An Efficient AIS based Feature Extraction Techniques for Spam Detection

Mafaz Mohsin Khalil Al-Anezi, PhD
Computer Sciences
Department, College of
Computer Sciences & Mathematics,
Mosul University, Mosul, 964, Iraq

ABSTRACT
Simultaneously with the development of networks, and with the increasing volume of unsolicited bulk e-mail especially advertising, indiscriminately has generated a need for reliable anti-spam filters. The problem for the traditional method of spam filtering cannot effectively identify the unknown and variation characteristics, therefore recently the researchers look at the artificial immune system exists diversity, immune memory, adaptive and self learning ability. The spam detection model describes an e-mail filtering is accomplished by extracting the characteristics of spam and ham (legitimate e-mail messages that is generally desired and isn't considered spam) that is been acquired from trained data set by feature extraction techniques. These techniques allowed to select subset of relevant, non redundant and most contributing features to have an added benefit in accuracy and reduced time complexity. The extracted features of spam and ham are then make a two types of antigen detectors, to enter then in series of cloning and mutation immune operations to built an immune memory of spam and ham. The experimental result confirms that the proposed model has a very high detection rate reach at 1 and a very low false alarm rate reach at 0 when using low numbers of feature extraction.

General Terms
Artificial Immune System (AIS), Feature Extraction Techniques and Security.

Keywords
Email, Spam, Ham (legitimate messages), Clonal selection, Information Gain, LDA, PCA

1. INTRODUCTION
The word "spam" is used to indicate the electronic equivalent of junk email. Exact definitions will vary, but it typically covers a range of unsolicited and undesired advertisements and bulk email messages.

The most common communication in the internet is using email communication. With the vast growth in email and its popularity unsolicited e-mail (spam) also emerged very quickly with almost 90% of all email messages. i.e., over 120 billion of these messages are sent each day. The cost of sending these e-mails is very close to zero being easy to reach a high number of potential consumers. In this context, spam consumes resources; time spent reading unwanted messages, bandwidth, CPU, disk, being also used to spread malicious content [1].

The email system design can easily be exploited by spammers who send inaccurate information. All email on the Internet is sent via a protocol called Simple Mail Transfer Protocol (SMTP). SMTP is designed to capture information about the route that an email message travels from its sender to its recipient. In actuality, the SMTP protocol provides no security, email is not private, it can be altered en route, and there is no way to validate the identity of the email source. In other words ,when a user receives an email message, there is no way to tell who sent the email and who has seen it. The lack of security in SMTP, and specifically the lack of reliable information identifying the email source, is regularly exploited by spammers and allows for considerable fraud on the Internet (such as identity theft or "phishing") [1].

Spam even provides various kinds of attacks and distributed harmful content or data such as viruses, worms, Trojan horses and other malicious code. Several technical solutions are available for dealing with these issues like commercial and open-source products [1].

Spam classification has contained the different machine learning classification. In supervised learning process text classification is very popular. In supervised learning process a task is assign to the text data or document and then classifies this text data according to predefined categories or classes according to their contents. According categories of their contents the data is automatically classified. Now days there are different types of algorithm are present to deal with automatic text classification [2].

There are two types of spam filtering are supervised and unsupervised. But the most different classifier method for detecting the spam mails are [2,3]:

I. Based on Non-machine learning:
- K-means Clustering Method
- Black list/White list
- Signature

II. Based on Machine learning:
- Support Vector Machine (SVM)
- Artificial Neural Network (ANN)
- Negative Selection Algorithm
- Naïve Bayesian Classifier
- Decision Tree
- Nearest Neighbor (NN)

Simple techniques including white and black list methods fail to categorize messages without user intervention. Even worse, a contacts inserted into the black list may send legitimate messages beside spam, e.g., a bank may send a spam message including new credit opportunities and a legitimate message containing online banking password as well. In this case,

smarter methods such as content based classification are needed.

One of the solution for the spam problem is the "machine learning" method. The ability of a machine to improve its performance based on the previous results is known as machine learning. In machine learning the existing data set training is used to differentiate between the spam & non spam emails. Feature extraction is the major concept used in machine learning. It extracts the feature from the email & then give the result whether it is spam or not & it takes the help of training & learning phase [3].

2. PREVIOUS WORKS

OzarkarP.andPatwardhanM. [4] used the spam dataset because it is possible to have large number of training instances. Based on this fact, they have made use of Random Forest and Partial Decision Trees algorithms to classify spam vs. non-spam emails. As a preprocessing step they have used feature selection methods such as Chi-square, Information gain, Gain ratio, Symmetrical uncertainty, Relief, One R and Correlation. So after using 70% of the feature set extracted, for spam base data set, the training accuracy is (99.918%) whereas the computation time reduced by 20%.

