Original Article

Email Spam Detection Using Logistic Regression

Manu Garg[1] , Parveen[2] , Muskan Gupta[3] , Ojasvi[4]

[1]UG Scholar Computer Science & Engineering Meerut Institute of Engineering and Technology, Meerut Meerut, India manugarg1478@gmail.com
[2]UG Scholar Computer Science & Engineering Meerut Institute of Engineering and Technology, Meerut Meerut, India parveenjangid2003@gmail.com
[3]UG Scholar Computer Science & Engineering Meerut Institute of Engineering and Technology, Meerut Meerut, India muskangupta891@gmail.com
[4]UG Scholar Computer Science & Engineering Meerut Institute of Engineering and Technology, Meerut Meerut, India ojassvi24@gmail.com
DOI: 10.47750/pnr.2022.13.S10.245

Abstract

More dependable and powerful anti-spam filters were urgently needed as the quantity of spam, or unwanted email, increased. Spam emails are detected and filtered effectively using contemporary machine learning techniques. We give a thorough analysis of a few well-liked machine learning-based spam filter systems. Our study provides a summary of key theories, experiments, results, and future directions for spam filtering work. In the study's backdrop, an initial study looks at how machine learning techniques are used by top Internet service providers (ISPs) including Gmail and Yahoo junk mail filters to filter email spam. Discussion of the overall spam filtering process shows that several researchers have tried to use machine learning approaches to reduce spam. In this analysis, we contrast the pros and cons of current machine learning for solving spam filtering issues. Keywords – Machine Learning, E-mails, Spam Filtering, research.

I. Introduction

Rapid technological advancement is occurring. Letters were the primary mode of communication for many years before it transitioned to wires and then, more recently, to other formats including messages, calls, SMS, and so on. A person transmits 75 messages each day, such as SMS texts. Every day, around 300 billion communications are sent, most of which are spam. Spam email is basically unsolicited, undesired mails that are distributed widely and clog mailboxes. Most of these communications are links to products that we might purchase and use to consume our own information, or they may be links and connections. On occasion, certain consumers' carelessness can make huge harm. Spam emails affect email traffic in addition to filling your mailbox with useless content. In Walk 2021, spam messages made up 45.1% of incoming messages. To put it simply, such messages may be dangerous and frustrating.

Crucial and important communications are ignored since the mailbox is 85% full of junk. Numerous analysts are developing various methods to find solutions to these problems and ensure correspondence. Important and worthwhile messages are referred to as "ham," just as unwanted emails are referred to as "spam."

Numerous methods have been developed to categorise such spam and junk mail. The combination of natural language processing and machine learning is however one method. It is feasible to categorise emails and create a model that can recognise spam emails by text classification techniques like stemming, lemmatization, svm algorithm, etc.

In this review, we developed a model that would categorise communications as either spam or ham depending on their content. When evaluating the proposed study, execution-related assessment metrics like exactness were taken into consideration. The analysis' findings supported the claim that the suggested study was very precise.

II. Literature Survey

The author of this study [1] authors have discussed the background and provided an explanation of the logistic regression idea. In any case, he provided detailed information on parallel strategic relapse. He also explained other types of calculated relapse, such as paired strategy relapse, multinomial computed relapse, and ordinal calculated relapse. This paper's main goal is to evaluate the interaction between the independent variable's effect and the dependent variables' influence. The dependent/target variable of the author's study, which involved 300 people from Ankara University, was critical thinking.

Author "Liu Lei" demonstrated how this model 'logistic regression' have been used to effectively and swiftly diagnose breast cancer in this research [2]. He used a breast cancer dataset and a logistic regression model. When "Maximum Textured" and "Max Perimeter" were used as the model's inputs, the author achieved the results with the highest accuracy (96.5%). On the other hand, when "mean texture" and "mean radius" were used as inputs for the model, accuracy was 90.48%. As a result, selecting a better set of characteristics will result in more accurate results.

