Research Article

Whale optimization algorithm-based email spam feature selection method using rotation forest algorithm for classification

Maryam Shuaib[1] · Shafi'i Muhammad Abdulhamid[1] · Olawale Surajudeen Adebayo[1] · Oluwafemi Osho[1] · Ismaila Idris[1] · John K. Alhassan[1] · Nadim Rana[2]

Abstract
Email has continued to be an integral part of our lives and as a means for successful communication on the internet. The problem of spam mails occupying a huge amount of space and bandwidth, and the weaknesses of spam filtering techniques which includes misclassification of genuine emails as spam (false positives) are a growing challenge to the internet world. This research work proposed the use of a metaheuristic optimization algorithm, the whale optimization algorithm (WOA), for the selection of salient features in the email corpus and rotation forest algorithm for classifying emails as spam and non-spam. The entire datasets were used, and the evaluation of the rotation forest algorithm was done before and after feature selection with WOA. The results obtained showed that the rotation forest algorithm after feature selection with WOA was able to classify the emails into spam and non-spam with a performance accuracy of 99.9% and a low FP rate of 0.0019. This shows that the proposed method had produced a remarkable improvement as compared with some previous methods.

Keywords Whale optimization algorithm · Metaheuristic algorithm · Email spam · Classification algorithms · Rotation forest · Feature selection

1 Introduction

Email has continued to take the lead in information dissemination globally. Its speed, cost-effectiveness and ease of use from personal computers, smartphones and other last-generation electronic gadgets have made emails popular [3, 14]. In spite of the growth in the usage of other means of online communication including instant messaging and social networking, emails have retained the lead in business communications and still serve as a prerequisite for other means of communications and e-transactions. The use of emails has led to visible improvement in group communications, the impact of which is seen in growing businesses around the world [9].

Emails are used as a means of easy communication regardless of the distance. In a recent study by Radicati [23], it is estimated that in 2017, 269 billion business and consumer emails will be sent and received daily and this figure is projected to keep growing at an intermediate annual rate of 4.4% spread across the next 4 years, to peak at 319.6 billion by the end of 2021. Approximately half of the world population is expected to own an email this year topping 3.7 billion, and by the end of 2021, the number of email users globally will be above 4.1 billion. The wide spread use of emails for communication and other dealings has led to an increase in the number of spam emails sent and received globally. Spam is a serious threat to the internet world and the email family. Although there is not

✉ Shafi'i Muhammad Abdulhamid, shafii.abdulhamid@futminna.edu.ng; Maryam Shuaib, maryambobi@gmail.com; Olawale Surajudeen Adebayo, waleadebayo@futminna.edu.ng; Oluwafemi Osho, femi.osho@futminna.edu.ng; Ismaila Idris, ismi.idris@futminna.edu.ng; John K. Alhassan, jkalhassan@futminna.edu.ng; Nadim Rana, nadimrana17@gmail.com | [1]Department of Cyber Security Science, Federal University of Technology, Minna, Nigeria. [2]College of Computer Science and Information Systems, Jazan University, Jazan, Kingdom of Saudi Arabia.

SN Applied Sciences (2019) 1:390 | https://doi.org/10.1007/s42452-019-0394-7

Received: 26 November 2018 / Accepted: 22 March 2019

Published online: 02 April 2019

a distinct description for spam emails (also referred to as junk emails) and how it is different from legitimate mail (also known as non-spam or ham), spam emails are unrequested messages that are not asked for and are almost indistinguishably sent to aggregate recipients through emails. The sender of spam mails has no existing interaction with the receivers but collects the addresses from different sources such as phonebooks and filled forms [1, 28]. During the period between 2014 and 2017, spam messages represented 59.56% of email traffic worldwide with above 14.5 billion spam messages sent daily around the globe [17, 31]. Email spam has remained a major problem that not only impacts internet users but also causes a persistent problem for companies and organizations. Previous techniques have been marred by the adaptive nature of unsolicited email spam [10], hence the reason for the continued improvement in existing techniques and development of new models.

