IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

Email Spam Detection Using Machine Learning Algorithms

ANGEL FELCIYA I [1], ESAKKI DEVI S[2], MAHESWARI.M[3], DR. ROSELIN MARY S[4]

Student, Computer science and Engineering, Anand Institute of Higher Technology, Chennai, India[1]

Student, Computer science and Engineering, Anand Institute of Higher Technology, Chennai, India[2]

Assistant Professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[3]

Head of Department, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[4]

Abstract: Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams are also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious links through spam emails which can harm our system and can also seek in into your system. Creating a fake profile and email account is much easier for the spammers, they pretend like a genuine person in their spam emails, these spammers target those people who are not aware about these frauds. So, it is necessary to Identify those spam mails which are fraudulent. This project will identify those spam by using techniques of machine learning, this paper will discuss the machine learning algorithms and apply all these algorithms on our data sets and the best algorithm is selected for the email spam detection having best precision and accuracy.

Keywords: Machine learning, Naïve Bayes, support vector machine-nearest neighbour, random forest, bagging, boosting, neural networks.

I. INTRODUCTION

Email spam, also known as electronic mail spam, is the practice of sending unwanted emails or commercial emails to a list of subscribers. Unsolicited emails signify that the recipient has not given consent to receive them.Since last decade, using spam emails has grown in popularity. Spam has grown to be a significant online problem. Spam wastes space, time, and message delivery. Although automatic email filtering may be the best way to stop spam, modern spammers may quickly get around all of these apps. Prior to a few years ago, the majority of spam that came from particular email addresses could be manually stopped.

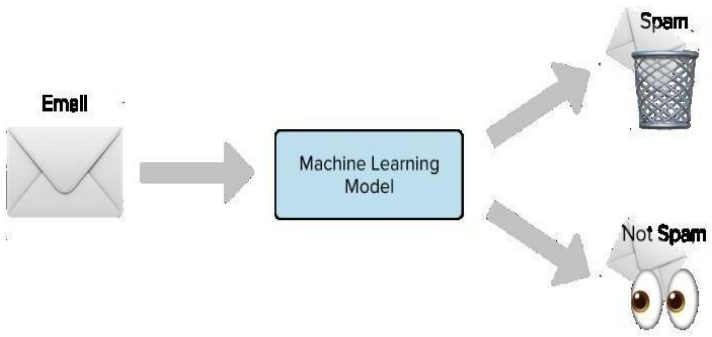


Fig.1. Classification into Spam and non-spam

II. LITERATURE REVIEW

There is some related research by A. Karim, S. Azam, B. Shanmugam, K. Kannoorpatti, and M. Alazab that uses machine learning techniques for email spam identification.[ii] They discuss a concentrated literature review of machine learning and Artificial Intelligence (AI) algorithms for email spam detection.

T. Kumar, K. Agarwal [3,] and Harisinghaney et alThe "image and textual dataset for the e-mail spam detection with the use of various methods" has been used by (2014) [4] and Mohamad & Selamat (2015) [v]. With experiments on a dataset, Harisinghaney et al. (2014) [iv] applied the KNN algorithm, Naive Bayes, and Reverse DBSCAN algorithm. OCR library"

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125124

[iii] is used for text recognition, although it doesn't work well. The feature selection hybrid strategy of TF-IDF (Term Frequency Inverse Document Frequency) and Rough pure mathematics is used by Mohamad & Selamat (2015) [v].

A. Data Set

This model utilises email data sets from many internet resources, including Kaggle and Sklearn, as well as some custom data sets. The spam.csv data set, which has 5573 lines and two columns, and the other data sets, which contain 574,1001,956 lines of email data set in text format, are utilised to train our model and to obtain results.

III. METHODOLOGY

A. Data preprocessing*:*

A really large data set with a significant number of rows and columns will always be noticed when the data is taken into account. However, this is not always the case because the data could be in the form of image, audio, or video files. Detailed tables, etc. Machines simply comprehend 1s and 0s; they are incapable of understanding photos, video, or text data.