The primary goal of this article [3] is to identify spam texts. To do this, vectors that identify comments as spam or not were created by taking into account ambiguous remarks with increased punctuation, word-stopping segmentations, non-ASCII letters, newlines, capital text, and inflammatory terms. They also included a stop keyword ratio, that is the amount of stop letters divided by total of words in the remark, as spam comments frequently contain repeated terms. This improved categorization precision. For the purpose of removing comments that are unrelated to a certain context, they finally supplied comment similarities and topic similarity. The authors also demonstrated that their method improves the decision tree classifier's performance.

The authors of this study [4] defined Term Frequency Inverse Document Frequency and described the operation of TI-IDF. They also spoke about the TI-advantages IDF's and disadvantages as well as solutions for each. They first gathered data from several domains and eliminated trace keywords from the data, after which they analysed the data using TI-IDF and reported the findings. The findings were given with the terms and their TF-IDF scores for various domains. Parts, presidency, years, and marketing were the most popular keywords from the domains ".business," ".com," ".edu," and ".org."

In order to identify spam communications, the authors[5] of this paper employed Random Forest and Term Frequency Inverse Document Frequency. The UCI Machine Learning Repository was used to get the information. They underwent some pre-processing before utilising TI-IDF, such as eliminating trail words. The authors evaluated many classification algorithms after applying TI-IDF and discovered that Random Forest offers superior accuracy, sharpness, and F-measure in comparison to other classification techniques.

In this paper[6], the writer put up a concept for combining deep neural networks and Tensor Flow to identify spam emails. This model illustrates the benefit of autonomous neural networks using a language approach. This study also looked at a number of publicly accessible datasets and described the model's fundamental structure. They also exposed several open research issues with spam filtering.

Spam filters only function to examine incoming data for undesired (Spam) or desirable content (Ham). Different kinds of filters have been developed by numerous researchers [7]. The model presented in this study makes use of Nave Bayes and natural language processing. To record and monitor spam and ham communications, a database is kept and a Bayesian spam filter is trained. Tokens are used to separate communications, and when a library of tokens has been created using a filter, messages may be evaluated. In order to preserve the spam filter's efficacy, the model additionally adds a threshold counter.

Data is categorised into categories using a variety of spam categorization techniques. [8] 2016's Emmanuel, Gbengadada, and Joseph These categories include artificial neural networks, support vector machines, random decision trees, and probabilistic methods. When used in conjunction with a content-based filtering method that detects particular traits, such classification methods have been demonstrated in the literature to be effective at reducing spam (keywords used inside spam emails). The frequency with which each characteristic appears in emails is used to calculate the likelihood of each feature in an email, which is compared to a threshold number. Email communications that have more recipients than allowed are considered spam.

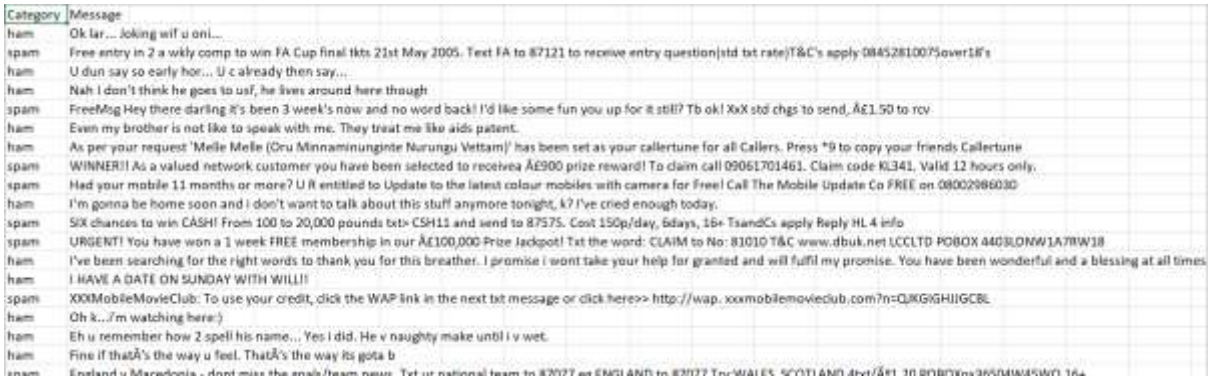
Before using an aggregate method—a voting classifier—the dataset was put through a variety of experiments, including label encode, tokenization, stemming, stop word removal, and feature creation. [9].The model's trials successfully classify the data set in every instance. The algorithms utilised in this investigation produced findings with high accuracy. The Support Vector Classifier, however, outperformed all other tests with an accuracy of

98.49% in predicting spam messages. The accuracy of the other techniques is equivalent, with a variation of about 3% [10][11].