Spam mails have continued to rise with the increase in the number of email users globally, and these unsolicited emails have a high cost in terms of time, storage space, network bandwidth consumptions and indirect costs to protect privacy and security breaches. Researches have been carried out to create new filtering techniques to eliminate or at least block spam emails from hindering activities of email users. In the same vain, spammers keep coming up with new techniques to evade detection. As such, the accuracy of the current filtering and detection techniques is gradually decreasing, and this is in addition to the increase in false positive detection rate. Thus, a constant update through the use of optimization algorithms is required over time. Features selection is among the most efficient means of improving system performance and enhancing data representation in terms of defined criteria [27]. It aims to determine the important features and remove the not so important ones from the actual dataset features. It also reduces data dimensionality and creates a good data understanding for various machine learning applications. The goal of the most optimization problems that are of theoretical or practical significance is to search for a "best" configuration of a set of variables to achieve some goals.

A metaheuristic is an iterative generation process that directs a subordinate heuristic by bringing together various concepts for exploring and exploiting the search space. They are an efficient means for achieving near-optimal solutions for large-scale problems [8]. Numerous metaheuristic algorithms have been proposed over the years, including genetic algorithms (GA), simulated annealing (SA), Tabu search (TS) and ant colony optimization (ACO). Each metaheuristic purports a different strategy, yet some basic principles remain the same. The aim of the optimization is to search for a solution that maximizes or minimizes a user-defined objective function. In the process of optimization, new solutions are usually brought forth through mutation. All metaheuristics generally take note of the current optimum and can be ended based on varying termination conditions (e.g. time or number of iterations). Metaheuristic optimization algorithms are getting more and more popular in different machine learning applications due to the fact that they: (i) rely on rather simple concepts and are easy to implement; (ii) do not require gradient information; (iii) can bypass local optima; (iv) can be utilized in a wide range of problems covering different disciplines. Nature-inspired metaheuristic algorithms solve optimization problems by mimicking biological or physical phenomena [27].

Whale optimization algorithm (WOA) is a metaheuristic optimization algorithm that is capable of avoiding local optima and getting a global optimal solution that makes it suitable for practical applications without structural alterations in algorithm for solving different constrained or unconstraint optimization problems. WOA unified with adaptive technique cuts down the computational times for extremely complex problems [35]. This research work seeks to present a metaheuristic optimization algorithm: WOA that mimics the hunting behaviour of humpback whales for features selection on the Spambase email dataset and the Enron-Spam corpus. Optimization results from different literature show that WOA is very competitive in comparison with other well-known optimization methods in improving system performance. Despite the effectiveness of the existing spam filtering methods used to curb the issues of spam, spammers periodically bypass these techniques through a change in their practice and behaviours. A continuous improvement to the existing techniques and stemming up new ones are important to control SPAM and ensure that genuine emails are not classified as spam by creating systems or enhancing existing systems to reduce the false positive (FP) rate.

The complexity of an algorithm in relation to time is proportional to how fast certain the WOA executes taking into account the whole operations, input parameters, the resources and time required to implement [4, 16]. The WOA has rather has low time complexity if we consider the amount of iterations and entire operations to be executed as compared to other previous propose algorithms in the literature. This makes the implementation of the WOA very easy as compared to similar schemes. Furthermore, the algorithm has few input parameter to be set, which decreases the ambiguity in selecting the best values that will produce faster convergence and exact results. In this research, the completion time has been used to determine the time complexity of the proposed application of optimized WOA in the area of email spam classification. Therefore, the WOA metaheuristic algorithm on integrated with

RF technique results in less computational time to reach optimum solution, local minima avoidance and faster convergence. It reduces the computational time for highly complex problems. The WOA has been used to enhance performances of classical engineering and machine learning problems [2, 6, 12, 15, 27, 35]. However, the WOA is yet to be used on email spam datasets. The contributions of this work are given as follows:

- We formulate the mathematical model of the proposed WOA–RF algorithm for email spam detection.
- We present a WOA-based email spam feature selection algorithm to enhance the extraction of salient features.
- We put forward a classification algorithm for email spam corpus using the RF Algorithm after features selection with WOA.
- We perform an experimental evaluation of the performance of the proposed WOA–RF algorithm using standard metrics.

