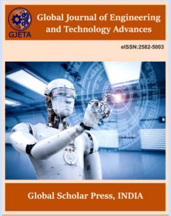
Global Journal of Engineering and Technology Advances

eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/


Global Journal of Engineering and Technology Advances
GJETA
eISSN:2582-5003
Global Scholar Press, INDIA

(Research Article)

Efficient genetic algorithm for spam mail detection and classification

B. Aruna Kumari * and C. Nagaraju

Department of CSE, YSR Engineering College of YVU, Proddatur. India.

Global Journal of Engineering and Technology Advances, 2024, 18(02), 035–041

Publication history: Received on 12 December 2023; revised on 05 February 2024; accepted on 08 February 2024

Article DOI: https://doi.org/10.30574/gjeta.2024.18.2.0017

Abstract

E-mail is the quickest tool to convey information from one to others. Individual users and organizations have become more reliant on e-mails as technology advances. At the moment, all email inboxes are swamped with spam and these cause significant and diverse difficulties, such as the loss of crucial information and the theft of the recipient's identity. Organizations may suffer significant losses. As a result, users cannot avoid spam emails, which come in various formats such as advertisements and messages. Spam filtering removes spam messages and prevents them from being accessed. This paper focuses on categorization of e-mails by using genetic algorithm. The proposed approach uses entropy to evaluate information gain, which is subsequently employed by the genetic algorithm classifier to choose important features from the spam database. The model performance is then assessed by using the testing dataset with all 57 features, a crossover probability of 0.3, and a mutation probability of 1.0 and obtained good accuracy.

Keywords: Spam Database; Crossover; Mutation; Fitness; Reproduction

1. Introduction

Email is regarded as a secure method of communication for exchanging information between individual users and organizations. Despite its benefits, the internet is increasingly being used to facilitate malicious acts [1][2]. As a result, e-mail users continue to expend significant time and effort on identifying spam and deleting it from their inbox. However, it is extremely inconvenient for people to check their email and discover thousands of emails from unknown senders with good headings but irrelevant content. As a result, this is only for spammers to promote their markets.

Spam emails are classified into two types: solicited e-mails and unsolicited e-mails. Solicited email is any commercial message, including those from companies, that the receiver chose to receive either through direct action or as an indirect consequence of another action, through an automatic or non-obvious procedure, like providing an email address while enrolling for an online shopping account from which the recipient has purchased things. Unsolicited email is one of the most common dangers to network integrity on the public internet. Unsolicited commercial email also known as spam, is sent to a large number of recipients who have not specifically requested the communication. An increasing number of unwanted emails are sent and received via the Internet every day.

To enhance the results of spam email filtering, the proposed work employs feature selection and a genetic algorithm classifier. One way of feature selection is information gain. The genetic algorithm is used to choose a population of solutions by employing crossover and mutation operators to discover the optimal model, which is then evaluated.

[*] Corresponding author: B. Aruna Kumari

2. Literature Survey

Table 1 Research gaps

Author, Year	**Aim**	**Method**	**Results**	**Research gaps**
Jitendra Nath Shrivastava, Maringanti Hima Bindu [3], 2013	Aim to use a genetic algorithm along with a heuristic function to classify emails.	Genetic algorithm based	The genetic algorithm achieved an efficiency of about 82%.	Because the selection of categories is mostly determined by the classification of emails, even if fewer categories are defined, an email can still be identified as spam. However, the FN/FP ratio rises.
Mandeep Singh, Prashant Sahai Saxena [4], 2017	To distinguish e-mails using fuzzy fusion of average probabilistic methods.	Genetic algorithm with probabilistic and average based methods	The recognition rate can be enhanced to 90% by using fuzzy fusion.	It is necessary to improve the rate of recognition.
Mandeep Choudhary, V.S. Dhaka [5], 2015	To classify e-mails using genetic algorithm	Genetic algorithm	The process efficiency is determined by the dataset and genetic algorithm parameters, and it is greater than 81%.	The number of false positive or false negative outcomes must be reduced.
Mandeep Choudhary, V.S. Dhaka [6], 2015	To filter spam e-mail using genetic algorithm.	Various variations of genetic algorithm	Genetic algorithm variation has a high impact on overall performance.	The efficiency increases as the number of mails in the corpus grows. This rise in efficiency, however, is not progressive.