III. Materials and Method

Dataset

Accurate spam detection was the major objective of our study. We used the "Spam Detection Collection Dataset" from Kaggle.com for this. There are 5574 messages in the dataset that include the categories legitimate/Ham or spam. The collection contains 5574 messages, from which 4825 were valid communications and 747 are spam. The texts were gathered from a variety sources. The dataset is in the form of .csv(comma separated values) file. The dataset is then converted into dataframes with the help of panda library present in python. Below is the screenshot of the dataset used [12][13].



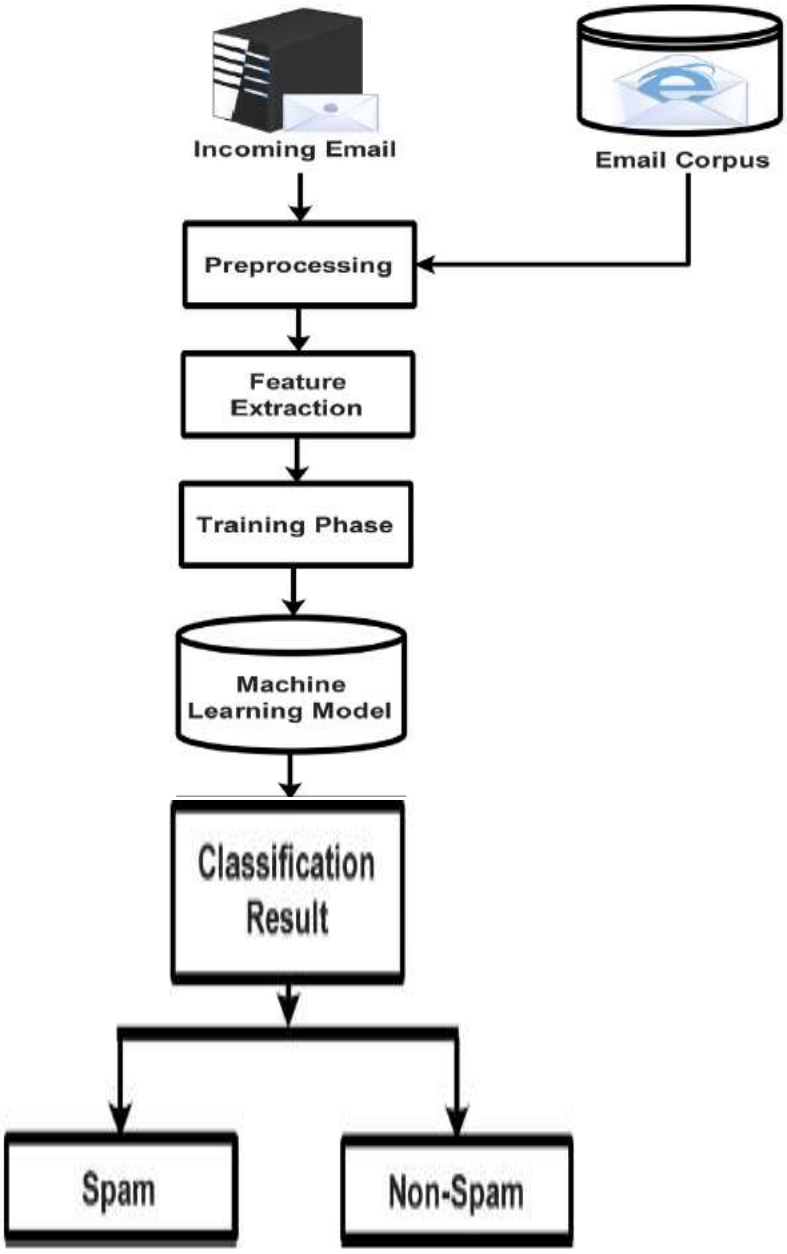
Category	Message
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, Â£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea Â£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
spam	URGENT! You have won a 1 week FREE membership in our Â£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
ham	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. You have been wonderful and a blessing at all times
ham	I HAVE A DATE ON SUNDAY WITH WILL!!
spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
ham	Oh k.../'m watching here:)
ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
ham	Fine if thatÃ's the way u feel. ThatÃ's the way its gota b
spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try WALES, SCOTLAND 4txt/Ã£1.20 POBOXox36504W45WQ 16+

