A Novel Spam Classification System for E-Mail Using a Gradient Fuzzy Guideline-Based Spam Classifier (GFGSC)

Vinoth Narayanan Arumugam Subramaniam
Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies (VISTAS), India
vinoth.nas89@gmail.com

Rajesh Annamalai
Department of Computer Science and Engineering, Vels Institute of Science, Technology and Advanced Studies (VISTAS), India
arajesh.se@velsuniv.ac.in

Abstract: *Spam messages have increased dramatically in recent years even as the number of email clients has grown. Email has already become a valuable way of communicating because it saves time and effort. However, numerous emails contain unwelcome content known as spam as a result of social platforms and advertisements. Despite the fact that many techniques have already been created for spam mails categorization, none of them achieves 100 percent efficiency in analyzing spam messages. So, in this research, we propose a novel Gradient Fuzzy Guideline-based Spam Classifier (GFGSC) for classifying the spam e-mails as spam or non-spam. This research uses four types of datasets and these datasets are pre-processed using normalization. Then the set of data can be extracted using Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) techniques. The aspects are selected using Information Gain (IG) and Chi-Square (ChS) techniques. And the GFGSC classifier can be used for classifying the data as spam or non-spam with better effectiveness. Finally, the performances are examined and these metrics are matched with the existing approaches. The results are obtained using the MATLAB tool.*

Keywords: *Spam e-mail, principal component analysis, latent semantic analysis, information gain, chi-square, gradient fuzzy guideline-based spam classifier, MATLAB tool.*

Received December 28, 2021; accepted September 8, 2022
https://doi.org/10.34028/iajit/20/3/12

1. Introduction

E-mail is indeed a convenient and rapid way to receive messages from anywhere in the globe and it may be used with desktops, cell phones, and some other next digital equipment [3]. Although the rise in popularity of other kinds of digital interaction like text messaging and networking websites, mails still remain the dominant mode of corporate communication and are still required for other types of interactions and payments [1]. Because humans are social creatures, they are often linked to the social circle. Because this is a digitalization period, Texts and e-mails are demonstrating to be among the most effective means for knowledge transfer. However, as these methods of information exchange get more popular, so does the speed at which spam becomes more prevalent. Spam may come via anywhere on the planet in which there is web service [5]. Spam is generating issues not just for consumers, but for businesses. Numerous anti-spam models have been developed for identifying such spam, but none of them are as effective as ham and spam clustering [10]. Several online services, like Google, Hotmail, and others, are now deploying anti-spam technologies to identify email spam so that customers never become victims of these messages, yet many consumers keep falling victim [21].

The method of extracting features could play a role inside the categorization procedure's improvement. Spam emails are categorized as such since they provide no value to the recipient. Promotion of mainly unlawful, non-existent, or useless things, advocacy of a message, as bait for fraudulent transactions, or transmission of viruses are the major reasons for the proliferation of spam messages. There are several eradication methods suggested by many authors such as machine learning, deep learning and other neural networks [23]. Also, the Bag of Words (BoW) was used to depict the meta-data framework of a spam file, however the sequence of term dependency inside the file is neglected, but only the phrase word count is regarded [18].

Inside this setting, clearing or obscuring explicit data is a difficult task that could seriously affect the spam filter successfulness. To reduce computation time and enhance accuracy, the vast bulk of information retrieval research employs different methods that interact with high-dimensionality. Term Strength (TS), Document Frequency (DF), and Mutual Information (MI) seem to be some cases of feature selection techniques that were widely used to recognize more substantial features that allow the classifier to classify spam messages. These tools, on the other hand, keep performing well in practice when it comes to email spam categorization [4].

Here, we propose the Gradient Fuzzy Guideline-based Spam Classifier (GFGSC) for classifying the e-mail spam as good or virus. Spam Base, Ling-spam, Spam Assassin, and Enron are utilized as datasets in this work. These data are pre-processed using the normalization method and the normalized data are extracted using the Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA) techniques. Then these datasets are selected using the IG and ChS tools. Finally, the datasets are classified by using the proposed classifier. The following is how the entire paper is organized: Topic 2 discusses related work and problem statements, while topic 3 illustrates the proposed approach. The performance examination is presented in topic 4. Lastly, topic 5 reports the conclusion.