3. Proposed Method

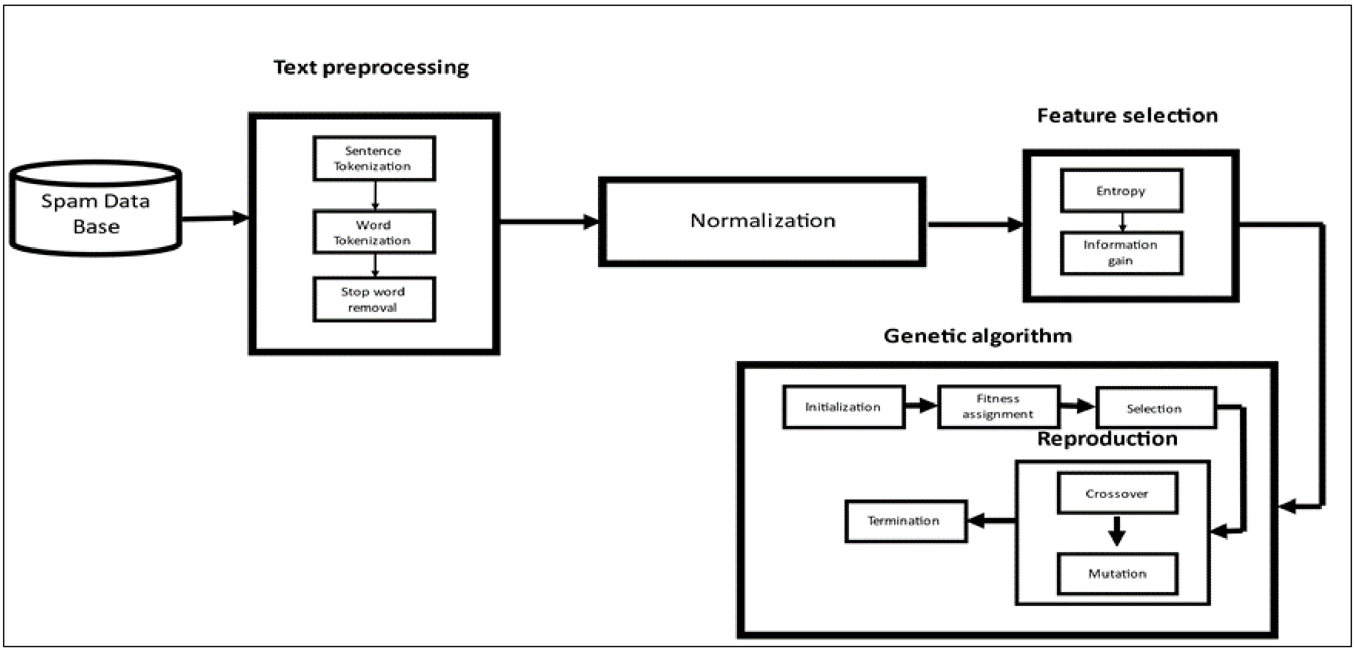
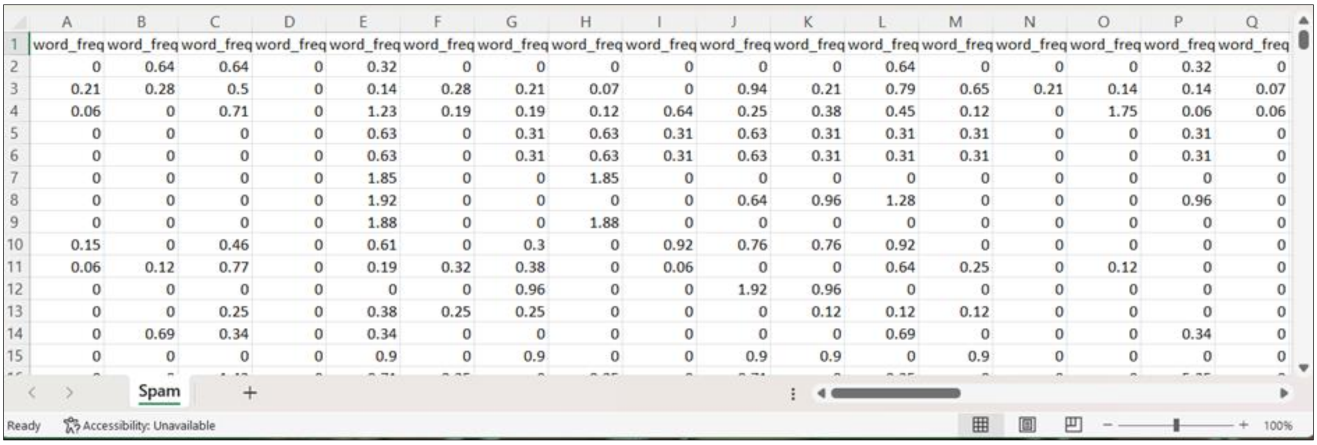