Packages

We acquired a number of files to complete our project. To process the data, it was essential to import the "pandas" package. The "get dummies" method was used to turn the category data into measurement items with the values 0 and 1. When processing tests using the "nltk" package, methods like "stopwords," "porterstemmer," "tfidvectorizer" are developed. The text data was processed using the "re" (Regular Expression Operations) package as well. The "train test split" and "logisticRegression" routines came from the package "sklearn." The data were divided into training and test datasets using the "train test split" function, and the prediction model was created using the "logisticRegression" function [14][15].

Logistic regression is among the most effective and probabilistic methods for categorising communications. In the event of classifying a dataset which has been defined as a spam base, logistic regression is the most adaptable decision-based technique for detecting spam emails from this type of dataset. In order to conduct a fundamental test on a given data distribution, logistic regression finds and computes various statistical ranges, including the standard deviation and mean. Additionally, it generates the useful information for operations like word and character counts, max and min operations. The logistic regression method receives the outcomes of these tests once they are made and delivered, and it has a tendency to correlate the findings. A fundamental tool in statistical analysis, logistic regression seeks to forecast the value of data using a previous distribution and observations. One factors is correlated with one or more other dependent variables using the logistic regression procedure [16][17][18].

Flowchart



IV. PROPOSED ANALYTICAL APPROACH

1. Analysis Statistics

Naïve Bayes Classification Accuracy: 81% === Confusion Matrix === a b <-- classified as 56 5 | a = spam 33 106 | b = ham	***SMO*** Classification Accuracy: 88.5 % === Confusion Matrix === a b <-- classified as 47 14 | a = spam 9 130 | b = ham	***C 4.5*** Classification Accuracy: 90.5 % === Confusion Matrix === a b <-- classified as 57 4 | a = spam 15 124 | b = ham
Bayesian Network Classification Accuracy: 80% === Confusion Matrix === a b <-- classified as 58 3 | a = spam 37 102 | b = ham	***k-NN (k=1)*** Classification Accuracy: 91.25 % === Confusion Matrix === a b <-- classified as 102 15 | a = spam 20 263 | b = ham	***k-NN (k=3)*** Classification Accuracy: 94.5 % === Confusion Matrix === a b <-- classified as 58 3 | a = spam 8 131 | b = ham
Neural Network Classification Accuracy: 93% === Confusion Matrix === a b <-- classified as 56 5 | a = spam 8 131 | b = ham	***Decision Table*** Classification Accuracy: 88 % === Confusion Matrix === a b <-- classified as 45 16 | a = spam 8 131 | b = ham	

2. Data preprocessing

Five columns made up the dataset, three of which were empty, and none of the columns were given the proper names. These three columns were eliminated since they had no use, while the other two were given new names. Because machine learning algorithms perform best when given quantitative data, columns having category "spam/ham" values were transformed into numeric values. The "get dummies" method of the "pandas" package was utilized for this.[19][20][21].

Label encoding is the process of transforming a value's numerical representation to one that is machine readable. Machine learning algorithms can choose how to use these labels after conversion. A basic component of supervised machine learning is this phase.[22].

2.1. Stop Words Removal

The messages (emails in the dataset) must be legible for a machine to comprehend, examine, and perform natural language processing mostly on data. Since computers cannot comprehend human speech, we must pre-process the data in order to make it comprehensible to computers. We must remove the extraneous information from the dataset in order for it to be intact. "Stopwords" are these extraneous words.

Ignore words include phrases like "is," "are," "and," "as," etc. In NLP and even text mining, ignore words are frequently employed to weed out pointless information.

2.2. Stemming

Stemming is the process of getting a word back to its original form, generally by deleting a suffix. The procedure is sped up since the vocabulary space is reduced. It is an additional technique for normalising sentences for machines.

