moved to the next generation. These processes continue until an optimal solution is achieved. Fitness function is also applied in the process. In their experiment, they considered 2448 emails out of which their system was able to detect 1346 mails as spam and the remaining 1102 mails as ham mails. Their results show that their system is 84% effective.

Sorayya and Seyed (2014) performed spam filtering using genetic based feature selection technique. They tried to evaluate spam detection in legal electronica letters, and their effect on several machine learning algorithms through presenting a feature selection method based on genetic algorithm. Their results indicate that feature selection by GA technique improves email spam classification.

Shahamat and Pouyan, (2014) performed feature selection using genetic algorithm for Classification of Schizophrenic using functional magnetic resonance imaging (fMRI) Data. They proposed a new method for classifying subjects into schizophrenia and control groups by making use of functional magnetic resonance imaging (fMRI) data. During the stage of processing, they reduced the number of fMRI time points using principal component analysis (PCA). For further data analysis, they made use of independent component analysis (ICA). Furthermore, local binary patterns (LBP) technique was used for feature extraction for the ICs. This made them able to estimate independent components (ICs) of PCA results. They used genetic algorithm for feature selection and they went further to get a set of features with large discrimination power. The test subject they used was classified into schizophrenia or control group. This they were able to do using a Euclidean distance-based classifier and a majority vote method. The result of their experiment showed that their proposed

method has an acceptable accuracy and is comparable to other state-of-the- art work.

Cerrada *et al.* (2015) performed multi-stage feature selection by using genetic algorithms for fault diagnosis in gearboxes based on vibration signal. They proposed a multi-stage feature selection technique for selecting the best set of condition parameters on the time, frequency and time frequency domains extracted from vibration signals for fault diagnosis purposes in gearboxes. They used three methods namely; wrapper, filtering and embedded methods to eliminate features that are not relevant. At each stage, they selected the best features from a subset of candidate features that improve classification metrics on the diagnosis model.

Sung *et al.* (2015) presented a feature selection method based on genetic algorithm for efficient of text clustering and text classification. They focused on selecting a set of optimized features from big data. They used Genetic Algorithm to extract these features as desired according to term importance calculated by the equation found. Their study revolved around feature selection method to lower computational complexity and to increase analytical performance. They were able to achieve the design of a new Genetic Algorithm to extract features in text mining. They also conducted clustering experiments on a set of spam mail documents to verify and to improve feature selection performance and they found that the proposal of the Feature Selection Method based on Genetic Algorithm (FSGA) for Efficient of Text Clustering showed better performance of Text Clustering and Classification than using all of features.

Priyanka and Kavita, (2016) performed feature selection using genetic algorithm and classification using weka for ovarian cancer. They investigated the performance of different classification methods on clinical

data. They used Genetic Algorithm to select relevant features before applying classification algorithm. The tool they used for the classification was Weka tool. They evaluated and investigated five selected and classification algorithms based on Weka. Their results showed that the best algorithm in WEKA is Bayesnet classifier with an accuracy of 61.57% because it takes 0.1 seconds for classifying the dataset.

Dahiya and Sangwan (2018) reviewed literatures on genetic algorithm. They elaborated on the fact that to use genetic algorithm, one has to encode possible model behaviours into genes after which the models are rated, and allowed to mate and breed based on their fitnesses. They also reviewed different problems that can be solved by genetic algorithm.

3. Methodology

The Methodology for the Proposed System Design is Object-Oriented Analysis and Design Methodology (OOADM). Object-Oriented Analysis and Design Methodology is a system approach to the analysis, and design of information systems from the point of view of software objects, and their interactions with the system, and its associated environment. In this methodology, the analysis tries to abstract components and assign names to them and then groups their operations into class abstracts. These class abstracts are then designed by identifying their roles in the system, and the actors that execute those roles. The data used in the execution of the roles and the operations of the roles are then bundled together as class designs which are presented using UML class designs. The system was divided into two different parts. The first part took care of the training while the second took care of the testing. The dataset used for the training and testing were downloaded from the internet. For the training, a fitness function is adopted so as to fit the data into it, and also to be able to

assign fitnesses to the emails for easy selection, and finally classification. From the architecture, it can be seen that the emails in the email corpus are pre-processed so as to remove redundant data. From there, they are assigned fitnesses based of the fitness function, and then the best fit individuals are selected, and trained. Thereafter, they are cross validated to check for accuracy, and to avoid overfitting. After that, the testing set is used to test the system, and classification of the mails into ham, and spam is achieved. The detailed description of the methodology is shown in figure 1.

3.1 Components of the System

The components of the Proposed System are:

A. **Email Corpus:** this is the database of the emails that will be used for training, and testing the model.

B. **Data Pre-processing:** at this stage, the emails from the corpus are cleaned to remove redundant data, and those that are not fit for purpose.