Figure 1 Architecture of Proposed System

3.1. Description of Spam database

The dataset Spam database consists of 4601 instances among those 1813 spam, 2788 are not spam. Besides email, the word or character that are frequently occurred are represented by 57 features. The first 48 features are continuous real numbers and the range is [0-100] like “word_freq_WORD” = words percentage that present in email and correspond to “word”. Here a “WORD” is any string of alphanumeric characters. The next six features are continuous real numbers. The range is [0-100] like “char_freq_CHAR”=characters percentage available in email and belongs to “CHAR”. The next features are continuous real and continuous integers. The last column is a class type, which indicate whether the email is a spam “1” or not “0”.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq	word_freq
2	0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	0	0	0	0.32	0
3	0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	0.65	0.21	0.14	0.14	0.07
4	0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0	1.75	0.06	0.06
5	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0
6	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	0.31	0	0	0.31	0
7	0	0	0	0	1.85	0	0	1.85	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1.92	0	0	0	0	0.64	0.96	1.28	0	0	0	0.96	0
9	0	0	0	0	1.88	0	0	1.88	0	0	0	0	0	0	0	0	0
10	0.15	0	0.46	0	0.61	0	0.3	0	0.92	0.76	0.76	0.92	0	0	0	0	0
11	0.06	0.12	0.77	0	0.19	0.32	0.38	0	0.06	0	0	0.64	0.25	0	0.12	0	0
12	0	0	0	0	0	0	0.96	0	0	1.92	0.96	0	0	0	0	0	0
13	0	0	0.25	0	0.38	0.25	0.25	0	0	0	0.12	0.12	0.12	0	0	0	0
14	0	0.69	0.34	0	0.34	0	0	0	0	0	0	0.69	0	0	0	0.34	0
15	0	0	0	0	0.9	0	0.9	0	0	0.9	0.9	0	0.9	0	0	0	0

Figure 2 Spam Excel database

Normalization, feature selection and genetic algorithm classifier are the three key components of the proposed system. Normalization is used on spam data bases to uniformly distribute the variant frequencies of words across datasets. Feature selection is utilized to pick the best features that improve the model's accuracy, and genetic algorithm classifier is used.

3.2. Normalization

Normalization is applied on a set of values to range from 0 to 1 by using the following equation (1).

$$X = (Y - \min)/(max - min) \text{……………Eq. (1)}$$

where X = new value and Y = old value

3.3. Feature selection

Generally, machine learning uses feature selection [7] to handle high dimensionality issues. This feature selection method chooses a subset of important features while excluding duplicate, irrelevant, and noisy features. Feature selection strategies include filter models, wrapper models and embedded models [8]. Attribute evaluation approaches include gain ratio, information gain, one-R, and symmetrical uncertainty. Entropy is utilized to compute the information gain in this case.

$$\text{Information Gain} = E(R) - E(R|S) = E(S) - E(S|R) \quad \text{………..Eq. (2)}$$

$$\text{Where } E(R|S) = -\sum_{b\epsilon S} P(S) \sum_{a\epsilon R} P(R|S) \log 2 \quad P(R|S) \quad \text{…………..Eq. (3)}$$

And

$$\text{Entropy} = E(R) = -\sum_{y\epsilon Y} P(R) log2P(R) \text{…………..Eq. (4)}$$

3.4. Genetic algorithm

The genetic algorithm [9][10] is the most well-known method in the subject of evolutionary algorithms. This procedure is used to provide high-quality solutions. The genetic algorithm method is both randomly search and robust. The numerous steps in a genetic algorithm are as follows:

3.4.1. Initialization

During the initialization procedure, a set of individuals is formed; this set of individuals is referred to as the population. In this case, each individual is seen as the solution to the given problem. Each individual has a set of parameters, which are referred to as Genes. Genes are linked together to form a string, which is then utilized to create chromosomes.