Steps in Data Preprocessing:

Data cleaning: The tasks of "filling in missing values," "smoothing noisy data," "identifying or removing outliers," and "resolving inconsistencies" are completed in this step.

Integration of data: Several databases, information files, or information sets are added in this step.

Data transformation: Scaling up to a particular value is accomplished by aggregation and normalisation.Data compression This part gets a summary of the dataset, which is really small yet consistently yields the same analytical conclusion thus far.

1. Stop words:

Stop words are English words that don't significantly add to a sentence's meaning. They can be safely disregarded without affecting the sentence's meaning. For instance, if a search for "how to make a veg cheese sandwich" is attempted, the search engine will attempt to look for online sites that contain the words "how", "to","make", "a","veg", "cheese","sandwich". The search engine looks for websites that contain the terms "how","to","a" rather than pages that contain veg cheese sandwich recipes because these terms are so often used in English.If these three terms are dropped or abandoned in favour of fetching web pages that contain the keywords "veg," "cheese," and "sandwich," it would result in

2.Tokenization:

Tokenization is the process of separating a stream of written text into tokens, which can be phrases, symbols, words, or other expressive elements. The list of tokens is also used to contribute to subsequent handling, like content mining and parsing. Tokenization is useful for lexical analysis in software engineering and building as well as semantics (where it serves as content separation). It might be challenging to define what is meant by the term "word" at times. As word-level tokenization takes place. A token frequently relies on simple heuristics, such as: Whitespace characters such as "line break" and "space" or "punctuation characters" are used to separate tokens.

Each adjacent group of alphanumeric letters, just like each group of digits, makes up a single token.

3.Bag of words :

A technique for extracting features from text texts is called Bag of Words (BOW). Furthermore, these traits may be employed in the development of machine learning algorithms. The training dataset's documents' unique words are collected into a vocabulary by Bag of Words.

B. CLASSIC CLASSIFIERS

Data analysis that uses classification extracts the models characterising significant data classes. The construction of a classifier or model enables the prediction of class labels, such as "A loan application as risky or safe."
The process of classifying data involves two steps

-learning and building a classification model.
- classification phase.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

1. NAVIES BAYES

Spam detection was accomplished in 1998 using a naive Bayes classifier. For supervised learning, there is an algorithm called the Naive Bayes classifier. The Bayesian classifier uses dependent events and calculates the likelihood that an event that has already happened can predict an event that will happen in the future. On the basis of the Bayes theorem, which presumes that features are independent of one another, naive Bayes was developed. The Naive Bayes classifier method can be used to categorise spam emails because word probability is the key factor at play . any word that frequently appears in spam but not in ham indicates that the classifier method can be used to categorise spam emails because word probability is the key factor at play. Any word that frequently appears in spam but not in ham indicates that an email is spam. For email filtering, the Naive Bayes classifier algorithm has emerged as the most effective method.

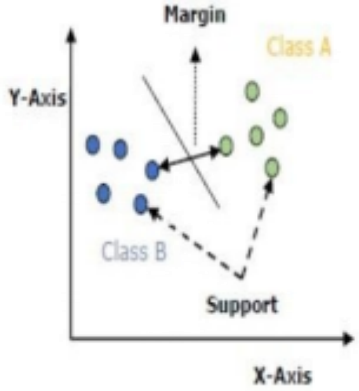


Fig.2. Support Vector Machine

1. SUPPORT VECTOR MACHINE

"The Support Vector Machine (SVM) is a popular Supervised Learning algorithm, the Support Vector model is used for classification problems in Machine Learning techniques. "Decision points served as the primary inspiration for Support Vector Machines. The Support Vector Machine algorithm's primary resolution is to draw a line or decision boundary. The Support Vector Machine algorithm produces a hyperplane that can categorise fresh samples. Each class is present in one side of a hyperplane, which divides a plane into two pieces in two dimensions. The Naive Bayes classifier uses dependent events and calculates the likelihood that an event that has already happened can predict an event that will happen in the future . On the basis of the Bayes theorem , which presumes that features are independent of one another , naïve bayes were developed . Spam emails can be classified using the Naive Bayes classification approach because word likelihood is a key in this process . any word that frequently appears is spam that not in ham indicate that email is spam

for filtering,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

(2)