In this research article, we propose a feature selection method called WOA with rotation forest (RF) for email spam classification. The rest of the paper is organized as follows: In Sect. 2, we review some related studies. The problem formulation is presented in Sect. 3. The design of the proposed methodology used in the study is discussed in Sect. 4. Section 4 presents the experimental set-up of the system, parameters of the evaluation and dataset used. Section 6 chronicles the results and discussion. The study is concluded with some future recommendations in Sect. 7.

2 Related works

In this section, we review some related researches that applied intelligent algorithms to address the problem of email spam detection and classification. We also surveyed related researches that utilize the WOA in addressing other optimization problem in other related fields.

2.1 Email spam detection and classification algorithms

Recent studies on spam emails show that several techniques are used to detect, filter or classify spam emails. According to Teli and Biradar [32], the solutions are designed to work either on the content of the emails (content-based) or on the links of the email (list-based). The list-based filters seek to block spam by grouping email senders as spammers or trusted users, and barricading or allowing their respective messages [32, 33], while the content-based filters are a widely used group of methods to filter spams. In content-based spam filtering, the focal point is in classifying the email as spam or as legitimate, based on the data that are seen in the body or the content of the mail. Therefore, the header section is ignored in the most cases of content-based spam filtering, and only a few mail headers like "subjects" are considered to classify the emails as spam or ham [22, 32].

A number of established anti-spam techniques for avoiding spam such as Bayesian-based sort, rule-based system, IP blacklist, heuristic-based filter, white list and DNS black holes were identified by Sharma and Kaur [29]. They used RBF, a neural network technique where they trained the neurons. Improved accuracy, precision, recall and Frr were achieved with the proposed technique. They compared their proposed technique with SVM and achieved a comparatively better result. They used a database of approximately 1000 spam words and suggest as future work to use larger dataset for spam detection. Rathi and Pareek [24] did a comparative study to analyse different data mining techniques in spam detection on some spam datasets in a bid to find out the best classifier for email classification. In their research, they examined the performance of different classifiers with and without feature selection algorithm. The best-first feature selection algorithm was used to select the desired features, and various classifiers were applied for classification. They ascertained that the improvement in results in accuracy with feature selection is employed in the experiment and also random tree was found to be the best classifier for spam mail classification with correct prediction rate of 99.72%. Though none of the techniques have achieved a hundred per cent accuracy in spam mail classification, random tree was close in achieving that.

In exploring the area of content-based spam filtering and detection algorithms, Malarvizhi and Saraswathi [18] centred their research on spam as a major problem faced daily by the internet community. The contents of every mail were presented as a set of descriptors or terms. Typically, the terms include words that appear in the document. User profiles are also presented with the same terms and built up by examining the content of mails seen by the user. Their research paper mainly adds to the comprehensive study of spam detection algorithms under the group of content-based filtering. Then, the implemented results were benchmarked to analyse how accurately they have been classified to their original groups of spam. Bayesian method was arrived at as the most efficient technique among the discussed techniques for spam filtering.

Yasin and Abuhasan [38] proposed an intelligent classification model for detecting phishing emails with the use of knowledge discovery, data mining and text processing methods. The paper presents the idea of phishing terms weighting that assesses the weight of phishing terms in each email. The pre-processing phase

is increased by applying text stemming and WordNet ontology to improve the model with synonyms of words. The model made use of knowledge discovery procedures with five popular classification algorithms (J48, naïve Bayes, support vector machine (SVM), multilayer perceptron and random forest). Two widely used datasets were used for training and testing of the proposed model with tenfold cross-validation test option to overcome the over-fitting problem. They acquired a notable improvement in classification accuracy with 99.1% accuracy using the random forest algorithm and 98.4% with J48. They compared their work with other research works in phishing emails.