3.4.2. Fitness assignment

Individuals are evaluated in each iteration based on their fitness function. For each individual, a fitness score will be assigned by the function and that fitness score determines the likelihood of being selected for reproduction. If the fitness score is high, the chances of being selected for reproduction are increased.

If the number of gene types (1) is larger than (X) on the chromosome, the email is spam. Otherwise, not spam, and we discovered that the minimum (X) estimated for the evolution of the fitness function was 3. An experiment was carried out on the spam database, where X is a number. Then, compare the new and old classes to calculate each of the (tp, tn,fn,and fp) and population evaluation.

3.4.3. Selection

Individual selection for offspring reproduction occurs during this phase. For reproduction, we employed the tournament selection process in our work. The best chromosomes are picked at random from a population where the tournament size is 2 i.e., two chromosomes are chosen at random.

3.4.4. Reproduction

A child is created during the reproduction step. The genetic algorithm on the parent population employs crossover and mutation operators during this production step.

Crossover

Crossover is critical in the genetic algorithm's reproduction step. A crossover point within the genes, is randomly selected. The genetic information of two current generation parents is then switched by the crossover operator to make a new offspring. As long as the crossover point is not reached, the genes of the parents are exchanged, and freshly produced children added to the population. One-point, two-point and uniform crossover are the various types of crossovers.

Mutation

Mutation is accomplished by flipping some bits in the chromosomes i.e., the mutation operator is responsible for maintaining population diversity by inserting random genes in the offspring. Flip bit, Gaussian and Exchange mutation are the different types of mutations.

To generate a new generation (child) use uniform crossover with probability p_c. Applying uniform crossover to the indicated bit's index position [11]. Because crossover cannot produce a new generation, the mutation process is used. The mutation procedure is done with probability p_m and by selecting a bit position and converting it to 1 instead of 0 or 0 instead of 1, bit's index location was selected.

Finally, the model with the highest accuracy is recognized as the best model, and we compare the actual class for the testing dataset with the previous class for the model followed by evaluation. More information can be found in below algorithm (3). Both crossover and mutation are used in our experiments. The variation in bit selection location results in a variety of effects.

3.4.5. Termination

After the reproduction phase, a stopping criterion is used to determine when to terminate the algorithm. The algorithm will terminate when the fitness solution threshold is reached, and the end solution will be the best population solution.

3.5. Genetic algorithm Classifier

The proposed algorithm is used to train the database and obtain the optimal model. The performance of the model is then evaluated by using the testing dataset.

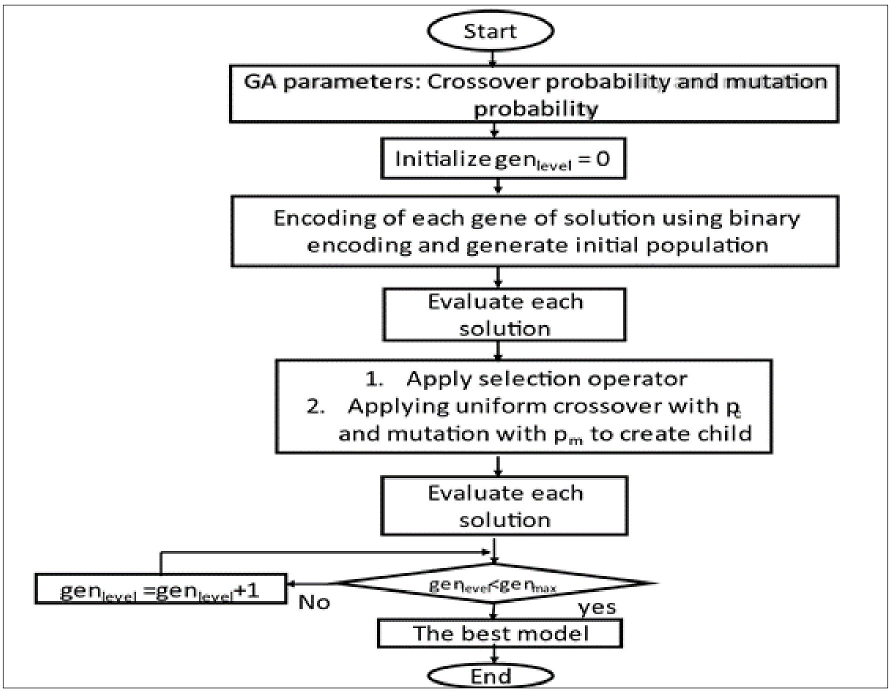

Start
GA parameters: Crossover probability and mutation probability
Initialize $gen_{level} = 0$
Encoding of each gene of solution using binary encoding and generate initial population
Evaluate each solution
1. Apply selection operator
2. Applying uniform crossover with p_c and mutation with p_m to create child
Evaluate each solution
$gen_{level} =gen_{level}+1$
No
$gen_{level}<gen_{max}$
yes
The best model
End