Entropy using the freq

$$P(B) = \sum_{y} P(B|A)P(A)$$

:

(3)

$$E(T,X) = \sum_{c \in X} P(c)E(c)$$

Entropy using the frequency table of two attributes:

(4)

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

3. DECISION TREEThe process of learning a decision tree from class- labelled training tuples is known as "decision tree induction." A decision tree is built similarly to a flowchart. Test on attribute for internal nodes or non-leaf nodes. Branch = displays the test's results A class label is held by a leaf node. Root node is the top node.

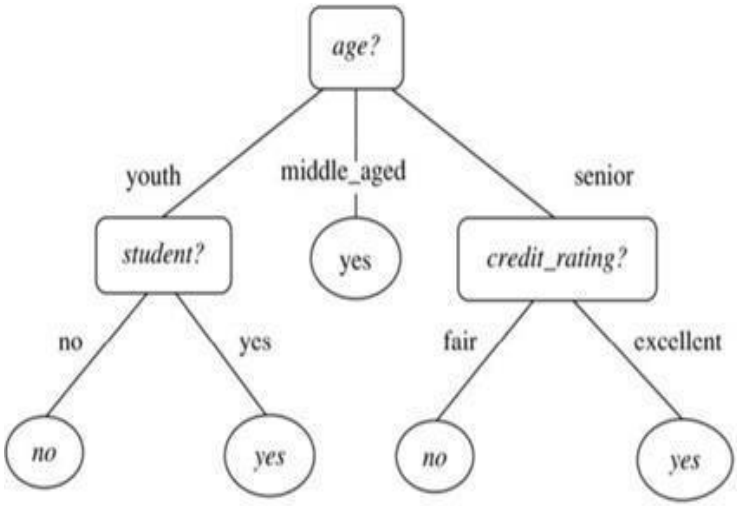


Fig .3.Decision Tree Structure

1. K- NEAREST NEIGHBOUR :

"The supervised classification method K-nearest neighbours. In order to forecast how a new sample point will be classified, this algorithm uses certain data points and a data vector that have been divided into a number of classes.

An inefficient algorithm is K- Nearest Neighbour. LAZY algorithms only attempt to memorise the steps that they cannot learn on their own. It doesn't make decisions on its own.

A new point is classified using the K-Nearest Neighbour algorithm using a similarity metric, which can be Euclidean distance.

The neighbours of an object are determined by the Euclidean distance measurement.

$$\text{Dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2} \qquad (5)$$

A. ENSEMBLE LEARNING METHODS

In order to reduce variability by reducing bagging bias by boosting predictions through stacking, "ensemble methods in machine learning" are an approach that uses multiple base models to create a predictive model. Two Sorts Base classifiers are constructed sequentially in this case. Base classifiers are running in parallel in this case.

1. RANDOM FOREST CLASSIFIER

A random forest classifier is an ensemble tree classifier made up of many decision trees of various sizes and shapes.

when a tree is being built, the random selection of the training data. When splitting at a node in a tree, input features are divided into random subgroups. The decision tree will appear less correlated if there is unpredictability.

2. BAGGING

Bagging classifiers are ensemble classifiers that fit base classifiers to individual random subsets of the original data sets and then average or vote their individual results to create a final prediction. Aggregating and bootstrapping are both used in bagging.Bagging= Bootstrap Aggregating, By simply resampling the data from the training data with the same cardinality as the original data set, bootstrapping reduces the classifier's variance and reduces overfitting. Having a high variance is bad for the model. Bagging is an extremely efficient method for little amounts of data, and you may estimate something by simply averaging the scores from samples. ensemble's generalisation error (the tree's features shouldn't be identical).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

3. BOOSTING AND ADABOOST CLASSIFIER

"Boosting is an ensemble method used to combine several weak classifiers into one strong classifier. Creating a model from training data sets and then another model to correct the previous model's flaws is how boosting is finished. [8] until the training set can be accurately predicted, are added in the boosting model. AdaBoost= Adaptive Bo**ost**ing The first successful boosting algorithm chosen for binary classification is called AdaBoost. AdaBoost is used to comprehend the boosting.