Idris I. and Selamat A. [5] proposed a new improved model that combines negative selection algorithm (NSA) with particle swarm optimization (PSO) has been proposed and implemented. The new model is called swarm negative selection algorithm (SNSA). The implementation of PSO with its fitness function improved the detector generation phase of NSA. The empirical report shows the superiority of the proposed SNSA improved model over the NSA model. At 8000 generated detectors with threshold value of 0.4, accuracy for negative selection algorithm is 68.863% while improved swarm negative selection algorithm is at 82.69%.

3. ARTIFICIAL IMMUNE SYSTEM

Artificial Intelligence System (AIS) is a research area which is used to build intelligence models & it takes the inspiration from Biological Immune System (BIS). BIS have several properties which consists of distributed detection, noise tolerance & reinforcement learning. Considering the immune processes related to BIS many AIS models have been developed to solve engineering problems. Examples are negative selection, clonal selection, immune network model & danger theory algorithm & these models are applied on real world problems which are pattern recognition, data mining, spam filtering & computer security. The main function of BIS is to protect the body from molecules which are known as antigens. The feature of BIS is that it has the pattern recognition capability which can be used to differentiate between foreign cells entering in the body (non-self or antigen) & the body cells (self) [3].

AIS is inspired by the human immune system which is a highly evolved, parallel and distributed adaptive system that exhibits the following strengths: immune recognition, reinforcement learning, feature extraction, immune memory, diversity and robustness. The artificial immune system (AIS) combines these strengths and has been gaining significant attention due to its powerful adaptive learning and memory capabilities.

The main search power in AIS relies on the mutation operator and hence, the efficiency deciding factor of this technique. The steps in AIS are as follows [6]:

1. Initialization of antibodies (potential solutions to the problem). Antigens represent the value of the objective function f(x) to be optimized.
2. Cloning where the affinity or fitness of each antibody is determined. Based on this fitness the antibodies are cloned that is the best will be cloned the most. The number of clones generated from the n selected antibodies is given byequation (1):

$$Nc = \Sigma \text{ round } (\beta * j/i), i = 1,2....n, \quad (1)$$

Where Nc is the total number of clones, β is a multiplier factor and j is the population size of the antibodies.

3. Hypermutation: The clones are then subjected to a hyper mutation process in which the clones are mutated in inverse proportion to their affinity; the best antibody's clones are mutated lesser and worst antibody's clones are mutated most. The clones are then evaluated along with their original antibodies out of which the best N antibodies are selected for the next iteration. The mutation can be uniform, Gaussian or exponential.

4. CLONAL SELECTION ALGORITHM CLONA

Clonal selection and expansion is the most accepted theory used to explain how the immune system copes with the antigens. In brief, the Clonal selection theory states that when antigens invade an organism, a subset of the immune cells capable of recognizing these antigens proliferate and differentiate into active or memory cells. The fittest clones are those, which produce antibodies that bind to antigen best (with highest affinity). The main steps of Clonal selection algorithm can be summarized as follows [7]:

Algorithm 1: Clonal selection

Step 1: For each antibody element

Step 2: Determine its affinity with the antigen presented

Step 3: Select a number of high affinity elements and reproduce (clone) them proportionally to their affinity.

5. E-MAIL SPAM DATASET

The dataset used for our experiment is spam base [8]. It is a multivariate and its contains 4601 instances, the attribute characteristics are integer or real. The last attribute of 'spam base. Data' denotes whether the e-mail was considered spam (1) or not (0). Most of the attributes indicate the frequency of spam related term occurrences. The first 48 set of attributes (1–48) give tf-idf (term frequency and inverse document frequency) values for spam related words, whereas the next 6 attributes (49-54) provide tf-idf values for spam related terms. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters, capital_ run_ length_average, capital_run_length_longest and capital_ run_ length_ total. Thus, our dataset has in total 57 attributes serving as an input features for spam detection and the last attribute represent the class (spam/non-spam).

6. FEATURES RANKING AND SUBSET SELECTION

Dimensionality reduction and feature selection is an important aspect of electroencephalography based event related potential detection systems such as brain computer interfaces [9].

Feature ranking further help us to:

1. Remove irrelevant features, which might be misleading the classifier decreasing the classifier interpretability by reducing generalization by increasing over fitting.
2. Remove redundant features, which provide no additional information than the other set of features, unnecessarily decreasing the efficiency of the classifier.
3. Selecting high rank features, which may not affect much as far as improving precision and recall is concerned; but reduces time complexity drastically. Selection of such high rank features reduces the dimensionality feature space of the domain. It speeds up the classifier there of improving the performance and increasing the comprehensibility of the classification result [4].

