prediction)

[ ] print('Accuracy score of test data : ', test_data_accuracy)

    Accuracy score of test data :  0.6410256410256411
```

Figure (xi):- Accuracy of Decision Tree

```
[ ] # prediction on training data

    prediction_on_training_data = model.predict(X_train_features)
    accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)

[ ] print('Accuracy on training data : ', accuracy_on_training_data)

    Accuracy on training data :  0.9670181736594121

[ ] # prediction on test data

    prediction_on_test_data = model.predict(X_test_features)
    accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)

[ ] print('Accuracy on test data : ', accuracy_on_test_data)

    Accuracy on test data :  0.9659192825112107
```

Figure (xii):- Accuracy of Logistic Regression

Thus, the comparative study between various algorithms with respect to their algorithms can be summarized as below in table 1:

Sl. No	Algorithm Name	Training Accuracy	Testing Accuracy	Rank Based on Accuracy
1	Decision Tree	100%	64.10%	4
2	Random Forest	96.75%	89.74%	3
3	Logistic Regression	96.70%	96.59%	2
4	Naïve Bayes	78.84%	64.10%	5
5	KNN	97.43%	97.43%	1

Table 1

So, from Table 1, it is clear that KNN performs best and Naïve Bayes has less accuracy on performance. Therefore, predictive system is developed with the help of Logistic Regression algorithm.

Building a Predictive System

```
[ ] input_mail = ["I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. Yo

    # convert text to feature vectors
    input_data_features = feature_extraction.transform(input_mail)

    # making prediction

    prediction = model.predict(input_data_features)
    print(prediction)


    if (prediction[0]==1):
      print('Ham mail')

    else:
      print('Spam mail')

    [1]
    Ham mail
```

Figure (xiii):- Predictive System

Figure(xiii) portrays a predictive system developed with the machine learning algorithm bearing highest accuracy for test data and yielding accurate results on that particular mail.

V. CONCLUSIONS

From this work, we can conclude that a predictive system is developed with the help of Machine Learning algorithm that performed best on this respective dataset to compute the fraud for a particular e mail. This system will help to check and identify any sort of fraud or spam email from any recipient.

VI. ACKNOWLEDGMENT

We would like to thank all the faculty and staff members from the Department of Biomedical Engineering and Department of Computer Applications for their support.

REFERENCES

[1] Ali, H. Faris, A. M. Al-Zoubi, A. A. Heidari et al., "An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks," Information Fusion, vol. 48, pp. 67–83, 2019.

[2] E. Blanzieri and A. Bryl, "A survey of learning-based techniques of email spam filtering," Artificial Intelligence Review, vol. 29, no. 1, pp. 63–92, 2008.

[3] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email classification using artificial neural network," International Journal for Academic Development, vol. 2, 2018.

[4] N. Udayakumar, S. Anandaselvi, and T. Subbulakshmi, "Dynamic malware analysis using machine learning algorithm," in Proceedings of the 2017 International Conference on Intelligent Sustainable Systems (ICISS), IEEE, Palladam, India, December 2017.

[5] S. O. Olatunji, "Extreme Learning machines and Support Vector Machines models for email spam detection," in Proceedings of the 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, Windsor, Canada, April 2017.

[6] J. Dean, "Large scale deep learning," in Proceedings of the Keynote GPU Technical Conference, San Jose, CA, USA, 2015.

[7] R. Talaei Pashiri, Y. Rostami and M. Mahrami, "Spam detection through feature selection using artificial neural network and sine–cosine algorithm," Mathematical Sciences, vol. 14, no. 3, pp. 193-199, 2020..

[8] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: dark of the social networks," Journal of Network and Computer Applications, vol. 79, pp. 41–67, 2017.

[9] A. Barushka and P. Hájek, "Spam filtering using regularized neural networks with rectified linear units," in Proceedings of the Conference of the Italian Association for Artificial Intelligence, Springer, Berlin, Germany, November 2016.

[10] F. Jamil, H. K. Kahng, S. Kim, and D. H. Kim, "Towards secure fitness framework based on IoT-enabled blockchain network integrated with machine learning algorithms," Sensors, vol. 21, no. 5, p. 1640, 2021.

[11] M. H. Arif, J. Li, M. Iqbal, and K. Liu, "Sentiment analysis and spam detection in short informal text using learning classifier systems," Soft Computing, vol. 22, no. 21, pp. 7281–7291, 2018.

[12] X. Zheng, X. Zhang, Y. Yu, T. Kechadi, and C. Rong, "ELM-based spammer detection in social networks," The Journal of Supercomputing, vol. 72, no. 8, pp. 2991–3005, 2016.

[13] W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, ''A support vector machine based Naive Bayes algorithm for spam filtering,'' in Proc. IEEE 35th Int. Perform. Comput. Commun. Conf. (IPCCC), Dec. 2016, pp. 1–8, doi: 10.1109/pccc.2016.7820655.

[14] 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, approachesandopenresearchproblems,''Heliyon,vol.5,no.6,Jun.2019, Art. no. e01802, doi: 10.1016/j.heliyon.2019.e01802.

[15] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study," Journal of Information Security and Applications, vol. 50, Article ID 102419, 2020.

[16] N. Kumar and S. Sonowal, "Email spam detection using machine learning algorithms," in Proceedings of the 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 108–113, Coimbatore, India, 2020.

[17] I. Santos, Y. K. Penya, J. Devesa, and P. G. Bringas, "N-grams-based file signatures for malware detection," ICEIS, vol. 9, no. 2, pp. 317–320, 2009.

[18] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzi, "On the capability of evolved spambots to evade detection via genetic engineering," Online Social Networks and Media, vol. 9, pp. 1–16, 2019.

[19] A. J. Saleh, A. Karim, B. Shanmugam et al., "An intelligent spam detection model based on artificial immune system," Information, vol. 10, no. 6, p. 209, 2019.

[20] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: a review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol. 160, pp. 3–24, 2007.

[21] E. Blanzieri and A. Bryl, E-mail Spam Filtering with Local SVM Classifiers, University of Trento, Trento, Italy, 2008.

[22] H. Bhuiyan, A. Ashiquzzaman, T. Islam Juthi, S. Biswas, and J. Ara, "A survey of existing e-mail spam filtering methods considering machine learning techniques," Global Journal of Computer Science and Technology, vol. 18, 2018.

[23] A. Asuncion and D. Newman, "UCI machine learning repository," 2007, https://archive.ics.uci.edu/ml/index.php.

[24] T. Vyas, P. Prajapati, and S. Gadhwal, "A survey and evaluation of supervised machine learning techniques for spam e-mail filtering," in Proceedings of the 2015 IEEE international conference on electrical, computer and communication technologies (ICECCT), IEEE, Tamil Nadu, India, March 2015.