2. Related Works

The presented method integrates the K-Nearest Neighbor (KNN) algorithm with the Support Vector Machine (SVM) algorithm to identify webpages as Malware, Valid, or Mysterious, as described in [2]. The method combines the strength of SVM with the efficacy and clarity of KNN. As a result, the suggested KNNSVM gets the benefits of incorporating KNN and SVM while avoiding their respective disadvantages if used individually. The goal of [6] is to develop a Stepsize-Cuckoo Search (SCS) as well as SVM-based method for spam mails identification. The SCS method is utilized to determine the optimal set of properties. SCS is utilized to identify the perfect collection of features, and then the SVM is utilized to classify spam. They are using 3 distinct kernels to improve categorization success: linear, polynomial, and quadratic. The focus of [9] is to use feature selection to facilitate the identification of malignant spam. Researchers suggest a framework that takes a novel set of data for selecting features, which is a process toward classification purpose in the future. The use of characteristics should reduce time for training as well as enhance the consistency of malignant spam filtering. At the tweet stage, Gibson *et al.* [8] suggest an ensemble method for spam filtering. On the basis of Convolutional Neural Networks (CNN), we exhibit a range of deep learning techniques. [15] In the ensemble, 5 CNN models and one feature-based prototype have been used. To train a model, every CNN employs a separate set of embeddings. Content-based, user-dependent, and n-gram characteristics are used in the feature-reliant prototype [22]. This technique uses a multi-layer neural network as a meta-classifier to incorporate deep learning as well as conventional feature-based designs. The methodology is evaluated on different datasets. Ghaleb *et al.* [7] Propose a strategy for dealing with spam scams. Semi-Automated Feature generation for Phish Classification (SAFE-PC) is a scheme they evolved to recognize new malware campaigns that have progressed from previous ones. SAFEPC actually uses spam email as well as valid email address sets of data from such a tier-1 study institution's central IT organization, totaling 425K spam and 158K valid messages. Rastenis *et al.* [19] they extracts information from every text's header and body, infused with knowledge of phishing frameworks. The RUS Boost classification model is then applied to the spam as well as valid email messages. Li *et al.* [14], propose a multi-view disagreement-reliant semi-supervised learning method for e-mail classification. The concept of multiple views can provide more data for categorization, which is almost always overlooked in the literature. Semi-supervised learning could be used to make use of supervised and unsupervised data [13]. Gibson *et al.* [8], propose a solution that incorporates numerous face detection, text retrieval, and language processing methods and modifies them using innovative approaches to classify images as essential or spam basis of user habits and preferences. That model divides the pictures on the users' the Smartphone into the separate types and then processes them correspondingly. A novel Feature-centric Spam Email Detection Model (FSEDM) is introduced in [8]. Material, impression, conceptual, user, and junk mail vocabulary are all components of the project management set of features. The CSDMC2010 Spam data was used to generate the proposed features. Testing was carried out in detail to implement the research framework. The benefits of effectiveness assessment methods show that sentiment characteristics were crucial in spam messages classification. Gibson *et al.* [8] offers a method depending on automated identification of email body texts into malware as well as spam scams. The preferred framework is utilized to categorize emails in three languages:

1. English.
2. Russian.
3. Lithuanian.

And they investigate the appropriateness of a computer-controlled set of data transcription to adjust it to email categorization signed in other texts, since most public email sets of data nearly solely gather English email messages. The purpose of [17] is to demonstrate how well an adaptive smart learning strategy relying on a visual anti-spam prototype for multi-natural language could be utilized to successfully identify anomalous circumstances. This strategy is used for phishing detection. The purpose of [12] is really to lessen the quantity of junk mail by detecting it with a classification model. The Machine Learning algorithms could be used to accomplish most precise spam categorization [16]. To analyze the message of an e-mail in order to locate spam, a language processing method was employed [20]. Jain *et al*. [11] aims to acquire emails via third-party APIs and effectively processing them for machine learning. On the suggested paradigm, many machine learning algorithms, both supervised and unsupervised,

could be trained and tested. For SMS spam categorization, Sah and Parmar [21] suggested a SVM algorithm. To accomplish spam categorization and determine the accuracy of classification.

- **Problem Statement**: Despite the numerous advantages of email, its use is hampered by the huge count of unsolicited and often spam emails that must be recognized and isolated as soon as possible using a spam detection system. Spam identification is critical for protecting email users and preventing a number of recent undesirable uses to which emails have been put. Unfortunately, the dynamic behavior of spam mails by the use of mailing techniques has restricted and mostly proved spam identification techniques ineffectual, forcing the creation of innovative spam detection techniques to obtain greater spam detection performance. In the research, many spam detection techniques have been presented and assessed; nonetheless, the reported accuracy shows that more research in this area is still required to enhance accuracy.

3. Proposed Work

In this study, we present the GFGSC is employed for accomplishing the better accurate classification of spam as good or malware. The dataset of Spam Base, Ling-spam, Spam Assassin, and Enron are initially pre-processed using the normalization method. Then these datasets are extracted with the PCA and LSA techniques and also these data are further selected using the IG and ChS approaches. Finally, the selected dataset is categorized using the proposed technique. Figure 1 depicts the proposed flow of this study.

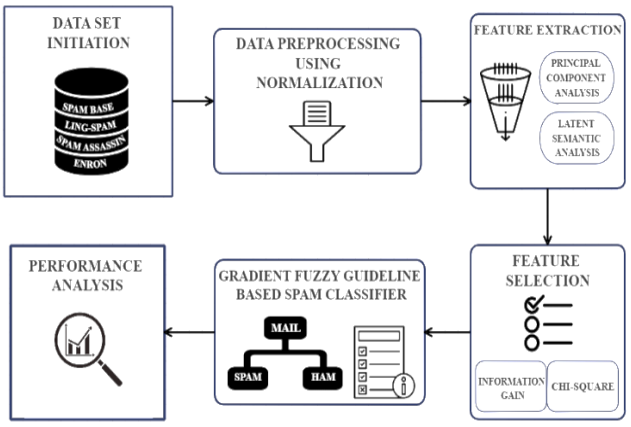

Figure 1. Proposed system using GFGSC classification framework.