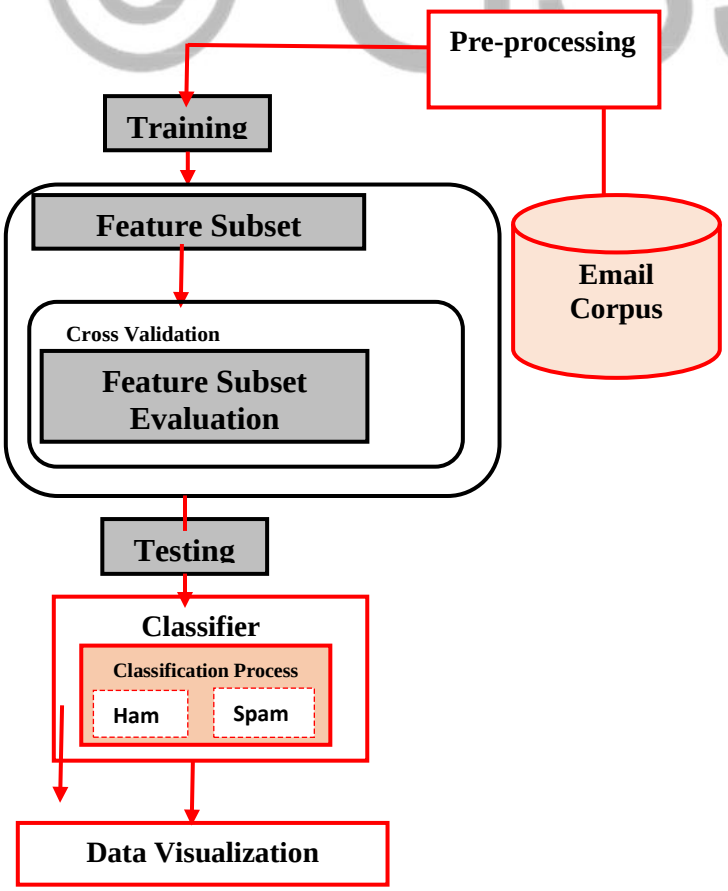

Figure 1. Detailed description of the methodology

C. Feature Selection:

Genetic Algorithm

(i) Fitness assignment: features of the emails are selected for training, and at this stage, the individuals are trained and their errors calculated. As individuals with their own chromosomes are allowed to operate in the environment, they are assigned fitnesses based on their performances.

(ii) The selection process: the selection of individuals for gene recombination is a mandatory step in genetic algorithm. Genetic algorithm is used to select the best fit individuals that will be used for reproduction.

D. Genetic operators:

(a) Cross over: the genes (selected individuals) are crossed over to get individuals of the next generation. This is done until an optimal result is achieved. In cross over, there is a high chance of generating offsprings that are similar to their ancestors which leads to low diversity.

(b) Mutation: due to the low diversity which is as a result of cross over, the offsprings are mutated by changing the value of some features of the offsprings at random. This is to create gene diversity. This is repeated until the generation does not significantly differ from the previous one.

3.2 Genetic Algorithm to be implemented for Classification

1. Start

2. Initialize the Population

3. Initialize the program size

4. Define the fitness fi of an individual program corresponds to the number of hits and is evaluated by specific formula:

5. Run a tournament to compare four programs randomly out of the population of programs

6. Compare them and pick two winners and two losers based on fitness

7. a) Copy the two winners and replace the losers

b) With Crossover frequency, crossover the copies of the winners

c) With Mutation frequency, mutate the one of the programs resulting from performing step 7(a)

d) With Mutation frequency, mutate the other of the programs resulting from performing step 7(a)

8. Repeat through step 5 till termination criteria are matched

4. Experiments and results

In order to experiment and evaluate the performance of the system, data from the dataset downloaded online was used to perform some iterations. The iteration was carried out over 50 generations. It was discovered that the system was able to effectively classify the spam email as shown in table 1.

Table 1: Result of the iterations

GENERATION/ITERATION	BEST SOLUTION SO FAR	WORST SOLUTION SO FAR	BEST SOLUTION INDEX SO FAR	MESSAGE
1	0.0625	4.6875	0	Spam
2	0.0625	4.6875	0	Spam
3	0.0625	4.6875	0	Spam
4	0.0625	4.6875	0	Spam
5	0.0625	4.6875	0	Spam
6	0.0625	4.6875	31	Spam
7	0.0625	4.6875	0	Spam
8	0.0625	4.6875	0	Spam
9	0.0625	4.6875	0	Spam
10	0.0625	4.6875	0	Spam
11	0.0625	4.6875	41	Spam
12	0.125	4.6875	0	spam
13	0.0625	4.6875	0	spam
14	0.0625	4.6875	45	spam
15	0.0625	4.6875	0	spam
16	0.0625	0.9375	0	spam
17	0.0625	4.6875	0	spam
18	0.0625	4.6875	0	spam

