Received: July 24, 2023. Revised: September 28, 2023.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

An Efficient Spam Mail Classification Approach Using Improved Moth Flame Optimization and Multi-class Support Vector Machine

Merly Thomas[1]* **Bandu B. Meshram[2]**

[1]*Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India*
[2]*Department of Computer Engineering, Veermata Jijabai Technological Institute, Mumbai, India*
* Corresponding author's Email: merly@fragnel.edu.in

Abstract: In the modern era, communication through email is increased dramatically due to the cost-effectiveness, usage of contexts, application advertising, and so on. However, e-mails are considered a professional way of communication that helps both commercial and non-commercial organizations to share important documents, reports, and so on. Since e-mail acts as a global pathway, it attracts more intruders to create spam messages which result in storage consumption and virus attacks. To overcome these issues, an improved classification approach is introduced to classify spam emails from ordinary emails. The data is obtained from four benchmark datasets such as Enron, Lingspam, Spamassassin, TREC, and the pre-processing is performed using tokenization, lemmatization, and stemming. After this feature extraction is performed using Bag-of n-grams, latent dirichlet analysis (LDA), term frequency, and inverse document frequency (TF-IDF). Then the feature selection is performed using the proposed improved moth flame optimization (IMFO) algorithm and finally, the classification is performed using multi-class support vector machine (MSVM). The results obtained through experimental analysis show that the proposed IMFO-MSVM has achieved better accuracy of 98.68% whereas the existing semantic graph neural network (SGNN) and fuzzy rule based long short term memory (LSTM) have obtained accuracy of 97.87% and 97% respectively.

Keywords: Improved moth flame optimization algorithm, Latent dirichlet analysis, Multi-class support vector machine, Spam mail classification and term frequency and inverse document frequency.

1. Introduction

An E-mail based server acts as a service-based application that helps users to send text messages and images. In recent days, more people tend to have an email account to use in their daily tasks such as banks, e-commerce websites, and so on [1, 2]. In other words, email is defined as a file that is comprised of text, files, web addresses, etc. E-mail services are widely used in the field of various applications related to transmitting bulk messages to an individual or a group of persons [3]. However, the global usage of the internet relies as a major reason that increases the count of spammers who creates spam emails. The word spam is defined as the undesired or harmful mail that evolves in the internet whereas ham is defined as the valid and significant mail which is sent by the recipient [4]. Moreover, this drastic increase in spam attacks creates more negative impacts over the recent time over the internet [5, 6]. Spamming is widely seen in web-based services such as emails and those spam emails are fake or junk mail which may rely as a significant reason to spread malware like viruses and Trojans. Some organizations may create spam emails which help to advertise their ads to more users. However, they are not harmful, it wastes the user's time and consumes more memory [7, 8].

More number of researchers have introduced various methodologies based on spam detection systems (SDS) to look after the spammers who create spam emails by detecting the pattern of the emails [9-11] The spam mail has increased among email users at the global level and these unsolicited emails have high cost in terms of storage space, time and consumption of network bandwidth. The researchers have put forward a step to generate new classification techniques to filter out spam emails and improvise the

user experience [12, 13]. Moreover, the usage of an effective feature selection approach can minimize the dimensionality of the data and helps various machine learning applications [14, 15]. Thus, this research introduced an optimization-based feature selection approach which helps to select the relevant features and helps in the process of classifying spam mail.

The major contributions of this research are listed as follows:

1. This research introduced an improved moth flame optimization algorithm to select the appropriate features and remove redundant information to ease the process of classification.
2. The features selected using IMFO algorithm are fed into the multiclass support vector machine which effectively classifies the spam mail from ordinary mail.

The remaining portion of this research paper is organized in the following manner: section 2 describes the related works of this research and the proposed methodology is described in section 3. Section 4 of the manuscript provides the results and finally, the overall conclusion of this research is described in section 5.

2. Related works

This section describes some of the recent existing approaches which are based on the spam mail classification approaches.

Pan [16] have introduced a semantic graph neural network (SGNN) to overcome the issues related to the classification of spam emails. The SGNN approach changes the email classification problem to a graph classification problem which exhibits the emails in the form of graphs. After the stage of conversion, the SGNN approach is utilized for classifying emails as spam or ham. The SGNN helps to create email features from the semantic graph which helps to embed the words into numerical vectors. However, the absence of a proper pre-processing technique leads to diminishing the accuracy. Hnini [17] have suggested a deep multimodal feature level fusion architecture which was comprised of two embedding vectors which reliably improvise the email presentation with improved classification performance. The feature extraction was performed using a paragraph vector distributed bag of words (PV-DBOW) and convolutional neural network (CNN). After the stage of feature extraction, the mined features were fed into the random forest classifier which effectively categorize the email as spam or non-spam. The PV-DBOW helps effectively preserve the semantic features during the time of creating feature vectors which effectively improvise the classification accuracy of the model. However, the proposed architecture was not valid to categorize the spam present in image-based e-mails.