3.1. Dataset Initialization

Four datasets namely Spam base, Ling-spam, Spam assassin, and Enron are considered for this study. The details related to the dataset are given in Table 1. All four datasets contain instances under two classes namely SPAM and HAM (non-spam).

Table 1. Dataset description.

Dataset	Total Instances	SPAM	HAM	Year
Spam base	4601	1813	2788	1999
Ling-Spam	2893	481	2412	2000
Spam Assassin	6047	1897	4150	2002
Enron	36715	20170	16545	2006

3.2. Dataset Preprocessing Using Normalization

Data preprocessing is a data-mining approach in which a series of procedures convert the raw e-mail dataset into a type, which shall be understood. It is responsible for preparing raw data for future processing. Data cleaning, data reduction, data transformation, data integration, and data discretization are few of the processes involved in data preprocessing. The process of converting values measured on a distinct scale to a common scale is known as normalization.

The method for data processing necessitates the normalization of calculated data on a separate balancing to a conceptually shared scale, which is commonly done before averaging. To acquire values linked to a distinct parameter, particular methods of normalization necessitate a rescaling step.

$$\hat{\sigma}^2{}_i = \frac{1}{l-k-1}\sum_{j=1_{j\neq 1}}^{l} \bar{\in}^2{}_j \tag{1}$$

Where k denotes the specification and σ represents the Standard Deviation (SD).

Next, the errors must not depend on each other. It is expressed as shown below,

$$m_i \sim \sqrt{o}\frac{T}{\sqrt{t^2+o-1}} \tag{2}$$

Where m is a random variable

After that, the motion of the specification requires to be normalized by the SD.

$$\mathrm{N}=\frac{\mu^n}{o^n} \tag{3}$$

Where n is the moment scale.

$$\mu^n = S(X-\mu)\hat{}N \tag{4}$$

Here X denotes a random parameter and s denotes the intended value

$$o^n = (\sqrt{s(X-\mu)\hat{}N\ )\hat{}2} \tag{5}$$

For standardizing the distribution of the parameter utilizing the mean, μ particularly for the normal orderly distribution.

$$Co_v = \frac{S}{\bar{X}} \tag{6}$$

Where C_{ov} is the coefficient of the variance

Then the function scaling procedure can be carried out to put in all values between 0 and 1. This method is called standardization, depending on the application.

$$X' = \frac{(X-X_{min})}{(X_{max}-X_{min})} \tag{7}$$

The preprocessed dataset contains 235 attributes that includes sender, receiver, Blind Carbon Copy (BCC), date of sending, receiving, number of receivers, etc. The preprocessed data is fed into the extractor for extracting the relevant features.

3.3. Feature Extraction

The feature extraction stage is a data pre-processing step that identifies features that could be used to define an item. Spam data extracted from datasets is stale data that can be cleaned up to provide more detailed information.

3.3.1. Principal Component Analysis (PCA)

The PCA methodology is a statistical technique that converts a collection of measurements of potentially correlated parameters into a collection of exponentially uncorrelated variables called principal components utilizing an orthogonal transformation. The count of original variables will be less than or equal to the number of primary components. The initial principal component has the maximum possible variance (that is, it accounts for as much variation in the data as feasible), and each subsequent element or component has the highest variance conceivable under the restriction of being orthogonal to the preceding components. The generated vectors constitute an orthogonal uncorrelated basis set. The eigenvectors of the symmetric covariance matrix are the primary components. A linear space is created by combining such orthogonal vectors into a matrix. The PCA is a valuable tool for analysis in high-dimensional spaces since each original data instance vector can be represented by a lower number of variables.

PCA is a method for weighing attributes that are used to choose clustering attributes. The grouping of e-mail spam data in accordance with the attribute equation possessed by each profile is the result of this study. Prior to clustering, PCA was performed to minimize dimensions and maximize the clustering results. PCA is used to rank features based on their relationship to other qualities. We can calculate the weight of each feature by using PCA to rank the features. If N is a matrix with rows corresponding to a point in space, we can compute N^TN and eigen pairs for that point. D, the matrix, with the columns acting as eigen vectors and the biggest eigen value coming first. Let K be a matrix with the N^TN eigen values along the diagonal, with the greatest value first and 0's in the other entries. Then, though $N^TNd=\lambda d =d\lambda$ for every eigen vector d and its related eigen value λ, it is understandable that:

$$N^T ND = DK \tag{8}$$

The points of N have been transformed into another coordinate space, in which the first axis, which corresponds to the biggest eigen value, is crucial. The axis with the most variance has the most points. In a similar fashion, the second axis, which is related to the second eigen pair, is the next notable axis, and this pattern continues for all eigen pairs. If it is desirable to translate N into a space with less dimensions, the most essential choice employs the eigen vectors associated with the highest eigen values and ignores the remaining eigen values, i.e., if D_j is the first j columns of D, then ND_j is the j-dimensional representation of N.