19	0.0625	4.6875	49	spam
20	0.0625	0.98375	0	spam
21	0.0625	4.6875	0	spam
22	0.0625	4.6875	0	spam
23	0.0625	4.6875	0	spam
24	0.0625	4.6875	0	spam
25	0.0625	4.6875	31	spam
26	0.0625	4.6875	0	spam
27	0.0625	4.6875	0	spam
28	0.0625	4.6875	0	spam
29	0.0625	4.6875	0	spam
30	0.0625	4.6875	41	spam
30	0.125	4.6875	0	ham
30	0.0625	4.6875	0	spam
30	0.0625	4.6875	45	spam
30	0.0625	4.6875	0	spam
31	0.0625	0.9375	0	spam
32	0.0625	4.6875	0	spam
33	0.0625	4.6875	0	spam
30	0.0625	4.6875	49	spam
30	0.0625	0.98375	0	spam
30	0.0625	4.6875	461	spam
30	0.0625	4.6875	37	spam
30	0.0625	4.6875	50	spam
34	0.0625	4.6875	0	spam
35	0.0625	4.6875	0	spam
36	0.0625	4.6875	0	spam
37	0.0625	0.9375	0	spam
38	0.0625	0.9375	0	spam
39	0.0625	0.9375	0	spam
40	0.0625	4.6875	0	spam
41	0.0625	4.6875	461	spam
42	0.0625	4.6875	37	spam
43	0.0625	4.6875	50	spam
44	0.0625	4.6875	0	spam
45	0.0625	4.6875	0	spam
46	0.0625	4.6875	0	spam
47	0.0625	0.9375	0	spam
48	0.0625	0.9375	0	spam
48	0.0625	0.9375	0	spam
50	0.0625	4.6875	0	spam

Table 1 shows the classification of spam email. It can be observed that at generation 30, the email was classified as ham which shows that the system is not a perfect one. The spam email was labeled 0 while the ham was labeled 1 as shown in the table below. The data is iterated over different generations, and the result of the previous

generation is moved into the next generation to continue the iteration. The best solution so far, and the worst solution so far are recorded. The best solution index so far is also recorded. Although the best individuals are selected to move into the next generation to continue to cross over and mutate, however, worst individuals are not discarded because they may eventually become relevant, and produce good offsprings in later generations. The accuracy of the system was calculated to be 98%. The accuracy and generation performance of the system is shown in table 2 while the graph of the iteration is also shown in figure 2.

Table 2: Accuracy and Generation Performance of the Proposed System

GENERATION/ITERATION	Accuracy
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
30	0
30	1
30	1
30	1

31	1
32	1
33	1
30	1
30	1
30	1
30	1
30	1
34	1
35	1
36	1
37	1
38	1
39	1
40	1
41	1
42	1
43	1
44	1
45	1
46	1
47	1
48	1
49	1
50	1