Srinivasarao and Sharaff [18] introduced a hybrid classifier that was developed based on classifying spam SMS. Initially, the data is pre-processed and the feature extraction is performed using the data augmentation approach. After this, the features were fed into six feature selection approaches and equilibrium optimization. Then, the classification was performed using the proposed hybrid classifier which was a combination of SVM and KNN. Furthermore, rat swarm optimization (RSO) was used to optimize the hyper-parameters and aid in better accuracy during the classification of SMS. However, the efficiency of the classifier is reduced when evaluated for smaller datasets. Hosseinalipour and Ghanbarzadeh [19] have introduced horse-herd optimization algorithm (HOA) to perform an effective detection of email spam with a minimal error rate. Initially, the continuous HOA is transformed into an algorithm with discrete components and transformed into a multi-objective. After this, spam detection takes place to identify email spam over the internet. The suggested HOA effectively selects the significant and relevant features which minimize the time complexity while classifying email spam. But, the usage of machine learning classifiers was not enough to evaluate the efficiency of the suggested approach.

Ismail [20] have introduced a hybrid processing mechanism which is a combination of genetic decision tree processing with natural language processing (GDTPNLP) to detect email spam in the form of text and voice. The GDTPNLP combines the benefits of both the genetic procedure and decision tree and works in a bidirectional manner to classify the spam present in text mail and voice mail. The features extracted using principle component analysis effectively minimize the data overhead during the classification process. However, the model must be properly trained to store multi-keywords. Khan [21] have introduced a fuzzy logic-based multi-criteria metric for evaluating the performance of spam detection techniques. The suggested approach integrates accuracy, precision and recall into a multi-criteria fuzzy function. The fuzzy rule based multi-criteria key is implemented with bidirectional encoder representations from transformers (BERT) and long short-term memory (LSTM) to evaluate the effectiveness of suggested fuzzy logic-based multi-criteria. However, the suggested fuzzy logic

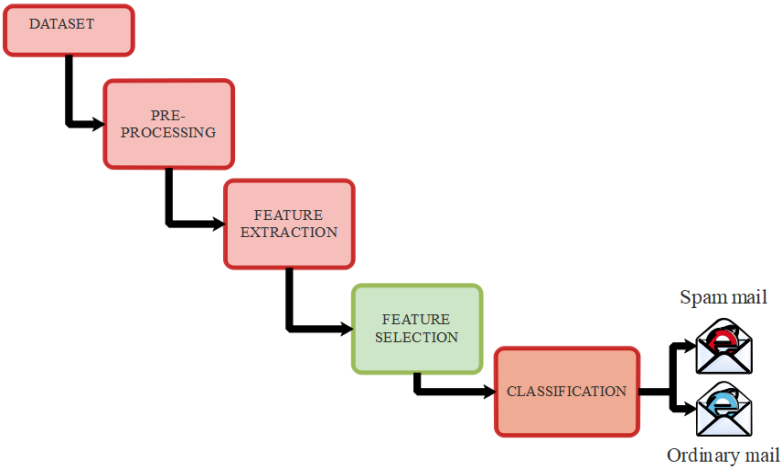

Figure. 1 Overall process involved in the classification of spam mail and ordinary mail

approach lacks the granularity of decision rules based on the fuzzy approach.

Sultan Zavrak and Seyhmus Yilmaz [22] have introduced a hierarchical attention hybrid neural network (HAN) to detect and classify the spam mail. The suggested approach was integrated with convolutional neural network (CNN), gated recurrent unit (GRU) with attention mechanisms. The suggested approach was flexible to detect temporal convolutions with receptive emails of varying sizes. However, the proposed approach does not suit for detecting the spam mails from real time scenarios. Sanaa A.A. Ghaleb [23] have introduced a spam detection system based on feature selection using multi-objective optimization approach. The suggested approach was comprised with multi-objective grasshopper optimization algorithm (MOGOA) for selecting the appropriate features and training the multi-layer perceptron (MLP) for detecting the spam mails. However, the combination of term frequency-inverse document frequency (TF-IDF) and bag of words (BoW) enhance the time of execution.

Sylwia Rapacz [24] have introduced a multi-stage meta-algorithm for evaluating the performance of the classifier. The suggested algorithm has the capability to select the and analyse the data with higher dimensionalities. The cross validation among supervised learning approach is validated using the proposed meta-algorithm. The inclusion of meta-algorithm was integrated with the classifiers to provide better accuracy of classifying spam messages. However, an appropriate pre-processing technique should be employed to normalize the unbalanced data.

3. Improved moth flame optimization algorithm for classifying spam emails

In this research, an effective classification to categorize spam mail is performed using the proposed feature selection method. This research introduced an improved moth flame optimization to select the appropriate features which have a significant role in diminishing the complexities that occur during classification. The classification of spam mail from normal mail undergoes various stages such as data acquisition, pre-processing, feature extraction, feature selection, and classification. Among these five stages, this research introduced an optimization-based feature selection approach to select the relevant features. Fig. 1 depicts the overall process involved in classifying normal mail and spam mail.