Figure 3 Flowchart of genetic algorithm classifier

To encode each gene solution, we use binary encoding. This involves representing gene values that are greater than or equal to 0.1 as 1, and all other values as 0. Suppose we have a vector of N of a certain length. Let's also consider a population P made up of N individuals, which can be represented as P = P1, P2, … , PN (where N is the overall population size). The population starts growing from a random starting point P0 and continues until it reaches the maximum number of iterations selected. At each GA iteration will employ three major operators: selection, crossover, and mutation. Following the use of the operator, the fitness function utilized in the proposed method for assessing the quality of genetic algorithm solutions.

4. Evaluation measures

To define the evaluation measures, a matrix with two classes, spam and not spam, is employed.

	True Class	
Texting Class	**Spam**	**Not spam**
Spam	TP	FP
Not Spam	FN	TN

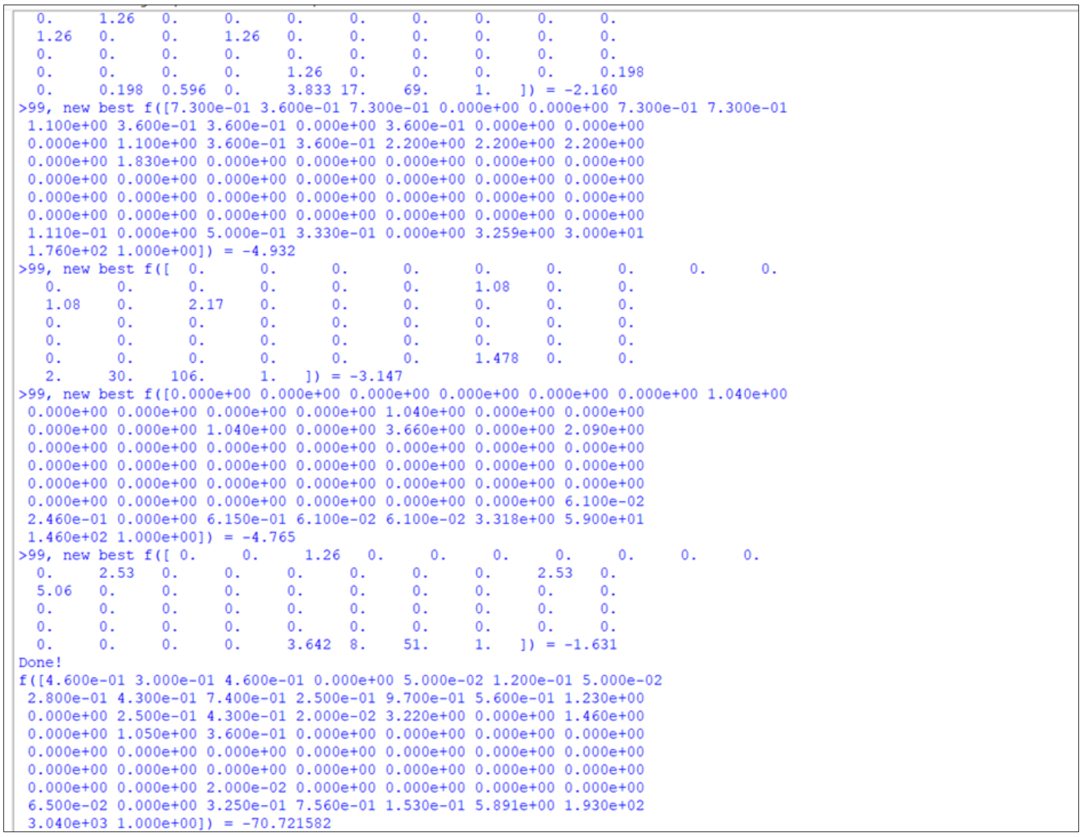
If e-mails in both the testing and true classes contain spam, the result will be TP (True Positive). If e-mails in both the testing and true classes are not spam, the result will be TN (True Negative) [12]. If an e-mail is not spam in the true class but is spam in the testing class, it is classified as FP (False Positive). If an e-mail is spam in the true class but not in the testing class, the result is FN (False Negative).