From the above defined feature vector of total 58 features, feature ranking and selection algorithms are used to select the subset of features. The given set of features are rankedusing the following distinct approaches.

6.1 Information Gain

Information Gain is the expected reduction in entropy caused by partitioning the examples according to a given attribute. Information gain is a symmetrical measure that is, the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y. The entropy of Y is given by equation (2): [4]

$$HY = -PY log2(PY) y \in Y \qquad (2)$$

If the observed values of Y in the training data are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. Equation (3) gives the entropy of Y after observing X

$$H\,Y = -(x)\,P\,y\,x\,log2(P\,y\,x)\,y \in Y x \in X \qquad (3)$$

The amount by which the entropy of Y decreases reflects additional information about Y provided by X and is called the information gain or alternatively, mutual information [4]. Information gain is given by equation (4):

$$Gain = H\,Y + H\,Y\,X$$
$$= H\,X + H\,X\,Y$$
$$= \mathrm{H\,Y} + \mathrm{H\,X} - (X,Y) \qquad (4)$$

6.2 Principal Component Analysis (PCA)

Perhaps PCA is one of the most commonly used dimensionality reduction methods. PCA seeks the linear combinations of the multivariate data that capture a maximum amount of variance. However, the projections that PCA seeks are not necessarily related to class labels; hence may not be optimal for classification problems [9].

Contributions to Principal Component Analysis is technique used for feature extraction, data used in intrusion detection problem are high dimensional in nature. It is desirable to reduce the dimensionality of the data for easy exploration and further analysis. The PCA is often used for this purpose [10].

The mathematics behind principle component analysis is statistics and is hinged behind standard deviation, eigenvalues and eigenvectors. The entire subject of statistics is based around the idea that you have this big set of data, and you want to analyze that set in terms of the relationships between the individual points in that data set [11].

PCA is concerned with explaining the variance-covariance structure of a set of variables through a few new variables. If there are M features in each datum and there are N data which is represented by x_{11}, x_{12}, x_{13}…..x_{1M}. x_{21},. x_{22}, x_{23}….x_{2M}. Similarly the final datum can be represented by xN1, xN2, xN3….xNM. The matrix A= [$Ø_1$, $Ø_2$,....., $Ø_M$] (N×M matrix) [10].

The sample covariance matrix C of the data set is defined as by equation (5):

$$C = \frac{1}{M}\sum_{i=1}^{M} \emptyset_i \, \emptyset_i{}^T \qquad (5)$$

Theeigenvalues(λ_1,λ_2, ...λ_N) and eignvectors (u_1, u_2,....., u_N) of covariance matrix C are computed. The K eigenvectors having the largest eigenvalues are selected. The dimensionality of the subspace K can be determined by using the following criterion.

$$\frac{\lambda_i}{\lambda_i} > threshold \ (\alpha)$$

The linear transformation RN>RK that performs the dimensionality reduction is by equation (6):

$$Z_n = U^T(x - x) = \bar{U}^T Ø_n \qquad (6)$$

6.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) searches for those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). More formally, given a number of independent features relative to which the data is described, LDA creates a linear combination of these which yields the largest mean differences between the desired classes. Mathematically speaking, for all the samples of all classes, the two measures are defined: 1) one is called within-class scatter matrix, as given byequation (7):

$$S_w = \sum_{i=1}^{c}\sum_{i=1}^{N_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^T \qquad (7)$$

where x^j_i is the ith sample of class *j*, μ_j is the mean of class j, *c* is the number of classes, and N_j the number of samples in class *j*; and 2) the other is called *between-class* scatter matrix, by equation (8):

$$S_b = \sum_{j=1}^{c}(\mu_j - \mu)(\mu_j - \mu)^T \qquad (8)$$

Where μ represents the mean of all classes.

The goal is to maximize the between-class measure while minimizing the within-class measure [12].

The standard LDA can be seriously degraded if there are only a limited number of observations N compared to the dimension of the feature space n. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn't change the location but only tries to provide more class separability and draw a decision region between the given classes [11]

7. PROPOSED SYSTEM

To solve spam detection and e-mail classification problem using Artificial Immune System. A new e-mail classification technique based on Clonal Selection Algorithm and feature extraction techniques shall be designed and implemented. At first of all the most important features will be extracted from each of two types spam and ham, then generate a spam and ham detector, after which e-mail classification will take place by utilizing the ham and the spam accordingly in other to successfully reduce the false rate. The experiment confirms the reliability and efficiency of our new techniques in

minimizing false positives and time consuming and maximizing true positive. The datasets used in this research is gotten from machine learning repository, Center for Machine Learning and Intelligent System.