3.1 Data acquisition

Data acquisition is defined as the process of obtaining raw data from various sources based on the user's needs. In this research, the raw data is obtained from four data sets such as Enron, Lingspam, Spamassassin, TREC datasets. The description of the fore mentioned datasets is mentioned as follows:

Enron: The Enron spam dataset [25] was gathered from Federal Energy Regulatory which is approximately comprised of 500000 emails which were created by employees of Enron.

Lingspam: The Lingspam dataset [26] is a collection of 2,893 spam and non-spam messages which is obtained from a linguist list. The messages present in the Lingspam dataset focus on linguistic interests related to job postings, research opportunities, and software discussions.

Spamassassin: The Spamassassin dataset [27] is a popular corpus benchmark dataset that is comprised of 3252 emails of which 2751 messages are legitimate and 501 are spam emails.

TREC: In TERC dataset [28], every individual mail is labeled as spam and non-spam using the chronological index. In the TREC dataset, there are about 92,189 emails of which 39,399 are considered as ham and the remaining 52, 790 emails are labeled as spam emails.

3.2 Pre-processing

After the stage of data acquisition, the raw data is pre-processed to make the text suitable for direct analysis. In the stage of pre-processing, the undesired information from the raw input data is removed to improve the quality of the input data. In this research, the pre-processing is performed using three techniques such as tokenization, stemming and lemmatization. The process involved in tokenization, stemming and lemmatization is described as follows:

Tokenization: It is one of the important stage in data pre-processing where all the words from the email are collected and the repetition of the words are counted. In the tokenization process, the count vectorizer is used to find the repeated words in the dataset and each word is allotted with a unified number known as tokens. These tokens are comprised of the type of feature values which helps in the process of creating feature vectors and in the stage of tokenization, every phase is assigned as tokens.

Stemming: After the stage of tokenization, the tokens need to be stemmed using the tokenization process. Stemming is defined as the process of converting the derived terms to their original forms. Initially, the base terms were exposed to prefixes and suffixes, then the stemming algorithm is used to change the modified words to the stemmed words. In this research, an effective stemming process is performed using the natural language Tool Kit (NLTK) library. After completion of an effective stemming process, the contents present in the email along with the spam words can be easily identified.

Lemmatization: Lemmatization is defined as the process of combining various inflected terms into a single term. The lemmatization reduces the word to its original form known as the lemma and the correct lemma can be identified by using the NLP tool to analyze the context, meaning, and intended meaning of the word. The process of lemmatization can effectively analyze the morphological characters of the words.

3.3 Feature extraction

After the stage of pre-processing, feature extraction is performed to transform the pre-processed data into numerical features while preserving the information from the original dataset. In this research feature extraction is performed using three techniques such as bag-of n-grams, latent dirichlet analysis (LDA), term frequency and inverse document frequency (TF-IDF).

Bag-of n-grams: The bag of n-grams is an extension of Bag of Words which is used to check the continuity of words that exist in the textual data. The bag of n-grams is comprised of three types such as unigram, bi-gram, and tri-gram. Single words are represented as unigrams, pair of words are represented as bigrams, and three or more words are represented as trigrams. The words which are extracted using a bag of n-gram is represented as T_r^{SE} and it is integrated with word2vec to provide a word vector representation. The actual root words can be identified using the bag of word concept without a lack of contextual meaning.

TF-IDF: After using a Bag of n-grams, the extracted words are subjected to the stage of TF-IDF and the value of TF is evaluated using the Eq. (1) as follows:

$$T_F = \frac{N_S}{TN} \tag{1}$$

Where the total number of terms are denoted as TN and the important term is evaluated using the IDF which is evaluated using the Eq. (2) as follows:

$$IDF = log_e\left(\frac{ND}{TD}\right) \tag{2}$$

Where the total number of emails present in the dataset is represented as TD and the characters present in the mail are denoted as N_D. Finally, the term weightage is evaluated using Eq. (3) as follows:

$$TF - IDF(s, DC) = TF(s, DC) \times IDF(s) \tag{3}$$

LDA: The extracted features from the TF-IDF are fed into the stage of LDA where the dimensionalities of the extracted features are minimized without affecting the original context information. The LDA is performed using the Eq. (4) as follows:

$$a_{opt} = argmax\frac{a^T S_b a}{a^T S_w a} \tag{4}$$

Where the between scatter matrix and the within scatter matrix is represented as S_b and S_w.