By using the above evaluation measures, accuracy will be calculated by using the formula,

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN)\text{...........Eq. (5)}$$

5. Experimental Results

During the experiment, a Spam database of 4601 emails is used. There are 1813 spam mails and 2788 ham mails among them. In addition to e-mail class type 57 features, which are expressed as specific words or character that appear frequently in an email. The chromosome size is fixed at 8 bits, the crossover probability is 0.3, and this crossover probability has a vast range of values up to 1.0. The mutation probability under consideration is 1.0. Taking these factors into account, the suggested approach of genetic algorithm classifier yields good accuracy, as shown in the figure below.



```
  0.      1.26    0.      0.      0.      0.      0.      0.      0.      0.
  1.26    0.      0.      1.26    0.      0.      0.      0.      0.      0.
  0.      0.      0.      0.      0.      0.      0.      0.      0.      0.
  0.      0.      0.      0.      1.26    0.      0.      0.      0.      0.198
  0.      0.198   0.596   0.      3.833 17.     69.      1.    ]) = -2.160
>99, new best f([7.300e-01 3.600e-01 7.300e-01 0.000e+00 0.000e+00 7.300e-01 7.300e-01
 1.100e+00 3.600e-01 3.600e-01 0.000e+00 3.600e-01 0.000e+00 0.000e+00
 0.000e+00 1.100e+00 3.600e-01 3.600e-01 2.200e+00 2.200e+00 2.200e+00
 0.000e+00 1.830e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 1.110e-01 0.000e+00 5.000e-01 3.330e-01 0.000e+00 3.259e+00 3.000e+01
 1.760e+02 1.000e+00]) = -4.932
>99, new best f([  0.       0.       0.       0.       0.       0.       0.       0.       0.
   0.       0.       0.       0.       0.       0.       1.08     0.       0.
   1.08     0.       2.17     0.       0.       0.       0.       0.       0.
   0.       0.       0.       0.       0.       0.       0.       0.       0.
   0.       0.       0.       0.       0.       0.       0.       0.       0.
   0.       0.       0.       0.       0.       0.       1.478    0.       0.
   2.      30.     106.       1.    ]) = -3.147
>99, new best f([0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 1.040e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 1.040e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 1.040e+00 0.000e+00 3.660e+00 0.000e+00 2.090e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 6.100e-02
 2.460e-01 0.000e+00 6.150e-01 6.100e-02 6.100e-02 3.318e+00 5.900e+01
 1.460e+02 1.000e+00]) = -4.765
>99, new best f([ 0.      0.      1.26    0.      0.      0.      0.      0.      0.      0.
  0.      2.53    0.      0.      0.      0.      0.      0.      2.53    0.
  5.06    0.      0.      0.      0.      0.      0.      0.      0.      0.
  0.      0.      0.      0.      0.      0.      0.      0.      0.      0.
  0.      0.      0.      0.      0.      0.      0.      0.      0.      0.
  0.      0.      0.      0.      3.642   8.     51.      1.    ]) = -1.631
Done!
f([4.600e-01 3.000e-01 4.600e-01 0.000e+00 5.000e-02 1.200e-01 5.000e-02
 2.800e-01 4.300e-01 7.400e-01 2.500e-01 9.700e-01 5.600e-01 1.230e+00
 0.000e+00 2.500e-01 4.300e-01 2.000e-02 3.220e+00 0.000e+00 1.460e+00
 0.000e+00 1.050e+00 3.600e-01 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 0.000e+00 0.000e+00 2.000e-02 0.000e+00 0.000e+00 0.000e+00 0.000e+00
 6.500e-02 0.000e+00 3.250e-01 7.560e-01 1.530e-01 5.891e+00 1.930e+02
 3.040e+03 1.000e+00]) = -70.721582
```

Figure 4 Accuracy result

6. Conclusions

The identification of spam email is crucial in the design of secure e-mail systems. To improve email spam filtering, the suggested approach uses feature selection which is subsequently followed by the genetic algorithm classifier to choose important features from the spam database. The selection of the fitness function involves careful experimentation with various data sets. The effectiveness of the process is reliant on both the dataset and the parameters of the genetic algorithm. The model's performance is then assessed by using the testing dataset. With all 57 features, a crossover probability, and a mutation probability, the proposed algorithm obtained 70.7% of accuracy. In addition, the suggested algorithm has the potential to be integrated with various classification algorithms in the future to improve the results.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Jose R. Mendez, et.al., (2019), A new semantic-based feature selection method for small filtering, Applied Soft Computing, Vol. 76, pp: 89-104.