Accuracy of the proposed system = number of spam mails * 100

Total number of emails

= 49 * 100 = 98%

50

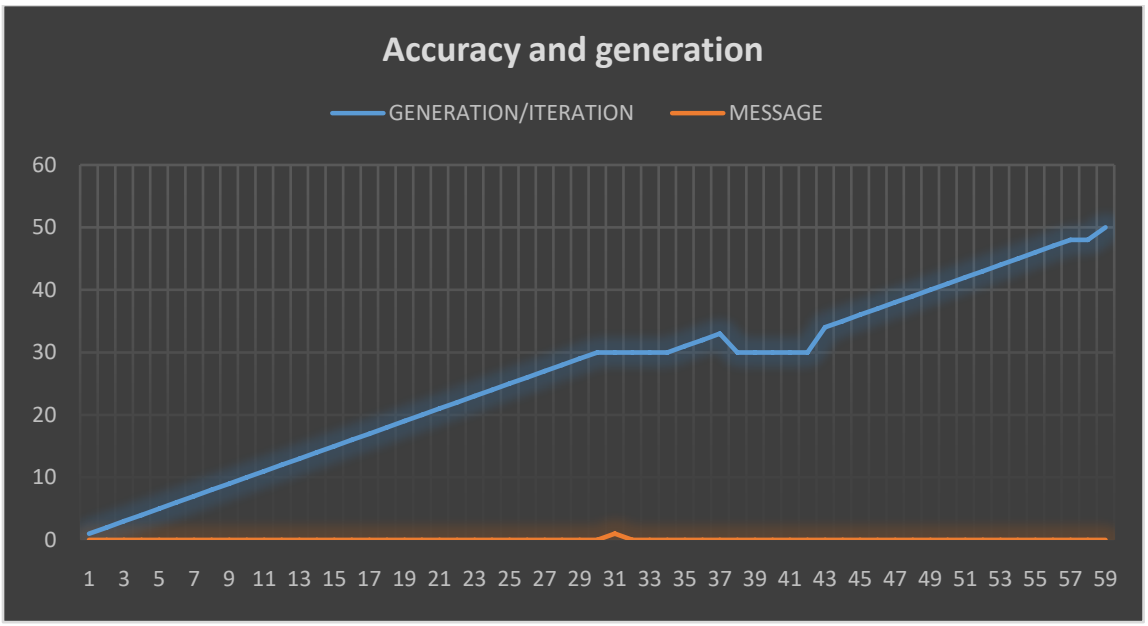

Figure 2: Accuracy and Generation graph of the Proposed System

5. Conclusion

In this paper, spam email selection and classification using genetic algorithm was proposed. The proposed approach is in two stages. In the first stage which is the training stage, genetic algorithm is used to select the best fit individuals based on their relative fitnesses. These best fit individuals are then allowed to cross over and mutate over different generations. They are then classified into spam or ham. During the testing stage, an incoming email goes straight to the testing phase, and gets classified by the genetic algorithm. A total of 2226 emails were taken from enron email dataset to train the system out of which 1022 0f them were spam emails. The accuracy of the system was calculated to be 98%. Our experimental results show that genetic algorithm can effectively be used to classify emails.

REFERENCES

Amira S. A. A., Ahmad T. A., Mostafa, A. S., Aboul, E.H and E.H. Sanaa. (2013) Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation. Proceedings of the 2013 Federated

Conference on Computer Science and Information Systems *769-774.*

Asti, D. I., Retantyo, W., Sri, H., and S. Endang (2016). Determination of Selection Method in Genetic Algorithm for Land Suitability. Presented at MTEC Web of Conferences *58 – 65.*

Cerrada, M., Sanchez, R., Cabrera, D. and Zurita, G., C. Li (2015). Multi-stage Feature Selection by using Genetic Algorithms for Fault Diagnosis in Gearboxes based on Vibration Signal Sensors. 15(9), *23903-23926*.

Dahiya, R.M. and Sangwan, S. (2018). Literature Review on Genetic Algorithm. International Journal of Research, 5 (16), *1142-1146.*

Noraini, M. R. and G. John (2011). Genetic Algorithm Performance with Different Selection Strategies in Solving TSP. Proceedings of the World Congress on Engineering (2) WCE, *1-6.*

Pei, M., Goodman E.D. and Punch W.F (1998). Feature Extraction Using Genetic Algorithms. Genetic Algorithms Research and Applications Group Michigan State University, 2325 Engineering Building, East Lansing, *1-14.*

Priyanka, K. and B. Kavita (2016). Feature Selection Using Genetic Algorithm and Classification using Weka for Ovarian Cancer. International Journal of Computer Science and Information Technologies, 7 (1) *194-196.*

Shahamat, H. and A.A. Pouyan (2014). Feature Selection using Genetic Algorithm for Classification of Schizophrenia using fMRI Data. Journal of ArtificiaI Intelligence and Data Mining 3 (1), *30-37*.

Sushmita, M., Sankar K. and M. Pabitra (2002). Data Mining in Soft Computing Framework: A Survey. IEEE Transactions on Neural Networks, 13 (1), *3-14.*

Shrivastava, J.N. and M.H. Bindu (2013). E-mail Classification Using Genetic Algorithm with Heuristic Fitness Function, International Journal of Computer Trends and Technology, 4 (8), *2956-2961*.

Sorayya M.K. and Seyed N.S. (2014) Spam filtering by using Genetic based Feature Selection, International Journal of Computer Applications Technology and Research, 3 (12), *839-843.*

Sung-Sam, H., Wanhee L. and H. Myung-Mook (2015). The Feature Selection Method Based on Genetic Algorithm for Efficient of Text Clustering and Text Classification. International Journal of Advance Soft Computing. Appl, 7, (1), ISSN 2074-8523 *22-40.*

Tan, F., Fu X., Zhang Y. and A.G. Bourgeois (2008). A Genetic Algorithm-Based Method for Feature Subset Selection. Soft Computing, 1, (12) *111–120.*

Venugopal, K.R., Srinivasa, K.G. and L.M. Patnaik. (2009) Soft Computing for Data Mining Applications ISBN 978-3-642-00192-5 e-ISBN 978-3-642-00193-2 DOI10.1007/978-3-642-00193-2 Studies in Computational Intelligence ISSN 1860949X Springer-Verlag Berlin Heidelberg *13-14.*