3.4 Feature selection

After the stage of feature extraction, feature selection is performed to minimize the number of input values by selecting the significant features. The process of neglecting redundant information helps to lower the computational cost and enhance the classification performance. Feature selection is a multi-objective process that helps to achieve a tradeoff between accuracy and the number of features. The feature selection helps to eliminate the inappropriate features and helps to improvise the convergence tendency. In this research, an improved moth flame optimization (IMFO) algorithm is introduced to select the relevant features which ease the process of classification. IMFO is an improvisation of the MFO algorithm where the moths are considered as the candidate solution. The MFO is based on swarm based optimization algorithm and the population of the candidate is represented in the following Eq. (5) as follows:

$$M = \begin{bmatrix} m_{1,1} & \cdots & m_{1,d} \\ \vdots & \ddots & \vdots \\ m_{n,1} & \cdots & m_{n,d} \end{bmatrix} \tag{5}$$

Where the number of moths are represented as n and the number of control variables are represented as d.

Every individual moth in the swarm gets surrounded by a respective flame and gets updated to the location of flame in the next iteration. The position of each moth to its corresponding flame is updated based on Eq. (6) as follows:

$$M_i = S(M_i, F_j) \tag{6}$$

Where the ith moth is represented as M_i, the $j$the flame is represented as F_j and the helical function is represented as S. The helical function of the moth flight path is represented in Eq. (7) as follows:

$$S(M_i, F_j) = D_i . e^{bt} . \cos(2\pi t) + F_j \tag{7}$$

Where the value of $t = (a-1) \times rand + 1$ and the value of $a = -1 + iteration \times \left(-\frac{1}{T_{max}}\right)$

Where the distance between ith moth and the jth flame is represented as D_i. The logarithmic helix shape is represented as b and the path co-efficient t lies among the range of $[-1,1]$. The value of D_i is evaluated using the Eq. (8) as follows:

$$D_i = |F_j - M_i| \tag{8}$$

The position of each moth relies on different location search space and the development capability gets diminished. The reduction of flame to the corresponding moth leads to inappropriate position updates based on the present fitness value.

3.4.1. Improved moth flame optimization (IMFO) algorithm

The inappropriate position updates and the poor search ability can be overwhelmed using the proposed IMFO algorithm. In IMFO, the Levy flight approach is used to perform global exploration and improve the search ability of the moth present in the swarm.

Levy flight mechanism

The Levy flight is a kind of random walk mechanism which involves a non-Gaussian stochastic process regarding Levy stable distribution. The combination of MFO in Levy flight probably enhances the search range and helps to jump out of the local optimum. By using Levy flight mechanism, search scope gets increased and prohibits the algorithm to fall into the local optimum. The improved formula for an effective search space is represented in Eq. (9) as follows:

$$S(M_i, F_j) = D_i . e^{bt} . \cos(2\pi t) + L(d) . F_j \tag{9}$$

Where the current iteration number is represented as t, the distance between the ith moth and the jth flame is represented as D_i and the levy flight mechanism is represented as $L(d)$. When the position of the moth spiral flight gets updated, then Levy flight mechanism can further improvise their search space. The formula of Levy flight is represented in Eq. (10) as follows:

$$L(d) = 0.01 \frac{r_1 \delta}{|r_2|^{\frac{1}{\varphi}}} \tag{10}$$

Where the random numbers which lie among the range of [0,1] are represented as r_1 and r_2 and the constant is denoted as φ. The formula of δ is evaluated using the Eq. (11) as follows:

$$\delta = \left(\frac{\tau(1+\varphi) \sin\left(\frac{\pi\varphi}{2}\right)^{\frac{1}{\varphi}}}{\tau\left(\frac{1+\varphi}{2}\right) \varphi 2^{\left(\frac{\varphi-1}{2}\right)}} \right) \tag{11}$$

In the Levy flight mechanism, an exploratory search performed over the shorter distance was the

same as the occasional long walks. The Levy flight approach embedded in Eq. (9) enhances the search efficiency in uncertain environments. Moreover, the Levy flight strategy helps IMFO from falling from local optimality. Thus, the IMFO helps in the process of extracting the appropriate features by performing an effective search to find the optimal solution which eases the process of classification.

3.2 Classification

After the stage of feature selection, classification is performed to categorize spam emails from ordinary emails. Since the classification of spam emails is based on various classes and multiple classes, the multi-class support vector machine (M-SVM) is the process of classification. Moreover, the MSVM classifier was based on linear and radial basis function (RBF) kernel, it acquired the ability to classify spam emails effectively with maximal accuracy. The MSVM classifier requires a minimal training period with reduced decompositions known as one-against All (OAA) and the procedure involved in classification is represented as $K = \{w_{1,}w_{2,}w_{3,} \dots \dots, w_{c,}\}$. The resultant of the OAA is compatible with a higher output value represented in Eq. (12) as follows:

$$F_i(x) = w_i^T \emptyset(x) + b_i \tag{12}$$

Where the weighted vector and the training data are represented as w_i^T and x respectively. The mapping function and scalar data are represented as $\emptyset(x)$ and b_i respectively. The input vector x is assigned to each class that lies nearer to the decision function and the sample x is categorized as spam and ordinary mail based on the number of classes, which is represented in Eq. (13) as follows:

$$x = arg_{i=1,2,\dots,n}^{max}(F_i(x)) \tag{13}$$

Thus, the MSVM is well suited for classifying emails as spam or ham and results in better classification accuracy.