7.1 Data Preprocessing

The data set is divided into the two parts; one is for training and the second is for testing. After the first division; our training data set will be further divided in to self detector (Ham) and nonself detector (Spam).

7.2 Feature Extraction by Information Gain or PCA or LDA

The information gain of each attribute are calculated and the attributes with low information gains are removed from the data set. The information gain of an attribute indicates the statistical relevance of this attribute regarding theclassificationPCA is a feature extraction method is unsupervised. This means that the class labels are not taken into account. Therefore, the presence of labels in the data set does not alter the resulting PCA projection.

Where LDA is a supervised feature extraction method that finds a linear subspace maximizing separability between classes. The dimensionality of the resulting subspace is fixed to the minimum between: number of features, number of samples, number of classes. Usually, the output dimensionality is determined by the number of classes.

7.3 Clonal

In the clonal selection method only a small set of best Artificial LymphoCytes ALCs (i.e., with the highest calculated affinity with a non-self pattern) is maintained so that the problem can be solved with the available minimal resources. The selected ALCs (i.e., detectors) are then cloned and mutated in an attempt to have a higher binding affinity with the presented nonself Ham pattern. The mutated clones compete with the existing set of ALCs, based on the calculated affinity between the mutated clones and the non-self pattern, for survival to be exposed to the next nonself Ham pattern.

The percent of data take off from dataset for training is divided in to spam training set and non-spam training set. The trained detectors is used to classify the rest of database email by obtaining feature vector after pre-processing when both e-mail and detectors affinity are calculated, and affinity that is greater than threshold, it is said to be Spam; otherwise it's a Ham.

8. EXPERIMENTAL SETUP AND RESULTS

During email classification, two mistakes occur by existing anti-spam method. It is either the email is recognized as self and is deleted or non-self and been accepted carelessly. This process is called false positive and false negative. The false positive occurs when the email or data that are needed to create a detector are classified as self while emails or data that are supposed to be discarded are recognized as non-self. Figure 1 depicts the functional block diagram of the proposed detection model.

Metrics are used as true negative rate, true positive rate, weighted accuracy, G-mean, precision, recall, and F-measure to evaluate the performance of learning algorithms assuming a total of N messages test set, the definition of variables: Spam, legitimate messages (Ham).

Then these Metrics can be defined to evaluate the mail classification system performance [13].

Detection Rate & False Alarm Rate, They are also called true positive rate (TPR) and true negative rate (TNR). To get optimal balanced classification ability, sensitivityand specificity are usually adopted to monitor classification performance on two classes separately.

$$TNR = \frac{TN}{TN+FP} \ , \qquad (9)$$

$$TPR = \frac{TP}{TP+FN} \qquad (10)$$

Precision, which is a spam probability of correct. The correct rate is higher; the misjudgment of legitimate messages as spam, the fewer the number of.

Precision (PPV) =TP/ (TP + FP) (11)

Accuracy, that is to judge all mail, and determine the probability of correct.

Accuracy (ACC) =

(TP + TN) / (TP + FN)+(FP +TN) (12)

Geometric Mean (G-mean): Is used to assess the performance based on the two metrics TPR and TNR, it is also the geometric means of classification accuracy on negative samples and classification accuracy on positive samples. Used if the target is to optimize classification performance with balanced positive class accuracy and negative class accuracy.

G-mean = $(TPR \times TNR)^{1/2}$ (13)

F-measure is used to integrate precision and recall into a single metric for convenience of modeling.

F-measure = (2 × Precision × Recall) /

(Precision + Recall) (14)

Where True positive value(TP), False positive value(FP), True negative value(TN), False negative value (FN).

In each experiment the number of Spam and Ham detectors generated are different for each cases of using Information Gain or PCA or LDA. And The overall time consuming for each experiment is computed (i.e. in cases of 10, 15, 21, 30, 40, 50, 57 features together).The following experiments were performed as follow:

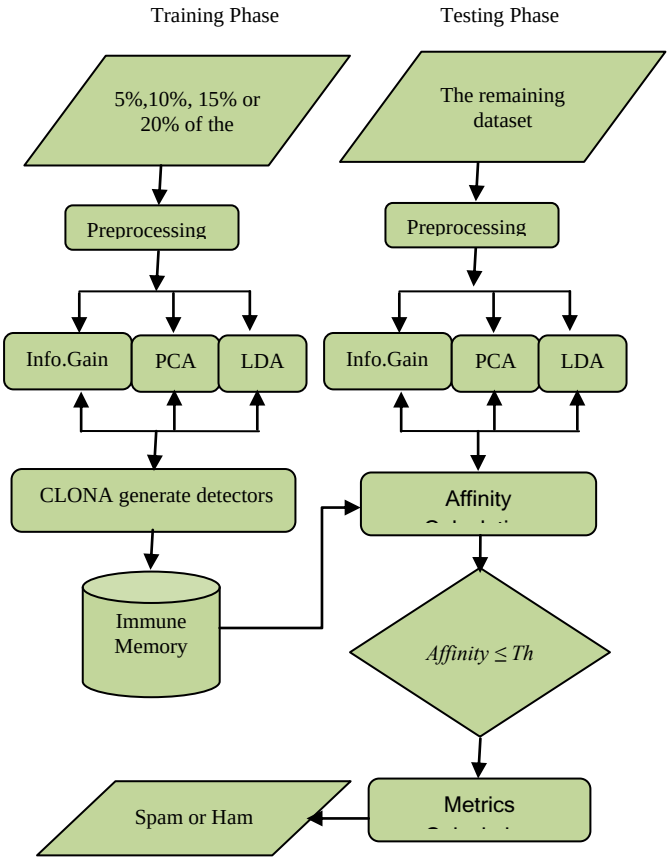

Fig. 1: Proposed Model for Spam/Ham detection.

Experiment 1: the input dataset is about 5% from the original which yield 139 Ham and 90 Spam, as show in table (1) and figure 2. Figure 3 shows the influence of number of features on time consuming, where the time increase Directly proportional as the number of features increase . The overall time consuming of this experiment are 12.23 secs, 12.55 secs, and 13.65 secs for Information Gain, LDA and PCA respectively.

Experiment 2: the input dataset is about 10% from the original which yield 278 Ham and 181 Spam, as shown in table (2) and figure 4, and figure 5 for other measures. The overall time consuming of this experiment are 58.22 secs, 1.02.15 mins, and 1.04.85 mins for Information Gain, LDA and PCA respectively.

Experiment 3: the input dataset is about 15% from the original which yield 418 Ham and 271 Spam, as shown in table (3). Figure 6 shows the differences between the accuracy of the algorithms on train and test dataset from table 3, it seems very convergent. Figure 7 shows the influence of the number of generated Spam and Ham detectors on the accuracy at test phase, by suggest that maximum Ham and Spam detectors are 5600 and 2500 respectively depending on the maximum number of detectors in table (3). The overall time consuming of this experiment are 2.36.48 mins, 2.30.30 mins, and 2.43.10 mins for Information Gain, LDA and PCA respectively.

Experiment 4: the input dataset is about 20% from the original which yield 557 Ham and 362 Spam. The overall time consuming of this experiment are 5.11.25 mins, 5.33.30 mins, and 5.59.90 mins for Information Gain, LDA and PCA respectively.

When a small (or nonrepresentative) training data set is used, there is no guarantee that Information Gain and LDA will outperform PCA.

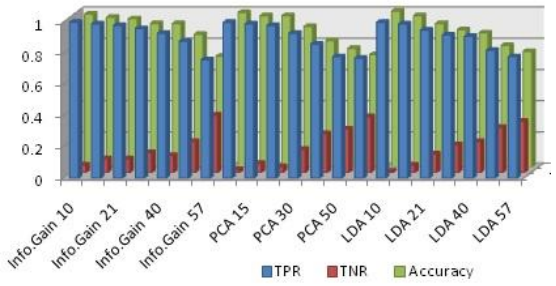

Fig. 2: Shows the experimental results of training phase of experiment 1 (5% of dataset).

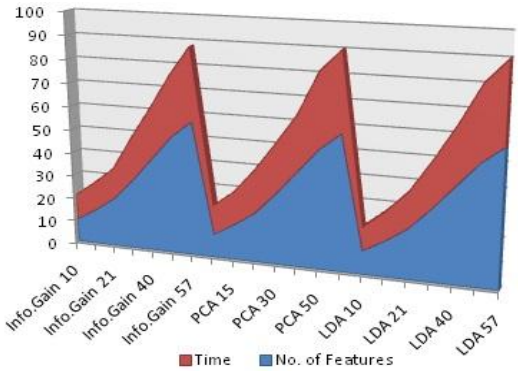

Fig. 3: Shows the time consuming of experimental results of training and testing phases of experiment 1 (5% of dataset).

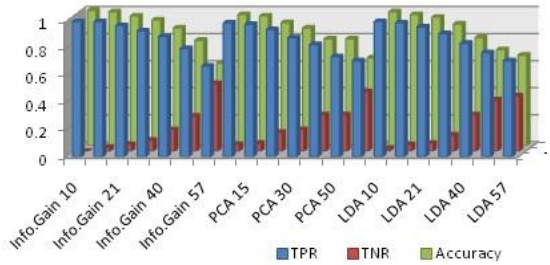

Fig. 4: Shows the experimental results of testing phase of experiment 2 (10% of dataset).