IV ALGORITHMS

1.1. Insert the training or testing dataset or file.
1.2. Verify the supported encoding in the dataset.
1.2.1. If one of the encodings is present, proceed to step 1.4.
1.2.2. Proceed to step 1.3 if one of the supported encodings is not available.
1.3 The inserted file's encoding should be changed to one of the supported encodings. Then try reading once more.
1.4 Make a decision for how you want to use the dataset to "Train," "Test," or "Compare" the models.
1.4.1 If "Train" is chosen, move on to step 1.5;
1.4.2 if "Test" is chosen, move on to step 1.6;
1.4.3. if "Compare" is chosen, move on to step 1.7.
1.5."Train" was chosen:
1.5.1. Determine which classifier should be trained with the added dataset.
1.5.2 Verify the data for NAN values and duplicates. Find the hyperparameter tuning values in
1..5.3.Process the text for a feature transform in
1.5.4. Train the model
1.5.5. Save the features and model. Display the outcomes.
1.5.6 Using the inserted dataset, decide which classifier to test.
1.5.7. A NAN value or duplication should be checked.
1.5.8 Load the model and features that were stored during the model's training phase.
1.5.9.Testing the dataset using the loaded values.
1.5.10.Show the outcomes
1.6. With "Compare" chosen:
1.6.1. Utilise the inserted dataset to compare each classifier.
1.6.2. Display the classifiers' findings

A. Implementation

The model is implemented using the Visual Studio Code platform, and the training dataset in this module is taken from the "Kaggle" website. To improve machine performance, the inserted dataset is first examined for duplicates and null values. The dataset is then divided into two smaller datasets, say the "train dataset" and "test dataset," in a ratio of 70:30. The "train" and "test" datasets are then provided as text-processing input parameters. Punctuation marks and terms on the stop words list are taken out during text processing and replaced with clean words.

The term "Feature Transform" is then passed using these clear phrases. The clean words that the text-processing procedure returns are employed in feature transform to produce a feature set. Additionally, the dataset is subjected to "hyperparameter tuning" to determine the best classifier settings based on the dataset.

The machine is fitted using those values along with a random state after obtaining the values from the "hyperparameter tuning". The characteristics and state of the trained model are retained for use for testing new data in the future. The machines are trained using the above-mentioned values using classifiers from python's learn package.

Data that is not utilised for training the trained system. In order to test it. Utilize the Save As Command to duplicate the template file.