4. Results and analysis

This section provides the results obtained from the proposed IMFO algorithm when it is evaluated with different existing optimization techniques for different datasets such as Enron, Lingspam, Spamassassin and TREC datasets. The result section is categorized into two sub-sections such as performance analysis and comparative analysis which are described as follows: Moreover, the results are evaluated with different classifiers with actual features and optimized features. The proposed approach is implemented in MATLAB software and the system with specifications such as Intel core i5 processor, 8 GB of random-access memory and windows 10 operating system. The performance of the proposed approach is evaluated by means of accuracy, sensitivity, specificity, Matthews Correlation Coefficient (MCC) and error rate. The fore mentioned performance metrics are evaluated using the formula mentioned in Eqs. (14-18).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \tag{14}$$

$$Sensitivity = \frac{TP}{(TP+FN)} \times 100 \tag{15}$$

$$Specificity = \frac{TN}{(FP+TN)} \times 100 \tag{16}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \times 100 \tag{17}$$

$$Error\ rate = Accuracy - 100 \tag{18}$$

Where TN is known as True Negative, TP is known as True Positive, FN is known as False Negative and FP is known as False Positive.

4.1 Performance analysis

In this sub section, firstly the performance of the classifier with actual and optimized features is evaluated. Secondly, the performance of the proposed optimization algorithm is evaluated for different datasets. Table 1 presented below shows the performance of the MSVM classifier when compared with the existing classifiers such as K-nearest neighbor (KNN), random forest (RF) and decision tree (DT).

The overall results obtained from Table 1-Table 4 shows that the MSVM classifier utilized in this research have obtained better performance for both actual and optimized features. For instance, when the MSVM classifier is evaluated for ENRON dataset, it obtained classification accuracy of 99.44% for optimized features whereas the other classifiers such as KNN, RF and DT have obtained classification accuracy of 89.63%, 91.91% and 89.67% respectively. The better result of the MSVM classifier is due to the proposed IMFO for feature selection which effectively selects the appropriate features and makes the classification process easier with better classification accuracy. Fig. 2 depicted below shows the graphical representation for performance of

Table 1. Performance analysis for various classifiers for ENRON dataset

	Classifiers	Accuracy	Sensitivity	Specificity	MCC	Error rate
Actual Features	KNN	87.51	86.48	85.38	83.49	12.49
	RF	86.30	87.86	85.91	89.31	13.70
	DT	87.53	86.96	85.08	85.69	12.47
	MSVM	96.53	95.16	94.08	92.97	3.47
Optimized Features	KNN	89.63	87.05	90.81	85.58	10.37
	RF	91.91	90.49	91.79	87.86	8.09
	DT	89.67	88.02	86.39	86.05	10.33
	MSVM	98.68	98.96	97.51	94.44	1.32

Table 2. Performance analysis for various classifiers for Lingspam dataset

	Classifiers	Accuracy	Sensitivity	Specificity	MCC	Error rate
Actual Features	KNN	90.26	88.21	91.86	92.65	9.74
	RF	92.14	90.90	89.45	89.34	7.86
	DT	90.41	88.00	90.02	91.57	9.59
	MSVM	96.90	94.38	93.63	95.47	3.10
Optimized Features	KNN	92.66	93.45	91.99	9.61	7.34
	RF	95.74	94.84	92.86	93.67	1.86
	DT	9.42	89.67	93.12	90.77	9.59
	MSVM	99.44	99.72	98.96	97.54	3.10

Table 3. Performance analysis of various classifiers for Spamassassin dataset

	Classifiers	Accuracy	Sensitivity	Specificity	MCC	Error rate
Actual Features	KNN	88.34	86.56	84.28	83.94	11.66
	RF	91.05	90.93	89.42	90.56	8.95
	DT	90.66	87.28	85.06	88.29	9.34
	MSVM	96.14	97.11	93.05	94.40	3.86
Optimized Features	KNN	92.13	90.90	91.61	90.91	7.87
	RF	95.29	96.17	94.35	93.56	4.71
	DT	93.20	90.43	92.56	91.13	6.80
	MSVM	99.85	98.81	97.83	96.42	0.15

Table 4. Performance analysis of various classifiers for TREC dataset

	Classifiers	Accuracy	Sensitivity	Specificity	MCC	Error rate
Actual Features	KNN	85.60	83.34	82.53	84.95	14.40
	RF	90.53	89.15	87.21	90.49	9.47
	DT	89.69	87.54	86.66	89.97	10.31
	MSVM	95.32	94.76	93.68	92.05	4.68
Optimized Features	KNN	89.05	88.64	87.36	86.52	10.95
	RF	92.16	90.70	91.58	92.42	7.84
	DT	93.39	93.26	90.21	91.04	6.61
	MSVM	97.88	97.94	96.53	95.94	2.12