3.3.2. Latent Semantic Analysis (LSA)

LSA is a method for assessing messages to uncover hidden meaning. LSA creates word vectors, which are then used to map words to concepts. The following is the procedure for creating these word vectors. Initially, a large matrix with e-mail documents as columns and index words as rows is constructed. Every cell in this matrix reflects the count of times the word occurs in a specific e-mail document. A word that appears in more than two e-mail documents and does not belong to stop words is called an index word. The TF-IDF algorithm is utilized to carry out the next stage in LSA, which is weighting. Term frequency-inverse document frequency is abbreviated as TF-IDF. It is employed to demonstrate the value of a term in a corpus. Vector operations are used to provide fewer common terms and more weight. The count in each cell of the raw matrix is substituted by the subsequent equation in TF-IDF:

$$TFIDF_{a,b} = \left(N_{a,b}/N_{*,j}\right) * log(E/E_i) \tag{9}$$

Where,
$N_{a,b}$- Actual cell count
$N_{*,j}$ - Count of overall words in document j
E- Overall count of e-mail documents
E_i - The count of e-mail documents in which word i occurs

From the 235 attributes, the feature extractor has extracted 120 attributes containing the sender, receiver, BCC, date of sending, receiving, and number of receivers.

3.4. Feature Selection

The extracted features are given as input to the feature selection process. It is a method for improving the performance of machine learning techniques and applications by deleting non-relevant and repeating characteristics from a data collection. Feature selection has improved the performance of data mining and machine learning approaches by dealing with the curse of dimensionality. One of the most essential data mining techniques in preprocessing is feature selection, which is used to choose a large number of features from a dataset. Its goal is to decrease data, which will speed up computing operations and result in more accurate models of the methods utilized. Feature selection is commonly used to choose the best features, minimize dimensions, increase algorithm accuracy, and eliminate unnecessary features.

3.4.1. Information Gain (IG)

Mutual information gain is another term for information gain. It is the information that can be used to calculate the mutual dependencies of two variables. Information gain is a technique for determining how much relevant information can be extracted from a random variable by combining it with another variable. In other words, information gain is a symmetrical measure of dependency.

Information Gain formula defined as:

$$IG\ (A;B) = H(A) - H(A|B) \tag{10}$$

3.4.2. Chi-Square (ChS)

Whenever the feature events are unconnected to the categorical variable, the Chi-squared testing is used as a statistical tool for determining departures from the anticipated distributions. The chi square value is calculated using true positives (t_p), false positives (f_p), true negatives (t_n), false negatives (f_n), likelihood of number of positive instances P_{po}, and likelihood of number of negative instances P_{ne}.

$$chi - square\ metric = T\left(t_{p,}\left(t_p + f_p\right)P_{po}\right) + T\left(f_{n,}(t_n + f_n)P_{po}\right) + T\left(f_{p,}\left(t_p + f_p\right)P_{ne}\right) + T\left(t_{n,}(t_n + f_n)P_{ne}\right) \tag{11}$$

Where T (count, expect)=(count–expect)^2/expect

The stages in the chi-square method are as follows:

1. State the hypothesis.
2. Create an analysis strategy
3. Analyze data from a sample
4. Draw conclusions

The assessment strategy outlines how to use model data to accept or reject the hypothesis when the hypothesis is stated. The following must be included in the plan:

1. Significance rank: investigators use significance levels of 0.01, 0.05, or 0.10, but any number between 0 and 1 can be used.
2. Test method: the chi-square test is done to evaluate if there is a significant association between two categorical features by determining their interdependence level.

To determine the degrees of freedom, predictable frequencies, test values, and P-value affiliated with the tests, the sample data must be evaluated.

$$DoF = (g - 1) * (c - 1) \tag{12}$$

Where *DoF* stands for degrees of freedom, *g* stands for one categorical variable's number of levels, and *c* for another categorical variable's number of levels.

$$X^2(f,c) = \left[\frac{N*(PS-RQ)^2}{(P+R)(Q+S)(P+Q)(R+S)}\right] \tag{13}$$

Where P=Number of times feature ‘t’ and class label ‘c’ co-exists.
Q=Number of times‘t’ occurs without ‘c’
R=Number of times ‘c’ occurs without‘t’.
S=Number of times neither ‘c’ nor‘t’ occurs.
N=Overall count of records.

Using the feature selection method, the best attributes that are used for classifying the spam and normal mails are selected. Table 2 shows the attributes selected for the classification of spam and normal emails. The feature selection technique has selected 60 attributes for the classification purpose.

Table 2. Features selected for classification purpose.