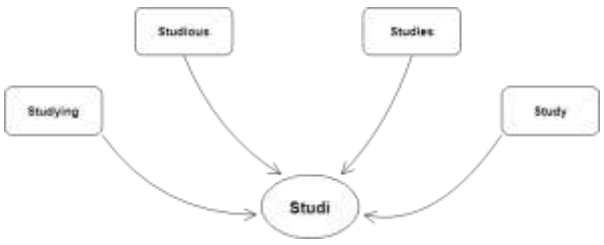


Fig 1 Stemming

2.3.TF-IDF

Due to the fact that the machine learning algorithm just functions on numeric input, we must now transform the text information to vectors. Frequency-inverse document frequency will be used to describe this (TF-IDF).

The word "frequency" is used to determine how often a phrase appears inside a text (TF). It may be computed by dividing a word's occurrence by the total number of words contained in the given document. Imagine that we are attempting to locate the TF for the term "Health," which appears 20 times in a manuscript with 1000 words. As a result, the TF for health in this study will be 0.02. Since certain meaningless words may occur frequently in a paper while still have little importance, the frequency of words alone does not offer a trustworthy indicator. Inverse document frequency (IDF) is used to address this problem since term frequency treats all words alike even if each word has a distinct meaning. IDF aids in lightening the weight of words that are used often across the document collection. The log of a total amount of documents divided by the total number of files containing a certain phrase is used to compute the IDF. The IDF of A and and B will be log(100/10) = 1 and log(100/60) = 0.22 in the provided sequence if, for example, A appears in 10 out of 100 documents and B appears in 60 out of 100 documents. The inverse document frequency is calculated by multiplying the term frequency (TF) by the inverse document frequency.

3. Implementing the algorithm

During cleaning and preprocessing, the "train test split" function can be used to separate the dataset into training and test data. To integrate the data for training in the model and evaluate if the mail is spam or not, we must install the logistic regression method from the "scikit-learn" module in addition to the performance measurements. In this work, we used the logistic regression method to categorise data. Logistic regression is a fantastic method for predictive modelling that forecasts probability for classification problems with a variety of potential outcomes. Logistic regression, which is similar to linear regression, generates an S-shaped line with an output of either 0 or 1. This S-shaped curve is created via logistic regression using a sigmoid function. Logistic regression in the model will tell us if the email is spam or ham. If the value is 1, it would be spam; if it is 0, it would be ham.

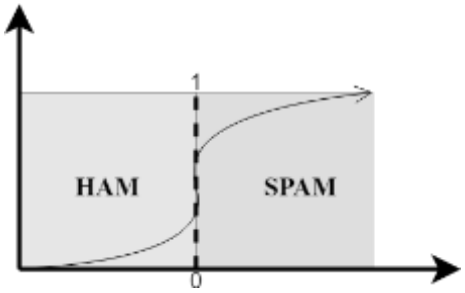


Fig 2 Logistic Regression

V. RESULTS

```
input_mail = ["We are glad to share this exciting opportunity with your esteemed campus."]

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

# making prediction

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


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

else:
  print('Spam mail')
```

```
[1]
Ham mail
```

As we can see in the above image we have inputted a testing mail which is “We are glad to share this exciting opportunity with your esteemed campus”. We are putting the mail in the trained model. If the output is ‘1’ that means it is a Ham mail and if the output is ‘0’ that means it is a spam mail. As we can see the output of the trained model is ‘1’ which is ham mail. The trained model is predicting the mail correctly.

VI. CONCLUSION

In this, we took into account the broad uses of NLP for spam identification. We also discussed the meticulous process the algorithm uses to categorise communications as spam or junk mail. Performance measurements were included to assess the model's accuracy. In the coming, we would be able to use neural networks and supervised neural algorithms to determine if a specific message is spam or not. In natural language processing, deep learning surpasses other traditional machine learning methods, but it needs a large amount of dataset to provide accurate results. The suggested spam detection and email filtering system can be further enhanced in the domain of internet security because natural language processing is still a relatively unexplored research area.

VII. References

[1] Sjarif, Nila, & Amir, N. (2019). SMS Spam Message Detection using Term Frequency-Inverse Document Frequency and Random Forest Algorithm. Procedia Computer Science , 509-515.