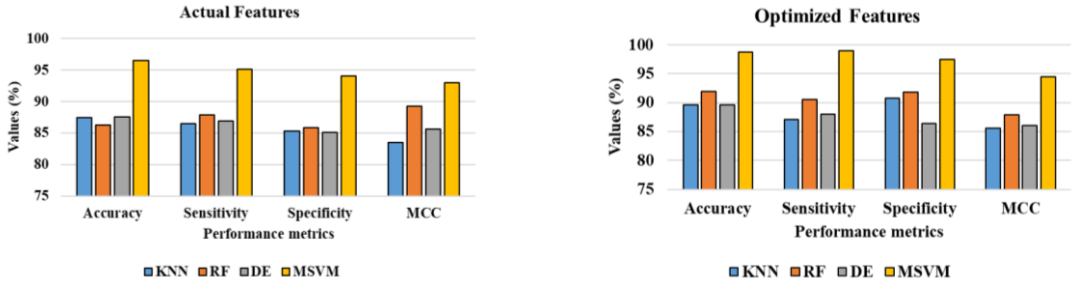

Figure. 2 Graphical representation for the performance of the classifier based on actual and optimized features for the ENRON dataset

Table 5. Performance analysis of various classifiers for ENRON dataset

Algorithms	Accuracy	Sensitivity	Specificity	MCC	Error rate
PSO	91.17	89.10	90.75	93.89	8.83
ACO	93.82	94.98	90.35	92.12	6.18
ABC	96.24	94.22	96.83	93.73	3.76
IMFO	98.68	98.96	97.51	94.44	1.32

Table 6. Performance analysis of various classifiers for Lingspam dataset

Algorithms	Accuracy	Sensitivity	Specificity	MCC	Error rate
PSO	92.59	94.77	93.35	91.15	7.41
ACO	93.58	92.17	94.26	92.18	6.42
ABC	95.45	96.80	94.94	95.83	4.55
IMFO	99.44	99.72	98.96	97.54	0.56

Table 7. Performance analysis of various classifiers for Spamassassin dataset

Algorithms	Accuracy	Sensitivity	Specificity	MCC	Error rate
PSO	93.35	92.53	91.15	89.68	6.65
ACO	92.42	91.81	92.57	89.31	7.58
ABC	90.38	89.06	85.71	81.28	9.62
IMFO	99.85	98.81	97.83	96.42	0.15

Table 8. Performance analysis of various classifiers for TREC dataset

Algorithms	Accuracy	Sensitivity	Specificity	MCC	Error rate
PSO	90.35	89.41	91.49	88.16	9.65
ACO	92.60	90.47	93.31	90.07	7.40
ABC	93.88	92.36	91.81	90.87	6.12
IMFO	97.88	97.94	96.53	95.94	2.12

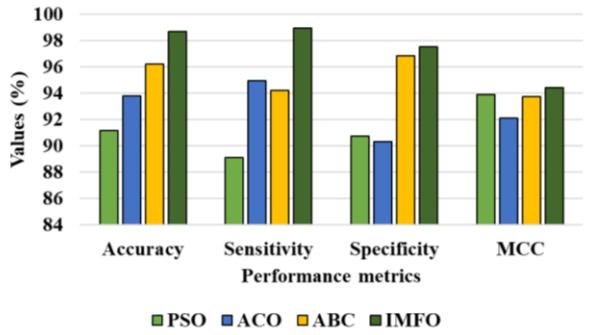

Figure. 3 Graphical representations for the performance of various classification methods for ENRON dataset

classifiers for ENRON dataset based on actual features and the optimized features.

Secondly, the performance of various optimization algorithms is evaluated for four datasets based on the performance metrics such as accuracy, sensitivity, specificity, MCC and error rate. Table 5-Table 8 depicted below presents the performance of the optimization algorithm when evaluated for the ENRON dataset, Lingspam, Spamassassin and TREC dataset respectively. The overall results from Table 5-Table 8 show that the proposed IMFO has obtained better results in terms of accuracy, sensitivity, specificity, MCC and error rate when compared to other optimization techniques. For instance, the error rate produced by the proposed IMFO for the ENRON dataset is 1.32% whereas the error rate of other optimization algorithms such as particle swarm optimization (PSO), ant colony optimization (ACO), and artificial bee colony (ABC) have obtained an error rate of 8.83%, 6.18%, 3.76%, and 1.32% respectively. The Levy flight mechanism involved in the IMFO algorithm relies as a reason to provide better performance by performing an advanced search The combination of MFO in the Levy flight probably enhances the search range and helps to jump out of the local optimum. By using the Levy flight mechanism, the search scope gets increased and prohibits the algorithm to fall into the local optimum. Thus, the proposed IMFO algorithm with an improved search ability identifies the appropriate features and helps in the process of classifying spam emails. Fig. 3 depicted below shows the graphical representation for performance evaluation of various optimization techniques for ENRON dataset.