S.No	Selected features	No. of attributes
1	word_freq_WORD	47
2	char_freq_CHAR	5
3	capital_run_length_average	2
4	capital_run_length_longest	2
5	capital_run_length_total	2
6	spam	2
	Total number of attributes	**60**

3.5. Classification using Gradient Fuzzy Guideline-based Spam Classifier (GFGSC)

This classification approach creates a set of fuzzy rules based on content-related factors such as word count and unique word proportion. Two threshold values are defined in these guidelines: one for the number of words and one for the proportion of unique words. The number of words in each email is counted, and the proportion of unique words is calculated and compared to the threshold value (Set threshold value=5). The words are assigned ranks depending on this threshold value. Words with uncertainties are ranked as one, two, three, etc., based on the proportion of the uncertainty. The proposed GFGSC algorithm will split the email messages according to the rank of the values calculated. Accordingly, they are classified into three sets namely: high risk, moderate risk and low risk mails. The email containing words with rank 1 is considered as high-risk email, rank 2 as moderate risk email, and rank 3 as low risk email. Based on these comparisons, the email is classed as either spam or legitimate. Equation (14) depicts how a message is classified as spam or legitimate.

$$E_L = \begin{cases} Truthful, & if\ c_w < T_{c_w}\ and\ P_{uw} > T_{P_{uw}} \\ Spam, & if\ c_w \geq T_{c_w}\ and\ P_{uw} \leq T_{P_{uw}} \end{cases} \tag{14}$$

Where,

E_L: Review label
c_w: Words count
P_{uw}: Unique words percentage
T_{c_w}: Words count threshold value
$T_{P_{uw}}$: Unique words percentage threshold value

The work of selecting guidelines for spam detection is critical. Chosen guidelines must be linked to the message type in order to improve the precision of spam email identification. In most cases, a guideline is stated in terms of an IF condition THEN action. It signifies that if the condition is met, the action of that particular

guideline is carried out. The following are the guidelines that are employed in the suggested technique:

- Guideline 1: IF there is a URL in the email, THEN it is most likely spam. Because hackers may deceive consumers by delivering a URL link in a text or email that when accessed could lead the users to a fraudulent login screen or download virus to the user's mobile phone, a URL analyzer examines for the existence of a URL in the text or email.
- Guideline 2: IF the email includes any algebraic expressions such as +, -, >, /, and so on, THEN it is most likely spam.
- Guideline 3: IF the communication includes any currency symbols such as "$," "£," and so on, THEN it is most likely spam. In the fraudulent reward messages, for instance, the sign "$" is utilized to symbolize cash. We chose two symbols that usually appear in spam messages: $ (Dollar) and £ (Pound).
- Guideline 4: IF there is a Phone number in the e-mail, THEN it is most likely a spam message. The hacker requests that consumers transfer their personal information, including bank account information, to a specific phone number.
- Guideline 5: Free, accidents, rewards, dating, awarded, services, lotteries, minutes, visiting, supply, money, claim, award, delivery, and other dubious keywords are deemed spam words. IF any of the suspected keywords appear in the email, THEN it is almost certainly spam.
- Guideline 6: IF the message size exceeds 150 characters, THEN they may be spam email. This includes spaces, symbols, special symbols, smiley faces, and other elements.
- Guideline 7: IF the message is self-addressed, THEN it is most certainly a spam message. Self-answering SMS prompts the user to enroll to or unsubscribe from any services.
- Guideline 8: IF the email comprises visual morphemes, it is most likely spam. Visual morphemes are numbers and other symbols used in texts, emails, and other forms of communication.
- Guideline 9: IF the email address is included in the message, THEN it is most likely a spam message. The hacker also obtains private data from the target source by using the email address in the message.

Algorithm 1: GFGSC

Input datasets: Spam base, Ling spam, Spam assassin, Enron
Output: Spam mail, Normal mail
Start
Read the email
Apply fuzzy set of guidelines
Set threshold value = 5
If uncertainty > threshold value
Then assign rank as 1
If uncertainty = threshold value
Then assign rank as 2
If uncertainty < threshold value
Then assign rank as 3
If uncertainty = 0;
Then assign rank as 0
Repeat ham
For every word
If rank 1
Return "High risk email"
If rank 2
Return "Moderate risk email"
If rank 3
Return "Low risk email"
If rank 0
Return "Normal email"
End

4. Performance Analysis

The suggested model is tested utilizing four datasets and the outcomes are analyzed utilizing the MATLAB simulation tool. The suggested spam detection technique's effectiveness is evaluated using a variety of performance metrics. Accuracy, sensitivity, specificity, precision, and F1 Score are among them.