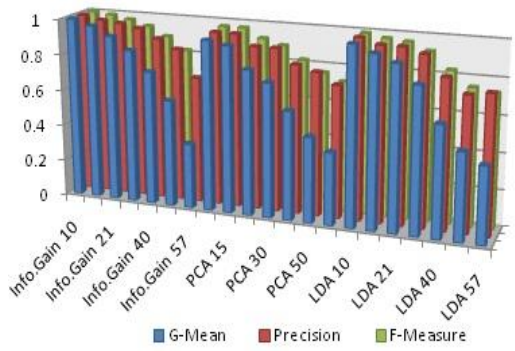

Fig. 5: Shows the experimental results of testing phase of experiment 2 (10% of dataset) for G-Mean, Precision and F-Measure.

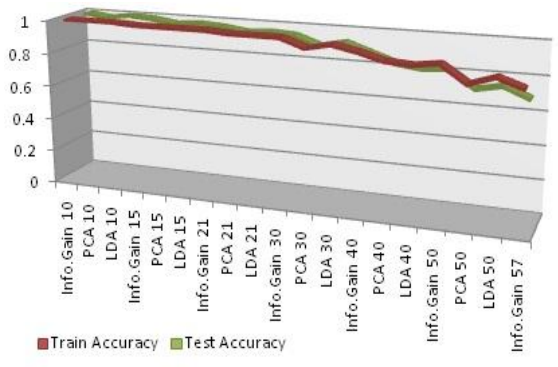

Fig. 6: The differences between the accuracy of the algorithms on train and test dataset from table 3 which applied training on 15% of dataset.

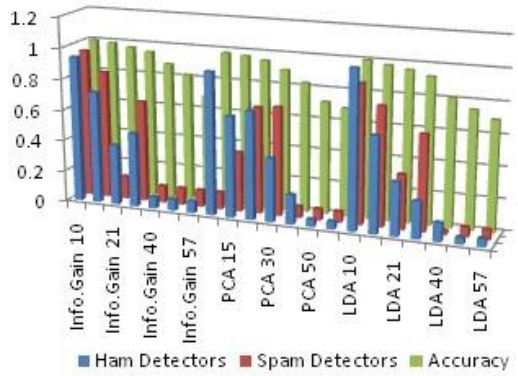

Fig. 7:The relation between the number of generated Spam and Ham detectors and the Accuracy at test phase of experiment 3 (15% of dataset).

9. CONCLUSIONS

An efficient email filtering approach which is consists of two phases are training and testing. This new model tries to increase the accuracy of a spam filtering and time consuming via combine a several well known feature extraction techniques with the artificial immune system by using its algorithm the Clonal selection algorithm. Experimental results showed an improvement in the performance of the new spam filtering than using each technique alone as it always seek to get the highest and fastest detectors to reduce the false positive rate and get highest accuracy. The experimental results applied on 4601 instances of email messages shows a high efficiency with the less number of false alarm 0 and High detection rate 1, especially when the experiment depend on low number of the most important attributes. These promising results of the immune-inspired method can be further developed and even integrated with other methods as an appealing future direction, and also as a model that could help us better understand the behavior of immune system and how it could be very useful in different fields of computer and network security.

Table 1: Results of training on 5% of data (139 Ham & 90 Spam) as antigens

Algorithm	No of used features	Ham Abs	Spam Abs	TPR		FNR		Accuracy		Time (Secs)
				Train	Test	Train	Test	Train	Test	
Information Gain	10	246	69	1	0.99	0.05	0.02	0.98	0.99	1.10
Information Gain	15	110	74	0.99	0.98	0.09	0.05	0.96	0.97	1.23
Information Gain	21	107	21	0.98	0.95	0.09	0.08	0.95	0.94	1.32
Information Gain	30	32	18	0.96	0.91	0.13	0.14	0.92	0.89	1.85
Information Gain	40	31	18	0.93	0.87	0.11	0.2	0.92	0.84	2.18
Information Gain	50	27	18	0.88	0.78	0.2	0.32	0.85	0.74	2.60
Information Gain	57	27	18	0.76	0.66	0.37	0.52	0.71	0.59	3.10
PCA	10	762	60	1	0.98	0.02	0.04	0.99	0.97	1.25
PCA	15	114	28	0.99	0.95	0.06	0.14	0.97	0.91	1.42
PCA	21	85	18	0.98	0.91	0.04	0.18	0.97	0.88	1.82
PCA	30	30	18	0.93	0.86	0.15	0.29	0.9	0.79	2.10
PCA	40	35	18	0.86	0.78	0.25	0.39	0.81	0.71	2.34
PCA	50	27	18	0.78	0.67	0.28	0.5	0.76	0.61	3.13
PCA	57	27	18	0.77	0.65	0.36	0.55	0.72	0.57	3.32
LDA	10	935	103	1	0.99	0.01	0.01	1	0.99	1
LDA	15	154	75	0.99	0.98	0.05	0.05	0.97	0.97	1.22
LDA	21	126	19	0.95	0.95	0.12	0.09	0.92	0.93	1.51
LDA	30	35	18	0.92	0.9	0.18	0.18	0.88	0.87	1.95
LDA	40	30	18	0.91	0.81	0.2	0.32	0.86	0.76	2.41
LDA	50	27	18	0.82	0.72	0.29	0.46	0.78	0.64	3.07
LDA	57	27	18	0.78	0.65	0.33	0.53	0.74	0.58	3.33