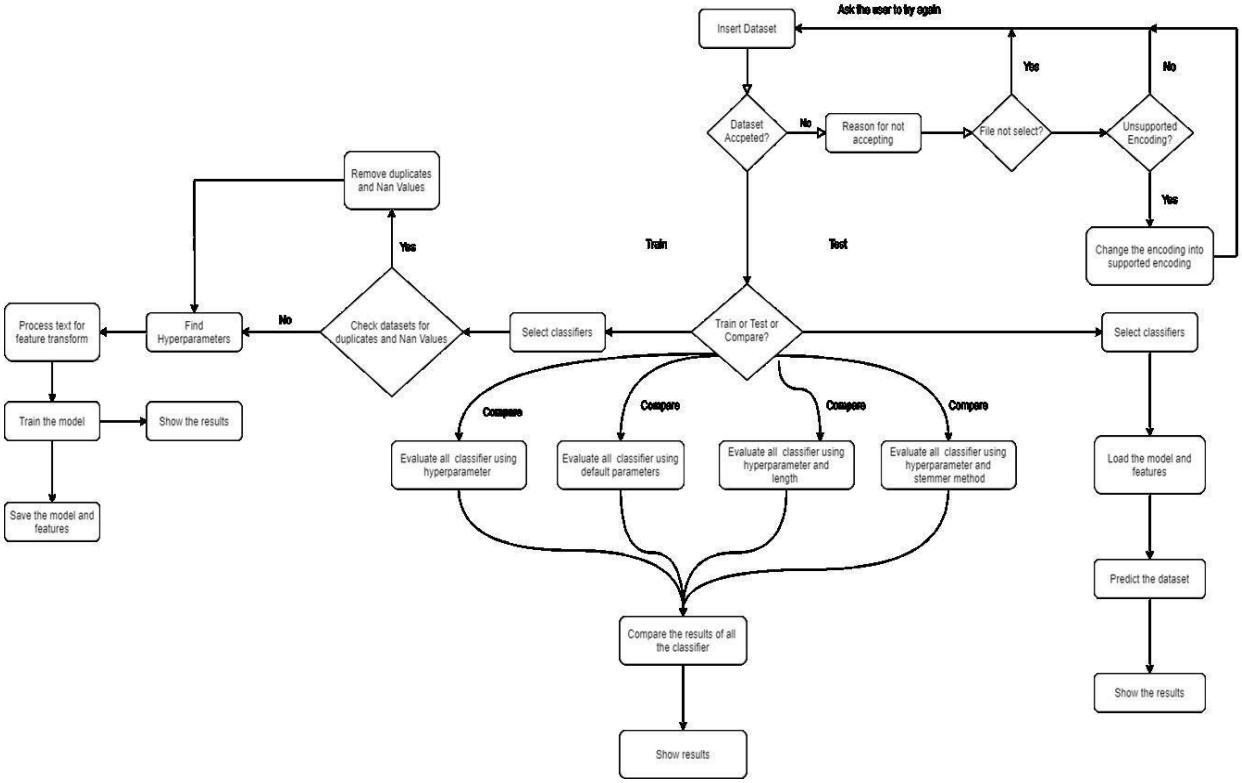
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

B. Flow Chart of the model



V. RESULT

once all classifiers have returned their findings to the user; In order to determine if the material is "spam" or "ham," the user can then compare it to other finding For easier understanding, graphs and tables will be used to display each classification result.For training, the dataset is used from the "Kaggle"website. "spam.csv" is the name of the dataset that was used. A new CSV file is created with unseen data, or to ensure greater accuracy, our model was trained using the naming convention prescribed by your conference for several classifiers, which were then checked and compared. The file used for training of the machine is named “email.csv”. After the paper has been ready, the newly created text file, and we will receive the results of each classifier after evaluation.

TABLE 1

COMPARISON TABLE

	Classifiers	***Score 1***	***Score 2***	***Score 3***	***Score 4***
1	Support Vector Classifier	0.81	0.92	0.95	0.92
2	K-Nearest Neighbour	0.92	0.88	0.87	0.88
3	Naïve Bayes	0.87	0.98	0.98	0.98
4	Decision Tree	0.94	0.95	0.93	0.95
5	Random Forest	0.90	0.92	0.92	0.92
6	AdaBoost Classifier	0.95	0.94	0.95	0.94
7	Bagging Classifier	0.94	0.94	0.95	0.94