[2] Sharifah MD Yasin and Iqbal Hadi Azmi (2023), Email spam filtering technique: Challenges and solutions, Journal of Theoretical and Applied Information Technology, Vol. 10, No. 13.

[3] Jitendra Nath Shrivastava and Maringanti Hima Bindu (2013), E-mail classification using genetic algorithm with heuristic fitness function, International Journal of Computer Trends and Technology (IJCTT) – Vol. 4.

[4] Mandeep Singh and Prashant Sahai Saxena (2017), E-mail classification using fuzzy fusion of average and probabilistic methods, International Journal of Applied Engineering Research, Vol.12, Number 18.

[5] Mandeep Choudhary and V.S.Dhaka (2015), E-mail spam filtering using genetic algorithm: a deeper analysis, International Journal of Computer Science and Information Technologies, Vol.6, Issue 3.

[6] Mandeep Choudhary and V.S. Dhaka (2015), Automatic e-mails classification using genetic algorithm, National Conference on Cloud Computing & Big Data.

[7] Sorayya Mirzapour Kalaibar and Seyed Naser Razavi (2014), Spam filtering by using genetic based feature selection, International Journal of Computer Applications Technology and Research, Vol. 3, Issue 12, pp: 839-843.

[8] Z. Hassani, et.al., (2020), A classification method for E-mail spam using a hybrid approach for feature selection optimization, Journal of Sciences, Islamic Republic of Iran, Vol. 31, Issue 2, pp:165-173.

[9] Pronaya Bhattacharya and Arunendra Singh (2020), E-mail spam filtering using genetic algorithm based on probabilistic weights and words count, International Journal of Integrated Engineering, Vol. 12, No. 1, pp: 40-49.

[10] Jitendra Nath Shrivastava and Maringanti Hima Bindu (2014), E-mail spam filtering using adaptive genetic algorithm, International Journal of Intelligent Systems and Applications, pp: 54-60, DOI: 10.5815/ijisa.2014.02.07.

[11] Dr. Soukaena H. hashem, Dr. Ekhlas Khalaf Gbashi (2018), Spam classification using Genetic algorithm, Iraqi Journal of Information Technology, Vol. 9, Issue 2.

[12] Shubham Mathur, Aakash Purohit (2022), Performance evaluation of machine learning algorithms on textual datasets for spam email classification, International Journal for Research in Applied Science & Engineering Technology, Vol. 10, Issue VII.

Author's short profile

	B. Aruna kumari pursuing her PhD in the field of Network Security under the supervision of Dr. C. Nagaraju, Professor of Computer Science and Engineering, YSR Engineering College of Yogi Vemana University, Proddatur, Kadapa, Andhra Pradesh. She received her B.Tech in Computer Science and Engineering at CBIT, Proddatur. M.Tech in Computer Science and Engineering at JNTUACE, Pulivendula. She has 4 years of teaching experience as an Assistant Professor in the department of Computer Science and Engineering at AITS, Rajampet and VITS, Proddatur. She has published six papers in various national and international journals. She has attended two conferences, three Faculty development programs and two workshops.
	C. Naga Raju is currently working as professor in the Department of Computer Science and Engineering at YSR Engineering College of Yogi Vemana University, Proddatur, YSR Kadapa District, Andhra Pradesh, India. He received his B. Tech. In Computer Science and Engineering from J.N.T. University, Anantapur, and M. Tech. In Computer Science and Engineering from J.N.T. University Hyderabad and Ph.D. in Digital Image Processing from J.N.T. University Hyderabad. He has got 25 years of teaching experience. He received research excellence award, teaching excellence award, outstanding scientist award and Rayalaseema Vidhyaratna award for his credit. He wrote text books on programming in C, Network Security and Digital Image processing. He has got one patent on his research work. He has completed two DST funding projects worth of forty lakhs. Under his guidance seven PhDs are completed and five scholars are pursuing. He has published ninety-He has published ninety-eight research papers in various National and International Journals and thirty research papers in various National and International Conferences. He has attended twenty seminars and workshops. He delivered more than seventy keynote addresses. He is member of various professional societies like IEEE, ISTE and CSI.