Parveen and Halse [21] made use of three data mining methods to the spam dataset in a bid to identify the most suitable classifier for email classification as spam and ham in their work 'Spam Mail Detection using Classification'. They analysed the performance of the classifiers and discovered that the naïve Bayes classifier provides better accuracy of 76% in comparison with the other two classifiers such as support vector machine and J48 and also the time taken for naïve Bayes classifier is less than time taken for J48 and SVM which implies that naïve Bayes classifier is the most suitable classifier among the three that were used for classifying the spam mails. They used a single dataset for testing the classifiers and suggested as future work the use of more high-quality filters to improve accuracy and also try the use of other classifiers. A combination of particle swarm optimization and artificial neural network was applied for feature selection with classification carried out using support vector machine to classify and separate spam by Zavvar et al. [39]. They compared their technique with other techniques such as data classification self-organizing map and *K*-means based on the criteria area under curve. The results showed that the area under curve (AUC used as threshold for performance evaluation) in their proposed method is better than other methods. Research gap is the use of a single dataset.

In a search for an effective spam filtering technique, Kumar [13] applied ten machine learning algorithms including random forest, averaged one-dependence estimators, Fisher's linear discriminate function, logistic model trees, LOGISTIC, radial basis function classifier, rotation forest with J48 base classifier, rotation forest with LMT as base classifier, simple logistic and sequential minimal optimization with the test option of tenfold cross-validation to classify the Spambase dataset from UCI repository. The result showed that the random forest classifier which is a combination of tree predictors such that all trees depend on the values of a random vector sampled autonomously and with the similar distribution for all trees in the forest had the best with AUC, accuracy and MCC value up to 0.987, 0.955 and 0.906, respectively.

Tuteja and Nagaraju [36] applied BPNN filtering algorithm, i.e. artificial neural network feed forward with back propagation, based on text classification to classify significant emails from unsolicited ones. They applied k-means clustering in the pre-processing stage and obtained an improved accuracy, better training time and a reasonable precision. Kaur and Kaur [11] in their paper titled 'Spam Detection Using Data Mining Tool in Matlab' focused on the reduction in error rate of data being misclassified. Unlike the previous researches reviewed where there were cases of misclassification, they modified the classification techniques to achieve better results and reduced error rates were found. In their research, they focused on further filtration of email data, calculated the linear re-substitution error, quadratic re-substitution error and cross-validation error and compared them. They made use of naïve Bayes algorithm and decision tree algorithm and successfully identified the misclassified mails and compared them all.

Negative selection algorithm and particle swarm optimization techniques were employed by Idris and Selamat [9] to improve email spam detection. They implemented a model to better the random generation of a detector in negative selection algorithm (NSA) with the use of stochastic distribution to model the data point using particle swarm optimization (PSO). They introduced a local outlier factor as the fittingness function to ascertain the local best (Pbest) of the candidate detector to arrive at the best solution. Distance measure was used to improve the disparateness between the non-spam and spam candidate detector. The detection process was halted when the expected spam coverage was arrived at. The experimental result showed that the rate of detection of the proposed improvement was higher with 91.22% than the standard negative selection algorithm with 68.86%. Hybridization of negative selection algorithm with differential equation (DE) was employed by Idris et al. [10] in building a spam detection model. A local selection differential evolution (DE) was used to yield detectors at the random detector generation phase of NSA. Local outlier factor (LOF) was employed as fitness function to increase the distance of generated spam detectors from the non-spam space. The issue of overlapping detectors was also solved by getting the minimum and maximum distance of two overlapped detectors in the spam space.