Accuracy refers to a system's ability to properly distinguish between spam and legitimate emails. It is calculated by dividing the fraction of true positive and true negative samples in all analyzed cases by the total number of cases.

$$Accuracy = \frac{t_p+t_n}{t_p+f_p+t_n+f_n} \quad (15)$$

The fraction of spam emails accurately recognized is measured by sensitivity. It demonstrates how effective a method is at identifying spam emails.

$$Sensitivity = \frac{t_p}{t_p+f_n} \quad (16)$$

The fraction of legitimate emails successfully recognized is measured by specificity, which indicates how effective a method is at eliminating false alarms.

$$Specificity = \frac{t_n}{t_n+f_p} \quad (17)$$

Precision is a metric that evaluates how many email messages are successfully predicted.

$$Precision = \frac{t_p}{t_p+f_p} \quad (18)$$

The weighted average of sensitivity and precision is the F1 score.

$$F1\ score = \frac{2t_p}{2t_p+f_p+f_n} \quad (19)$$

The email spam classification results of the GFGSC technique are examined on four datasets in Tables 3, 4, and Figure 2. The experimental results showcased that the GFGSC technique has gained effective outcomes on all the applied datasets. For instance, on the Spambase dataset, the GFGSC technique has offered an accuracy of 0.93, sensitivity of 0.912, specificity of 0.904, precision of 0.92, and F1-score of 0.909. Besides, on the Ling-spam dataset, the GFGSC approach has presented an accuracy of 0.943, sensitivity of 0.943, specificity of 0.95, precision of 0.937, and F1-score of 0.95.

Additionally, on the Spam Assassin dataset, the GFGSC method has an accuracy of 0.99, sensitivity of 0.988, specificity of 0.991, precision of 0.98, and F1-score of 0.997. Lastly, on the Enron dataset, the GFGSC methodology has offered an accuracy of 0.98, sensitivity of 0.983, specificity of 0.985, precision of 0.979, and F1-score of 0.988.

Table 3. Results analysis with sensitivity, specificity, precision and F1 score of proposed GFGSC model on applied dataset.

Datasets	Metrics			
	Sensitivity	Specificity	Precision	F1-Score
Spam base	0.912	0.904	0.92	0.909
Ling-spam	0.943	0.95	0.937	0.95
Spam Assassin	0.988	0.991	0.98	0.997
Enron	0.983	0.985	0.979	0.988

Table 4. Results analysis with accuracy of proposed GFGSC model on applied dataset.

Data Sets	Accuracy
Spam base	0.93
Ling-spam	0.943
Spam Assassin	0.99
Enron	0.98

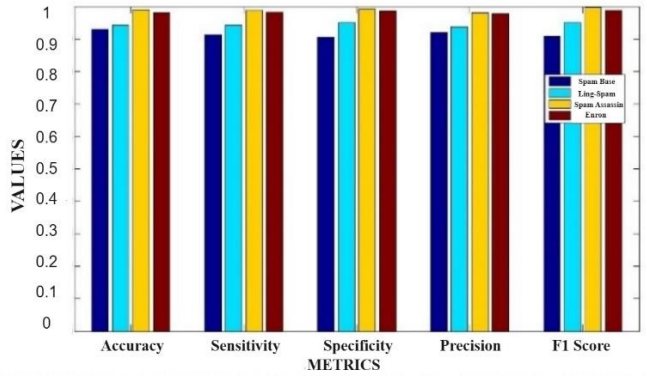

Figure 2. Result analysis of GFGSC Model with different measures.

For exhibiting the enhanced performance of the GFGSC technique, a brief comparative study is made. Figures 3, 4, 5, and 6 show the comparative analysis of all the metrics for existing and proposed methods using the Spam base, Ling-spam, Spam Assassin, and Enron datasets respectively. It is evident from the graphs that the proposed classifier outperforms the traditional methods.

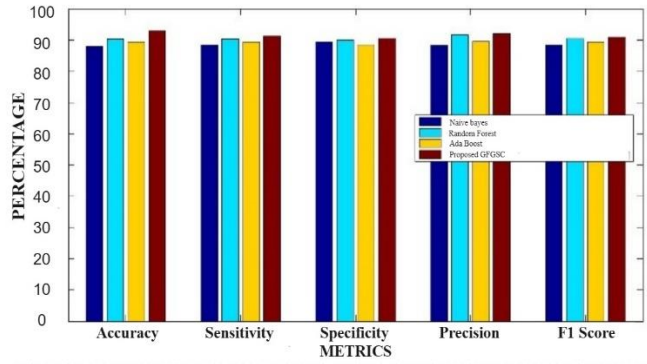

Figure 3. Comparative analysis of existing and proposed methods for spam base dataset.

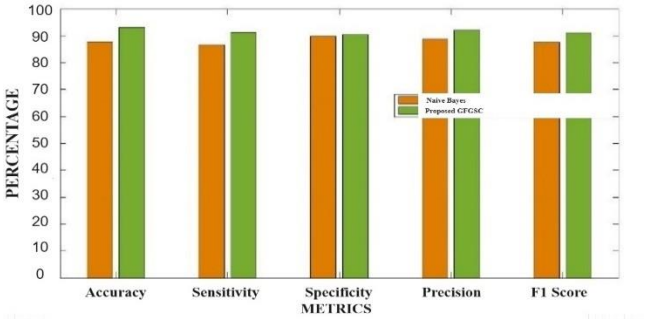

Figure 4. Comparative analysis of existing and proposed methods for Ling-spam dataset.