[2] Narayan, Vipul, et al. "To Implement a Web Page using Thread in Java." (2017).

[3] Srivastava, Swapnita, and P. K. Singh. "HCIP: Hybrid Short Long History Table-based Cache Instruction Prefetcher." International Journal of Next-Generation Computing 13.3 (2022).

[4] Srivastava, Swapnita, and P. K. Singh. "Proof of Optimality based on Greedy Algorithm for Offline Cache Replacement Algorithm." International Journal of Next-Generation Computing 13.3 (2022).

[5] Smiti, Puja, Swapnita Srivastava, and Nitin Rakesh. "Video and audio streaming issues in multimedia application." 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2018.

[6] Qaiser, Shahzad, & Ali, R. (2018). Text Mining: Useof TF-IDF to Examine the Relevance of Words toDocuments. International Journal of ComputerApplications , 25-29.

[7] Awasthi, Shashank, et al. "A Comparative Study of Various CAPTCHA Methods for Securing Web Pages." 2019 International Conference on Automation, Computational and Technology Management (ICACTM). IEEE, 2019.

[8] Narayan, Vipul, and A. K. Daniel. "FBCHS: Fuzzy Based Cluster Head Selection Protocol to Enhance Network Lifetime of WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.3 (2022): 285-307.

[9] Narayan, Vipul, et al. "E-Commerce recommendation method based on collaborative filtering technology." International Journal of Current Engineering and Technology 7.3 (2017): 974-982.

[10] Srivastava, Swapnita, and Shilpi Sharma. "Analysis of cyber related issues by implementing data mining Algorithm." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019.

[11] Narayan, Vipul, and A. K. Daniel. "Multi-tier cluster based smart farming using wireless sensor network." 2020 5th international conference on computing, communication and security (ICCCS). IEEE, 2020
[12] Pandey, S., & Yadav, R. (2020). Email Spam Detection using Machine Learning and Deep Learning.IJRASET .
[13] Narayan, Vipul, and A. K. Daniel. "Design consideration and issues in wireless sensor network deployment." (2020): 101-109.
[14] Choudhary, Shubham, et al. "Fuzzy approach-based stable energy-efficient AODV routing protocol in mobile ad hoc networks." Software Defined Networking for Ad Hoc Networks. Cham: Springer International Publishing, 2022. 125-139.
[15] Narayan, Vipul, and A. K. Daniel. "RBCHS: Region-based cluster head selection protocol in wireless sensor network." Proceedings of Integrated Intelligence Enable Networks and Computing: IIENC 2020. Springer Singapore, 2021.
[16] Narayan, Vipul, and A. K. Daniel. "CHOP: Maximum coverage optimization and resolve hole healing problem using sleep and wake-up technique for WSN." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal 11.2 (2022): 159-178.
[17] Lei, L. (2018). Research on Logistic Regression Algorithm of Breast Cancer Diagnose Data by Machine Learning. ICRIS, (pp. 3-4).
[18] Emmanuel, Gbengadada, & Joseph. (2016). Machine learning for email spam filtering: review, approaches and open research problems. Heliyon
[19] Narayan, Vipul, and A. K. Daniel. "CHHP: coverage optimization and hole healing protocol using sleep and wake-up concept for wireless sensor network." International Journal of System Assurance Engineering and Management 13.Suppl 1 (2022): 546-556.
[20] Narayan, Vipul, and A. K. Daniel. "IOT based sensor monitoring system for smart complex and shopping malls." Mobile Networks and Management: 11th EAI International Conference, MONAMI 2021, Virtual Event, October 27-29, 2021, Proceedings. Cham: Springer International Publishing, 2022.
[21] Narayan, Vipul, and A. K. Daniel. "Energy Efficient Protocol for Lifetime Prediction of Wireless Sensor Network using Multivariate Polynomial Regression Model." Journal of Scientific & Industrial Research 81.12 (2022): 1297-1309.
[22] Narayan, Vipul, A. K. Daniel, and Ashok Kumar Rai. "Energy efficient two tier cluster based protocol for wireless sensor network." 2020 international conference on electrical and electronics engineering (ICE3). IEEE, 2020.