4.2 Comparative analysis

This section provides the comparative results which are evaluated by comparing the proposed approach with the existing methodologies for various datasets. Table 9 depicted below presents the comparative results of the proposed IMFO-SVM when evaluated with existing approaches such as Semantic Graph Neural Network (SGNN) [16],

Table 9. Comparison of the proposed method with existing methods for various datasets

Methodologies	Dataset	Accuracy (%)	Sensitivity (%)
SGNN [16]	Enron	97.87	NA
	TREC	96.57	NA
Hybrid KNN-SVM [18]	Spamassassin	99.69	NA
Fuzzy rule based LSTM [21]	Enron	97	98
	Lingspam	98	98
HAN [22]	Enron	95.8	93.7
	Lingspam	98.0	94.8
	Spamassassin	95.5	97.8
MOGOA [23]	Spamassassin	98.3	98
IMFO-MSVM	Enron	98.68	97.51
	TREC	97.88	96.53
	Spamassassin	99.85	97.83
	Lingspam	99.44	98.96

Hybrid KNN-SVM [18], fuzzy rule based LSTM [21], HAN [22] and MOGOA [23]. Accuracy and sensitivity are considered as the common performance metric among various methodologies to evaluate the efficacy of the proposed approach.

The results from Table 9 show that the proposed IMFO-MSVM outperforms well when compared with the existing methodologies, where NA is the value which is not available in the corresponding researches. For example, for the Enron dataset, the proposed IMFO-MSVM achieves an accuracy of 98.68% whereas the existing SGNN, and fuzzy rule based LSTM have obtained accuracy of 97.87% and 97% respectively. Similarly, for Spamassassin dataset, the proposed approach achieved accuracy of 99.85% whereas the Hybrid KNN-SVM, HAN and MOGOA obtained accuracy of 99.69%, 95.5% and 98.68% respectively. The better result of the proposed IMFO-MSVM is due to the effective feature selection performed by the proposed IMFO which selects the appropriate features by performing an effective search using the Levy flight mechanism integrated with it.

5. Conclusion

Unnecessary mail or spam mail relies as an issue among internet users and data centres due to its large storage consumption and acts as a gateway for cyberattacks. The major objective of this research paper is to develop an effective classification approach by selecting the relevant features using the proposed IMFO algorithm, the improvisation is made by introducing the Levy flight mechanism. The Levy flight mechanism improvises the search ability and prohibits the solution to fall into the local optimum. Moreover, the proposed IMFO can overwhelm the problems related to poor position updates which is seen in the existing MFO algorithm. The IMFO effectively selects the relevant features by eliminating redundant data and aids in better classification accuracy performed by the MSVM classifier. The experimental results show that the proposed approach achieved better accuracy of 98.68% whereas the existing SGNN and fuzzy rule based LSTM have obtained accuracy of 97.87%, and 97% respectively. In the future, the deep learning based classifiers can be utilized with the proposed approach to achieve better classification accuracy.

Conflicts of interest

The authors declare no conflict of interest

Author contributions

For this research work all authors' have equally contributed in Conceptualization, methodology, validation, resources, writing—original draft preparation, writing—review and editing.

References

[1] M. K. Islam, M. A. Amin, M. R. Islam, M. N. I. Mahbub, M. I. H. Showrov, and C. Kaushal, “Spam-Detection with Comparative Analysis and Spamming Words Extractions”, In: *Proc. of 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, pp. 1-9, 2021.

[2] S. Rapacz, P. Chołda, and M. Natkaniec, “A Method for Fast Selection of Machine-Learning Classifiers for Spam Filtering”, *Electronics*, Vol. 10, No. 17, p. 2083, 2021.

[3] S. A. A. Ghaleb, M. Mohamad, W. A. H. M. Ghanem, A. B. Nasser, M. Ghetas, A. M. Abdullahi, S. A. M. Saleh, H. Arshad, A. E. Omolara, and O. I. Abiodun, “Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks

and Grasshopper Optimization Algorithm", *IEEE Access*, Vol. 10, pp. 98475-98489, 2022.

[4] Z. F. Sokhangoee, and A. Rezapour, "A novel approach for spam detection based on association rule mining and genetic algorithm", *Computers & Electrical Engineering*, Vol. 97, p. 107655, 2022.

[5] A. Sharaff, C. Kamal, S. Porwal, S. Bhatia, K. Kaur, and M. M. Hassan, "Spam message detection using Danger theory and Krill herd optimization", *Computer Networks*, Vol. 199, p. 108453, 2021.

[6] J. D. P. Rosita, W. S. Jacob, "Multi-objective genetic algorithm and CNN-based deep learning architectural scheme for effective spam detection", *International Journal of Intelligent Networks*, Vol. 3, pp. 9-15, 2022.

[7] N. Bacanin, M. Zivkovic, C. Stoean, M. Antonijevic, S. Janicijevic, M. Sarac, and I. Strumberger, "Application of Natural Language Processing and Machine Learning Boosted with Swarm Intelligence for Spam Email Filtering", *Mathematics*, Vol. 10, No. 22, p. 4173, 2022.