2.2 Differences between spam emails and phishing emails

While phishing emails are out of the scope of our work, it is important to look at the similarities and differences between spam emails and phishing emails. Dangerous and inauthentic emails are often classified as either spam emails or phishing emails. Many of the characteristics of

spam and phishing emails are interwoven. Characteristics such as using company logos and links, creating a plausible premise and requiring a quick response are all examples of how hackers spoof reputable companies and trick email users into engaging with their emails [26]. A spam email is basically unsolicited emails that are sent out, mostly with the purpose of selling a service or product. Spammers usually send such emails to a long list of recipients, in the hope that at least a few of them will respond back [30], while phishing emails are sent out specifically to extract useful personal and financial information from a user or its' system with the use of emails.

In more general terms, not all spam emails are phishing emails, while phishing emails are a kind of spam emails. This implies that a phishing email is only a subset of a spam email. The impact of spam emails is targeted on random individuals, while the impact of phishing email attacks is clearly on the security of individual computer users as well as whole organizations [20]. Spam emails do not request for sensitive information but for commercial purpose [30], while phishing emails always require a reply to an email or by phone, or a link to a Web site which most times redirect to Web pages that are identical to the legitimate ones. Spam emails are not usually malicious but can be vicious, while phishing emails are always malicious. The motive behind sending spam emails is simply to lure recipients into buying dubious products or participate in fraudulent and quasi-legal schemes [26], while the motives for phishing emails are specifically to get out personal information about individuals or company [5].

Due to the differences in both characteristics and features, spam emails are very much different from phishing emails. Therefore, the approach in research and methodology of handling the two issues are quite different in both attacks. We wish to state that the scope of this research will strictly only cover spam email attacks and wish not to discuss about phishing email attacks beyond what have been enumerated above to avoid overlap of discussions.

2.3 Recent applications of WOA

Whale optimization algorithm (WOA) recently developed by Mirjalili and Lewis [19] is a swarm-based optimization algorithm inspired by imitating the hunting behaviour of humpback whales. It is an improvement in the field of metaheuristic algorithms and is prompted by the bubble-net hunting strategy which is a unique behaviour that can only be seen in humpback whales. They prefer to hunt krill or small fishes close to the surface. WOA mimics the hunting behaviour with random or the best search agent to pursue the prey and the use of a spiral to simulate bubble-net attack mechanisms of humpback whales [12, 35]. The numerical efficiency of the WOA algorithm designed and implemented was tested by solving 29 mathematical optimization problems, 23 are classical benchmark functions and 6 composite benchmark functions. Six structural engineering problems including design of a tension/compression spring, design of a welded beam, design of a pressure vessel, design of a 15-bar truss, design of a 25-bar truss and design of a 52-bar truss were solved for more information gathering. WOA has also been compared with popular swarm-based optimization algorithms. WOA is very competitive with metaheuristic optimizers and superior over conventional techniques [19].