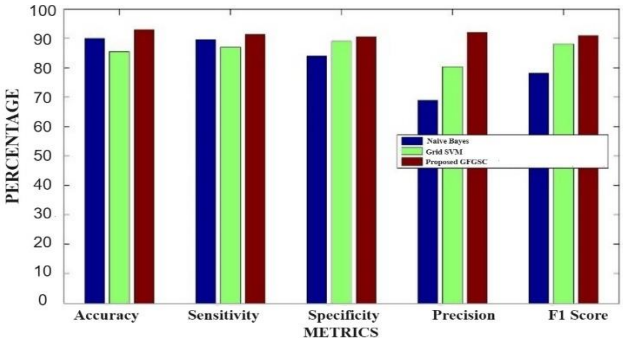

Figure 5. Comparative analysis of existing and proposed methods for spam assassin dataset.

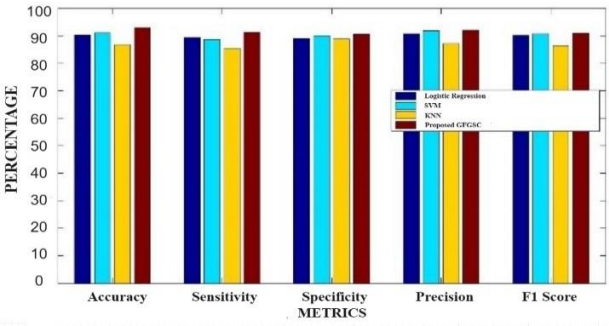

Figure 6. Comparative analysis of existing and proposed methods for Enron dataset.

Also, the computation time in seconds is estimated and compared with the traditional classification algorithms. The timescale necessary to complete a computational process is known as computation time (sometimes known as "execution times"). The computation time is proportionate to the count of rule applications when a computation is represented as a series of rule applications.

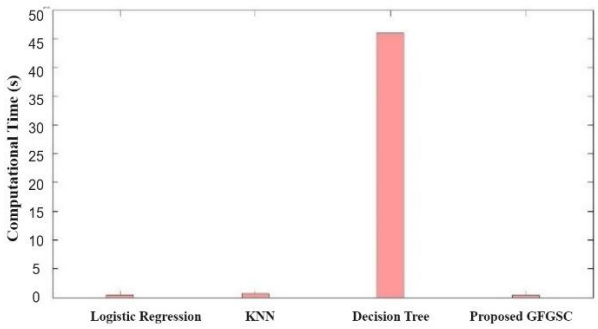

Figure 7. Comparison of computation time (s) for existing and proposed GFGSC Method.

Figure 7 and Table 5 shows the comparison of computation time for the proposed method with the existing methods like Logistic Regression (LR), (KNN), and Decision Tree (DT). The graph clearly indicates that the computation time is less for the proposed system. As a consequence of the experiments, GFGSC appears to be the best classifier for correctly classifying spam.

Table 5. Computational Time Performance analysis of proposed GFGSC method with existing methods.

S.NO	Technique	Computational Time (in Seconds)
1.	Linear Regression	0.6
2.	K-NN	1
3.	Decision Tree	46
4.	GFGSC(Proposed)	0.5

5. Conclusions

In this study, a new email spam detection and classification model has been developed by the GFGSC technique. The proposed GFGSC technique encompasses different processes namely pre-processing, PCA and LSA based feature extraction, IG and ChS based feature selection, and GFGSC based classification. A comprehensive simulation analysis is carried out to point out the superior outcomes of the GFGSC technique. The simulation values demonstrated the betterment of the GFGSC technique over the other state of art techniques. In future, the clustering and outlier detection approaches can be designed to improve the email spam filtering outcomes.

References

[1] Abdulhamid S., Shuaib M., Osho O., Ismaila I., and Alhassan J., "Comparative Analysis of Classification Algorithms for Email Spam Detection," *International Journal of Computer Network and Information Security*, vol. 10, no. 1, pp. 60-67, 2018.

[2] Altaher A., "Phishing Websites Classification Using Hybrid SVM and KNN Approach," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 90-95, 2017.

[3] Bhuiyan H., Ashiquzzaman A., Juthi T., Biswas S., and Ara J., "A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques," *Global Journal of Computer Science and Technology*, vol. 18, no. C2, pp. 21-29, 2018.

[4] Cohen A., Nissim N., and Elovici Y., "Novel Set of General Descriptive Features for Enhanced Detection of Malicious Emails Using Machine Learning Methods," *Expert Systems with Applications*, vol. 110, pp. 143-169, 2018.

[5] Douzi S., AlShahwan F., Lemoudden M., and Ouahidi B., "Hybrid Email Spam Detection Model Using Artificial Intelligence," *International Journal of Machine Learning and Computing*, vol. 10, no. 2, pp. 316-322, 2020.

[6] Faris H., Al-Zoubi A., Heidari A., Aljarah I., Mafarja M., Hassonah M., and Fujita H., "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.

[7] Ghaleb S., Mumtazimah M., Fadzli S., and Ghanem W., "Training Neural Networks by Enhance Grasshopper Optimization Algorithm for Spam Detection System," *IEEE Access*, vol. 9, pp. 116768-116813, 2021.