Table 2: Results of training on 10% of data (278 Ham & 181 Spam) as antigens

Algorithm	No of used features	Ham Abs	Spam Abs	TPR		FNR		Accuracy		G-M	Prec.	F-M
				Train	Test	Train	Test	Train	Test	Test	Test	Test
Information Gain	10	2550	239	1	1	0.01	0	1	1	1	1	1
Information Gain	15	1909	1110	1	1	0.02	0.03	0.99	0.99	0.97	0.98	0.99
Information Gain	21	746	169	0.99	0.97	0.01	0.05	0.99	0.96	0.92	0.97	0.97
Information Gain	30	934	144	0.98	0.93	0.04	0.08	0.97	0.93	0.850	0.95	0.94
Information Gain	40	185	44	0.95	0.89	0.1	0.16	0.93	0.87	0.74	0.9	0.89

Algorithm	No of used features	Ham Abs	Spam Abs	TPR		FNR		Accuracy		G-M	Prec.	F-M
				Train	Test	Train	Test	Train	Test	Test	Test	Test
Information Gain	50	184	36	0.87	0.8	0.13	0.26	0.87	0.78	0.59	0.85	0.82
Information Gain	57	183	36	0.78	0.67	0.32	0.5	0.74	0.61	0.36	0.7	0.68
PCA	10	1906	1083	1	0.99	0.02	0.05	0.99	0.97	0.94	0.96	0.97
PCA	15	2437	1106	1	0.98	0.04	0.06	0.98	0.96	0.92	0.96	0.97
PCA	21	1504	130	0.97	0.94	0.05	0.14	0.96	0.91	0.8	0.9	0.92
PCA	30	1533	92	0.94	0.88	0.12	0.16	0.92	0.87	0.74	0.9	0.89
PCA	40	196	36	0.87	0.83	0.22	0.27	0.83	0.79	0.6	0.82	0.82
PCA	50	189	36	0.84	0.74	0.23	0.27	0.81	0.79	0.47	0.79	0.76
PCA	57	192	36	0.78	0.71	0.35	0.44	0.73	0.65	0.40	0.73	0.72
LDA	10	2548	1035	1	1	0.03	0.02	0.98	0.99	0.98	0.99	0.99
LDA	15	1876	1075	0.99	0.99	0.02	0.05	0.99	0.97	0.94	0.96	0.97
LDA	21	1769	157	0.96	0.96	0.03	0.06	0.97	0.95	0.9	0.96	0.96
LDA	30	1475	74	0.94	0.91	0.08	0.12	0.93	0.9	0.8	0.93	0.92
LDA	40	203	39	0.9	0.84	0.16	0.27	0.88	0.8	0.61	0.82	0.83
LDA	50	187	37	0.85	0.77	0.27	0.38	0.8	0.71	0.48	0.74	0.75
LDA	57	173	36	0.78	0.71	0.3	0.41	0.75	0.67	0.42	0.76	0.73

Table 3: Results of training on 15% of data(418 Ham & 271 Spam) as antigens.