Aljarah et al. [2] in their work proposed a novel training algorithm using whale optimization algorithm (WOA). The approach was based on WOA for training the multilayer perceptron (MLP) network and is called as WOA–MLP taking into consideration two key aspects which are the representation of the search agents in the WOA and the selection of the fitness function. The research was the first based on its literature to test the proposed WOA-based trainer on a set of 20 datasets with different levels of difficulty based on classification accuracy and MSE evaluation measures. The results obtained showed that WOA trainer outperforms all other trainer optimizers and BP for blood, breast cancer, diabetes, hepatitis, vertebral, liver, diagnosis I, diagnosis II, Australian, monk, tic-tac-toe, ring, wine and seeds datasets with an average accuracy of 0.7867, 0.9731, 0.7584, 0.8717, 0.8802, 1.000, 1.000, 0.6958, 0.8535, 0.8224, 0.6733, 0.7729, 0.8986 and 0.8894, respectively. Hu et al. [7] in their research introduced a new control parameter, an inertia weight $\omega \in [0, 1]$, which is added to the WOA to achieve an improved WOA (IWOA) with the intention of tuning the influence on the current best solution. The IWOA was tried with 31 high-dimensional continuous benchmark functions and compared with the WOA, artificial bee colony (ABC) algorithm and the fruit fly algorithm (FOA). The numerical results showed that the IWOA is a strong search algorithm. IWOA does not only improve the basic WOA but also outperform both the artificial bee colony algorithm (ABC) and the fruit fly optimization algorithm (FOA). The WOA as was used by Touma [34] delivered optimum or close optimum solutions for solving the economic dispatch problem. The algorithm was tested on system of IEEE 30-bus with six generating thermal units. The results indicated the proposed technique using WOA produced optimal or near-optimal solutions. The obtained results explain and verify the closeness between the (WOA) method and particle swarm optimization, ant colony optimization and genetic algorithm as it is proved in the case studied. Lacosere [15] implemented a novel power system planning strategy through a combination of whale optimization algorithm (WOA) with pattern search algorithm (PS).

The WOA–PS was implemented to solve several optimal power flow (OPF) problems. IEEE 30-bus test system and various cases were tried in order to reduce the total fuel cost, reduce total power losses, the total emission and voltage profile. The acquired results from the WOA–PS approach were compared with those reported in related literatures. Based on the results arrived at, the effectiveness, robustness and performance of WAO-PS attained the best of all objective functions.

A new randomization approach referred to as adaptive technique was hybridized with WOA and tested on ten unconstraint test benchmark functions by Trivedi et al. [35]. The WOA algorithm shows improved feature that it makes use of logarithmic spiral function to cover a wide area in exploration phase, that is in addition to powerful randomization adaptive technique for whale optimization algorithm (AWOA) to achieve global optimal solution and faster convergence with less parameter dependency. Horng et al. [6] put forward a new multi-objective whale optimization algorithm (MOWOA) for optimization of the vehicle fuel consumption problem. The environment of current traffic condition, the path distances and the locations of vehicle were modelled to the objective functions, and search agents are mapped to a parsing solution in every iteration of a vehicle travelling during optimization. In the proposed technique, a grid network of vehicle transportation was used to represent solution constraint, and MOWOA. It handles two objectives simultaneously: the shortest distance and path-travel gasoline of the vehicle travelling. The optimal solution of these two objectives was got from the Pareto-optimal solution by calculating the probability of the obtained non-domination solution Pareto-frontier. The path of the globally best whale is selected in each iteration and reached by the vehicles in sequence. Simulations results showed that the MOWOA effectively gives the vehicle travelling optimization with a convincing performance. Compared with the obtained results of Dijkstra algorithm, and the MOGA methods, the quality of the proposed method MOWOA is slightly improved performance, and a reduction in the error rate.

The whale optimization algorithm was also proposed to ascertain the optimal DG size in the paper by Reddy et al. [25]. Distributed generator (DG) resources are small-scale electric power generating plants that can supply power to homes, businesses or industrial facilities in distribution systems. Power loss reductions, voltage profile improvement and increased reliability are some strengths of DG units which can be obtained through optimal placement of DGs. Reduction in system power losses and improvement in voltage profile were the main objectives of the research. The proposed method was tested on typical IEEE 15, 33, 69 and 85-bus radial distribution systems with different types of DGs and compared with other algorithms. Better results were obtained with WOA when compared with other algorithms which proves further that the WOA is efficient and robust.