[8] T. O. Omotehinwa, and D. O. Oyewola, "Hyperparameter Optimization of Ensemble Models for Spam Email Detection", *Applied Sciences*, Vol. 13, No. 3, p. 1971, 2023.

[9] N. Ghatasheh, I. Altaharwa, and K. Aldebei, "Modified Genetic Algorithm for Feature Selection and Hyper Parameter Optimization: Case of XGBoost in Spam Prediction", *IEEE Access*, Vol. 10, pp. 84365-84383, 2022.

[10] P. Pirozmand, M. Sadeghilalimi, A. A. R. Hosseinabadi, F. Sadeghilalimi, S. Mirkamali, and A. Slowik, "A feature selection approach for spam detection in social networks using gravitational force-based heuristic algorithm", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 14, No. 3, pp. 1633-1646, 2021.

[11] N. Saidani, K. Adi, and M. S. Allili. "A semantic-based classification approach for an enhanced spam detection", *Computers & Security*, Vol. 94, p. 101716, 2020.

[12] R. T. 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.

[13] S. Venkatraman, B. Surendiran, and P. A. R. Kumar, "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.

[14] M. Shuaib, S. M. Abdulhamid, O. S. Adebayo, O. Osho, I. Idris, J. K. Alhassan, and N. Rana, "Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification", *SN Applied Sciences*, Vol. 1, No. 5, p. 390, 2019.

[15] W. Park, N. M. F. Qureshi, and D. R. Shin, "Pseudo NLP joint spam classification technique for big data cluster", *Computers, Materials and Continua*, Vol. 71, No. 1, pp. 517-535, 2022.

[16] W. Pan, J. Li, L. Gao, L. Yue, Y. Yang, L. Deng, and C. Deng, "Semantic graph neural network: a conversion from spam email classification to graph classification", *Scientific Programming*, Vol. 2022, p. 6737080, 2022.

[17] G. Hnini, J. Riffi, M. A. Mahraz, A. Yahyaouy, and H. Tairi, "MMPC-RF: A Deep Multimodal Feature-Level Fusion Architecture for Hybrid Spam E-mail Detection", *Applied Sciences*, Vol. 11, No. 24, p. 11968, 2021.

[18] U. Srinivasarao and A. Sharaff, "Machine intelligence based hybrid classifier for spam detection and sentiment analysis of SMS messages", *Multimedia Tools and Applications*, 2023.

[19] A. Hosseinalipour and R. Ghanbarzadeh, "A novel approach for spam detection using horse herd optimization algorithm", *Neural Computing and Applications*, Vol. 34, No. 15, pp. 13091-13105, 2022.

[20] S. S. I. Ismail, R. F. Mansour, R. M. A. E. Aziz, and A. I. Taloba, "Efficient E-mail spam detection strategy using genetic decision tree processing with NLP features", *Computational Intelligence and Neuroscience*, Vol. 2022, p. 7710005, 2022.

[21] S. A. Khan, K. Iqbal, N. Mohammad, R. Akbar, S. S. A. Ali, and A. A. Siddiqui, "A Novel Fuzzy-Logic-Based Multi-Criteria Metric for Performance Evaluation of Spam Email Detection Algorithms,, *Applied Sciences*, Vol. 12, No. 14, p. 7043, 2022.

[22] S. Zavrak and S. Yilmaz, "Email spam detection using hierarchical attention hybrid deep learning method", *Expert Systems with Applications*, Vol. 233, p.120977, 2023.

[23] S. A. A. Ghaleb, M. Mohamad, W. A. H. M. Ghanem, A. B. Nasser, M. Ghetas, A. M. Abdullahi, S. A. M. Saleh, H. Arshad, A. E. Omolara, and O. I. Abiodun, "Feature Selection by Multiobjective Optimization: Application to Spam Detection System by Neural Networks and Grasshopper Optimization Algorithm", *IEEE Access*, Vol. 10, pp. 98475-98489, 2022.

[24] S. Rapacz, P. Chołda, and M. Natkaniec, "A method for fast selection of machine-learning classifiers for spam filtering", *Electronics*, Vol. 10, No. 17, p.2083, 2021.
[25] S. Douzi, F. A. AlShahwan, M. Lemoudden, and B. E. Ouahidi, "Hybrid email spam detection model using artificial intelligence", *International Journal of Machine Learning and Computing*, Vol. 10, No. 2, pp. 316-322, 2020.
[26] Link to download Lingspam dataset: http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz
[27] A. S. Rajput, V. Athavale, and S. Mittal, "Intelligent model for classification of SPAM and HAM", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 6S, pp. 773-777, 2019.
[28] M. Dredze, R. Gevaryahu, and A. E. Bachrach, "Learning Fast Classifiers for Image Spam", In: *Proc. of the Fourth Conference on Email and Anti-Spam*, Mountain View, CA, USA, pp. 2-3 August 2007.