[8] Gibson S., Issac B., Zhang L., and Jacob S., "Detecting Spam Email with Machine Learning Optimized with Bio-Inspired Metaheuristic Algorithms" *IEEE Access*, vol. 8, pp. 116768-116813, 2021.

[9] Gupta V., Mehta A., Goel A., Dixit U., and Pandey A., *Harmony Search and Nature Inspired Optimization Algorithms: Theory and Applications*, Springer, 2019.

[10] Gutierrez C., Kim T., Della Corte R., Avery J., Goldwasser D., Cinque M., and Bagchi S., "Learning from the Ones that Got Away: Detecting New Forms of Phishing Attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 6, pp. 988-1001, 2018.

[11] Jain V., Kapoor R., Gulyani S., and Dubey A., "Categorization of Spam Images and Identification of Controversial Images on Mobile Phones Using Machine Learning and Predictive Learning," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 22, no. 2, pp. 293-307, 2019.

[12] Kontsewaya Y., Antonov E., and Artamonov A., "Evaluating the Effectiveness of Machine Learning Methods for Spam Detection," *Procedia Computer Science*, vol. 190, pp. 479-486, 2021.

[13] Kumaresan T. and Palanisamy C., "E-Mail Spam Classification Using S-cuckoo Search and Support Vector Machine," *International Journal of Bio-Inspired Computation*, vol. 9, no. 3, pp. 142-156, 2017.

[14] Li W., Meng W., Tan Z., and Xiang Y., "Design of Multi-view-based Email Classification for IOT Systems via Semi-supervised Learning," *Journal of Network and Computer Applications*, vol. 128, pp. 56-63, 2019.

[15] Madisetty S. and Desarkar M., "A Neural Network-Based Ensemble Approach for Spam Detection in Twitter," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 973-984, 2018.

[16] Maroofi S., Korczyński M., Hölzel A., and Duda A., "Adoption of Email Anti-Spoofing Schemes: A Large-Scale Analysis Technique," *Applied*

Machine Learning for Smart Data Analysis in IEEE Transactions on Network and Service Management, vol. 18, no. 3, pp. 3184-3196, 2021.

[17] Mohammed M., Ibrahim D., and Salman A., "Adaptive Intelligent Learning Approach Based on Visual Anti-spam Email Model for Multi-Natural Language," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 774-792, 2021.

[18] Nagwani N., "A Bi-Level Text Classification Approach for SMS Spam Filtering and Identifying Priority Messages," *The International Arab Journal of Information Technology*, vol. 14, no. 4, pp. 473-480, 2016.

[19] Rastenis J., Ramanauskaitė S., Suzdalev I., Tunaitytė K., Janulevičius J., and Čenys A., "Multi-Language Spam/Phishing Classification by Email Body Text: Toward Automated Security Incident Investigation," *Electronics*, vol. 10, no. 6, pp. 668, 2021.

[20] Rehman A., Javed K., and Babri H., "Feature Selection Based on a Normalized Difference Measure for Text Classification," *Information Processing and Management*, vol. 53, no. 2, pp. 473-489, 2017.

[21] Sah U. and Parmar N., "An Approach for Malicious Spam Detection in Email with Comparison of Different Classifiers," *International Research Journal of Engineering and Technology*, vol. 4, no. 8, pp. 2238-2242, 2017.

[22] Venkatraman S., Surendiran B., and Arun Raj Kumar P., "Spam E-mail Classification for the Internet of Things Environment Using Semantic Similarity Approach," *The Journal of Supercomputing*, vol. 76, no. 2, pp. 756-776, 2020.

[23] Zamir A., Khan H., Mehmood W., Iqbal T., and Akram, A., "A Feature-centric Spam Email Detection Model Using Diverse Supervised Machine Learning Algorithms," *The Electronic Library*, vol. 38, no. 3, pp. 633-657, 2020.

Vinoth Narayanan Arumugam Subramaniam is currently working as Assistant Professor in Computing Technologies, SRM Institute of science and Technologies and also pursuing research at Vels Institute of Science Technology and Advanced Studies. He has 7 years of Academic and Research experience. His research interest includes Network Security, Big Data and Machine Learning. He has presented nearly 10 research articles in National and international conferences. He has published 9 Research Articles in various reputed journals. He is an IELTS Scorer and also a member of IEI as well as in ACM.

Rajesh Annamalai received the M.Tech. in Computer science and Engineering, from the VIT University, Vellore Tamil Nādu, India (2004) and his Ph.D. from Anna University, Chennai, Tamil Nādu, India (2017). He is presently the Associate Professor of Computer Science Engineering at the School of Engineering of VISTAS University, Chennai, Tamil Nādu, India, where he has established an advanced research Virtual Reality laboratory, emphasizing AR-VR visual hybrid tracking approach and the guiding part of a reliable indoor navigation requests for 3D model of the environment. His research interest includes technology and applications of Machine Learning, AR-VR Technology and Trusted Network telecommunication.