In machine learning, whale optimization algorithm (WOA) was proposed and applied for the selection of best feature subsets for classification purposes of the Wisconsin Breast Cancer Database with 32 attributes and 596 instances [27]. The proposed model involved the use of WOA bio-inspired algorithm to select salient features, and the selected features were fed into the SVM using different kernel functions including RBF, linear, polynomial and quadratic. The results obtained were measured using different metrics including precision, accuracy, recall and f-measure. The best result was achieved with quadratic kernel function and showed that WOA is competitive for breast cancer diagnosis and obtains high degree of accuracy over some existing feature selection algorithms including genetic algorithm (GA), principle component analysis (PCA), mutual information (MI), statistical dependency (SD), random subset feature selection (RSFS), sequential floating forward selection (SFFS) and sequential forward selection (SFS). WOA achieved an overall 98.77% accuracy, 99.15% precision, 98.64% recall and 98.9% f-score.

The literature review shows clearly recent techniques used by researchers in the detection, filtering and classification of emails as legitimate or spam. Most of the techniques implemented in the literature are yet to achieve the peak performance or level of accuracy. The WOA has shown its feasibility in the area of optimization through the various works in which it was used where the WOA proved to be better or highly competitive with existing algorithms, and hence, it is selected to improve the accuracy of the existing spam detection techniques as it has not been used for the filtering and classification of email corpus.

3 Problem formulation

In email spam classification and filtering, a spam filter is sought, mathematically: a decision function f indicates if a given email message M is spam (S) or legitimate mail (H). The set of email messages can be represented by M.

A function $f: M \rightarrow \{S,H\}$ is searched.

This function is obtained by training one of the machine learning algorithms on a set of pre-classified messages $\{(m_1, c_1), (m_2, c_2),\ldots, (m_n, c_n)\}$, $m_i \in M$, $c_i \in \{S, H\}$.

This research involved two critical aspects: feature selection (using the WOA as the feature selector to extract features $v_j = f_1, f_2, f_3,\ldots, f_n$ from an email message and construct feature vectors $V = < v_1, v_2, v_3,\ldots, v_i >$ that

was then fed and used in the classification phase) and classification which was done with the rotation forest classifier.

4 Design of the proposed method

The proposed model involved the application of a bio-inspired metaheuristic optimization algorithm, WOA, for feature selection and the RF algorithm for classification of the email spam dataset. In this section, the algorithms and steps followed for the research work are presented.

4.1 Feature selection with WOA

The WOA depicts the special hunting behaviour of humpback whales as shown in Fig. 1. The whales follow distinctive bubbles to cause the formation of circular or '9-shaped path' while encircling prey in the hunting process. Simply bubble-net feeding/hunting behaviour is such that humpback whales go down in the water approximately 10–15 m and then start to produce bubbles in a spiral shape to encircle prey and then follow the bubbles and move upward to the surface. This algorithm was applied in getting features that are more useful and of significant value to the classification process.

4.1.1 Mathematical model of WOA

Mirjalili and Lewis's [19] mathematical model for WOA is given as follows:

```
Initialize the whales population Xi (i=1,2,...n)
Calculate the fitness of each search agent
X*=the best search agent
While (t<maximum number of iterations)
   For each search agent
   Update a, A, C, l and p
      If1 (p<0.5)
         If2 (|A| < 1)
            Update the position of the current search agent by Eq.
         Else if2 (|A| ≥ 1)
            Select a random search agent (Xrand)
            Update the position of the current search by Eq.
         End if2
      Else if1 (p ≥ 0.5)
         Update the position of the current search by Eq
   End if1
End for
   Check if any search agent goes beyond the search space and amend it
   Calculate the fitness of each search agent
   Update X* if there is a better solution
   t=t+1
end while
return X*
```

Fig. 1 Pseudocode of the WOA

a. Encircling prey equation

Humpback whale encircles the prey (small fishes) and then updates its position towards the optimum solution over the course of increasing number of iteration from start to a maximum number of iteration.

$$\vec{D} = \left|C\vec{X} * (t) - X(t)\right| \tag{1}$$

$$\vec{X}(t+1) = \vec{X} * (t) - A\vec{D} \tag{2}$$

where $\vec{A}, \vec{D}$ are coefficient vectors, t is a current iteration, $\vec{X} * (t)$ is position vector of the optimum solution so far and $X(t)$ is position vector.