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.125124

a. score 1: using default parameters

b. score 2: using hyperparameter tuning

c. score 3: using stemmer and hyperparameter tuning

d. score 4: using length, stemmer and hyperparameter tuning

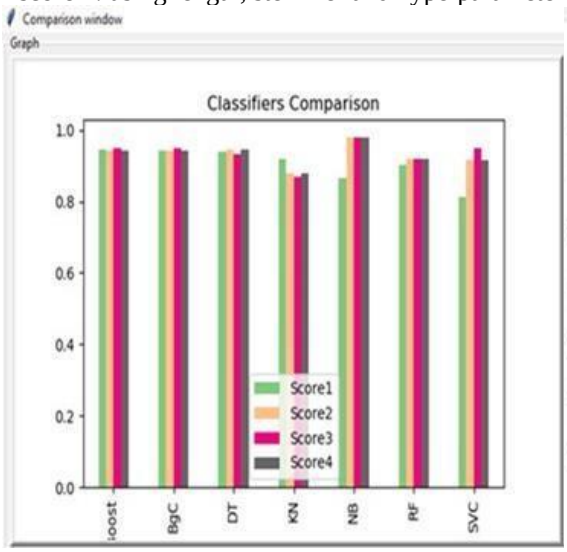


Fig.5 Comparison of all algorithms

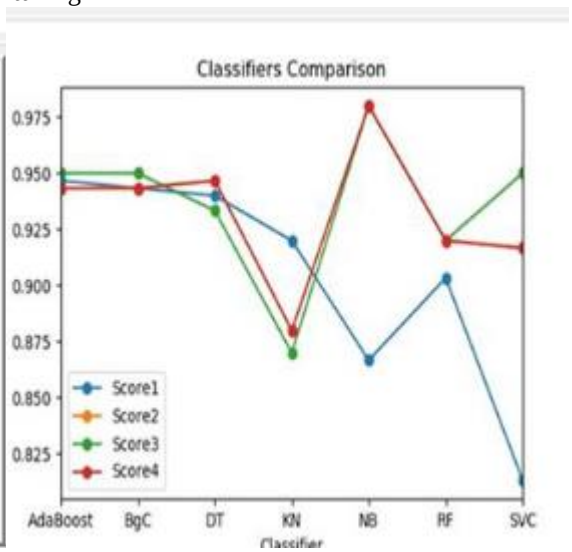


Fig.6. Comparison Graph

VI. CONCLUSION

With this result, it can be concluded that the Multinomial Naïve Bayes gives the best outcome but has limitations due to class- conditional independence which makes the machine misclassify some tuples. Ensemble methods on the other hand proved to be useful as they used multiple classifiers for class prediction. Nowadays, lots of emails are sent and received and it is difficult as our project is only able to test emails using a limited amount of corpus. Our project, thus spam detection, is proficient in filtering mails giving to the content of the email and not according to the domain names or any other criteria. Therefore, at this it is an only limited body of the email. There is a wide possibility of improvement in our project. The subsequent improvements can be done:

"Filtering of spams can be done on the basis of the trusted and verified domain names."
"The spam email classification is very significant in categorizing e- mails and to distinguish e-mails that are spam or non-spam". "This method can be used by the big body to differentiate decent mails that are only the emails they wish to obtain."

REFERENCES

[1] Karim, A., Azam, S., Shanmugam, B., Krishnan, K., & Alazab , M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access,7,168261-168295.[08907831]. https://doi.org/10.1109/ACCESS.2019.2954791

[2] K. Agarwal and T. Kumar, "Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2018, pp. 685-690.

[3] Harisinghaney , Anirudh, Aman Dixit, Saurabh Gupta, and Anuja Arora. "Text and image-based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN algorithm." In Optimization, Reliabilty , and

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 8.102 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 12, Issue 5, May 2023
DOI: 10.17148/IJARCCE.2023.125124

Information Technology (ICROIT), 2014 International Conference on, pp.153-155. IEEE, 2014

[4] Mohamad, Masurah, and Ali Selamat. "An evaluation on the efficiency of hybrid feature selection in spam email classification." In Computer, Communications, and Control Technology (I4CT),

2015 ernational Conference on, pp. 227-231.IEEE,2015

[5] Shradhanjali, Prof. Toran Verma “E-Mail Spam Classification Using SVM and Feature Extraction”in International Jouranl Of Advance Reasearch, Ideas and Innovation In Technology,2017 ISSN: 2454-132X Impact factor: 4.295

2016 W.A, Awad & S.M, ELseuofi. (2011). Machine

2017 Learning Methods for Spam E-Mail

Classification. International Journal of ComputerScience & Information Technology. 3. 10.5121/ijcsit.2011.3112.

2018 A. K. Ameen and B. Kaya, "Spam detection in online social networks by deep learning," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-4.

2019 Diren, D.D., Boran, S., Selvi, I.H., & Hatipoglu, T. (2019). Root

Cause Detection with an Ensemble Machine Learning Approach in the Multivariate Manufacturing Process.

2020 Tasnim Kabir, Abida Sanjana Shemonti, Atif Hasan Rahman. "Notice of Violation of IEEE Publication Principles: Species

Identification Using Partial DNA Sequence: A Machine Learning Approach”, 2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), 2018.