Algorithm	No of used features	Ham Abs	Spam Abs	TPR		FNR		Accuracy		G-M	Prec.	F-M
				Train	Test	Train	Test	Train	Test	Test	Test	Test
Information Gain	10	5271	2393	1	1	0.01	0	1	1	1	1	1
Information Gain	15	4052	2063	1	0.99	0.01	0.2	0.99	0.99	0.97	0.98	0.98
Information Gain	21	2178	363	1	0.99	0.01	0.05	0.99	0.97	0.94	0.97	0.98
Information Gain	30	2680	1632	0.99	0.96	0.03	0.07	0.98	0.95	0.89	0.95	0.95
Information Gain	40	418	271	0.95	0.89	0.9	0.13	0.93	0.88	0.77	0.92	0.9
Information Gain	50	418	271	0.92	0.85	0.13	0.23	0.9	0.82	0.65	0.85	0.85
Information Gain	57	418	271	0.84	0.73	0.27	0.38	0.8	0.69	0.45	0.77	0.75
PCA	10	5131	271	1	0.99	0.01	0.03	1	0.98	0.96	0.98	0.98
PCA	15	3633	959	1	0.98	0.3	0.04	0.99	0.97	0.94	0.97	0.97
PCA	21	3865	1710	0.99	0.97	0.05	0.08	0.98	0.95	0.89	0.95	0.96
PCA	30	2292	1740	0.96	0.93	0.1	0.15	0.93	0.9	0.79	0.9	0.91
PCA	40	1064	187	0.91	0.86	0.14	0.21	0.89	0.83	0.68	0.86	0.86
PCA	50	260	180	0.82	0.75	0.22	0.33	0.8	0.72	0.5	0.81	0.78
PCA	57	274	179	0.82	0.73	0.27	0.39	0.78	0.69	0.45	0.77	0.75
LDA	10	5647	2225	1	1	0	0	1	1	1	1	1
LDA	15	3432	1903	0.99	0.99	0.01	0.4	0.99	0.98	0.95	0.97	0.98
LDA	21	1912	871	0.98	0.98	0.03	0.07	0.98	0.96	0.91	0.95	0.96
LDA	30	1326	1527	0.97	0.96	0.06	0.11	0.96	0.93	0.85	0.93	0.94
LDA	40	662	65	0.91	0.84	0.17	0.24	0.88	0.81	0.64	0.84	0.84
LDA	50	239	157	0.87	0.79	0.19	0.31	0.85	0.75	0.56	0.8	0.97
LDA	57	284	159	0.84	0.74	0.24	0.37	0.81	0.7	0.47	0.78	0.76

10. REFERENCES

[1] Geerthik S., 2013. “Survey on Internet Spam: Classification and Analysis”, Pages 384-391 in Int.J.Computer Technology & Applications,Vol 4 (3), May-June.

[2] Sao P., Singh A., 2015.”Survey on Email Spam Classification using Different Classification Method”, Pages 680-684 in JETIR, Volume 2, Issue 3.

[3] Sharayu S A., Irabashetti P., 2014, "Efficient Spam Filtering Based on Artificial Immune System (AIS)", International Journal of Ignited Minds (IJIMIINDS), Volume: 01 Issue: 12 | Dec.

[4] Ozarkar P.& PatwardhanM., 2013. “Efficient Spam Classification by Appropriate Feature Selection”, Pages 48-57 in Global Journal of Computer Science and TechnologySoftware & Data EngineeringVolume 13 Issue 5 Version 1.0.

[5] Idris I. and Selamat A., 2015, " A Swarm Negative Selection Algorithm for Email Spam Detection", Journal of Computer Engineering &Information Technology, March 17.

[6] Kathiravan A. V. and Vasumathi B., 2015, " Artificial Immune System Based Classification Approach for Detecting Phishing Mails", Pages 4308 – 4315, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 5, May.

[7] Mahmoud T. M., El Nashar A. I., Abd-El-Hafeez T. and Khairy M., 2014, "An Efficient Three-phase Email Spam FilteringTechnique", Pages 1184-1201, British Journal of Mathematics & Computer Science.

[8] “UCI repository of Machine learning Databases”, Department of Information and Computer Science,University of California, Irvine, CA,http://www.ics.uci.edu/~mlearn/MLRepository.html, Hettich, S., Blake, C. L., and Merz, C. J.,1998.

[9] LanT., Erdogmus D., BlackL., and SantenJ., 2014. “A Comparison of Different Dimensionality Reduction andFeature Selection Methods for Single Trial ERP Detection”, Conf Proc IEEE Eng Med Biol Soc.

[10] Singh S., Silakari S. and Patel R., 2011, "An efficient feature reduction technique for intrusion detection system", International Conference on Machine Learning and Computing, IPCSIT vol.3.

[11] Khan A. and Farooq H., 2011, "Principal Component Analysis-Linear Discriminant Analysis Feature Extractor for Pattern Recognition", pages 267-270, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 2, November.

[12] Martinez A. M. and Kak A. C., 2001, " PCA versus LDA", Pages 228-233, IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 23, NO. 2, February.

[13] Al-Anezi M. M. and Al-Dabagh N. B., "Multilayer Artificial Immune Systems for Intrusion and Malware Detection", LAP 2012.