Coefficient vectors $\vec{A}, \vec{D}$ are calculated as follows:

$$\vec{A} = 2\vec{a} * r - \vec{a} \tag{3}$$

$$\vec{C} = 2 * r \tag{4}$$

where $\vec{a}$ is a variable linearly decreasing from 2 to 0 over the course of iteration and r is a random number [0,1].

b. Bubble-net attacking method

In order to come up with the mathematical equation for bubble-net behaviour of humpback whales, two methods are modelled as:

(i) Shrinking encircling mechanism

This technique is employed by decreasing linearly the value of $\vec{a}$ from 2 to 0. Random value for a vector $\vec{A}$ is in range between [− 1, 1].

(ii) Spiral updating position

Mathematical spiral equation for position update between humpback whale and prey that was helix-shaped movement is given as follows:

$$\vec{X}(t+1) = \vec{D} * e^{bl} * \cos(2\pi l) + \vec{X} * (t) \tag{5}$$

where l is a random number [− 1, 1], b is constant defining the logarithmic shape, $\vec{D} = \left|\vec{X} * (t) - X(t)\right|$ expresses the distance between ith whale to the prey means the best solution so far.

Note: Assume that there is 50–50% probability that whale either follows the shrinking encircling or logarithmic path during optimization. Mathematically, it is modelled as follows:

$$\vec{X}(t+1) = \begin{cases} \vec{X} * (t) - \vec{A}\vec{D}, & \text{if } p < 0.5 \\ \vec{D} \cdot e^{bl} \cdot \cos(2\pi l) + \vec{X} * (t), & \text{if } p \geq 0.5 \end{cases} \tag{6}$$

where p expresses random number between [0, 1].

(iii) Search for prey

The vector $\vec{A}$ can be used for exploration to search for prey; vector $\vec{A}$ also takes the values greater than one or less than −1. Exploration follows two conditions.

$$\vec{D} = \left|\vec{C}\vec{X}_{rand} - \vec{X}\right| \quad (7)$$

$$\vec{X}(t+1) = \vec{X}_{rand} - \vec{A}\vec{D} \quad (8)$$

Finally, it follows these conditions:

$|\vec{A}| > 1$enforces exploration to WOA to find out global optimum avoiding local optima.
$|\vec{A}| < 1$ for updating the position of current search agent.

4.2 Rotation forest algorithm

RF is a technique for extracting classifier ensembles with the use of feature extraction. In creating the training data for a base classifier, the feature set is randomly split into *K* subsets (*K* is a parameter of the algorithm) and principal component analysis (PCA) is applied to every subset. All principal components are maintained in order to preserve the variability information in the data. Therefore, *K* axis rotations take place to create the new features for a base classifier. The aim of the rotation approach is to encourage simultaneously individual accuracy and diversity within the ensemble. Diversity is achieved through the feature extraction for each base classifier.

Decision trees were chosen here because they are sensitive to rotation of the feature axes, hence the name "forest". Accuracy is sought by keeping all principal components and also using the whole dataset to train each base classifier as shown in Fig. 2.

4.3 The proposed WOA–RF model

The proposed spam detection model consists of two main stages, features selection and classification as shown in Fig. 3. The model first starts by taking the Spambase dataset as input and then, WOA bio-inspired algorithm was adopted to select salient features; then, the selected features were used to feed the rotation forest, and the results obtained were evaluated using five different performance metrics including precision, accuracy, recall, *F*-measure and root mean squared error (RMSE).The same procedure was repeated for the Enron-Spam dataset.

4.3.1 Features selection phase

WOA was used as feature selection algorithm. The entire Spambase dataset was fed into MatLab R2012a, and 55 out of the 58 features were selected as the most relevant features retained with the use of the best position. Same procedure was applied to the Enron-Spam corpus, and 426 features out of the 1054 features were selected as the most relevant.

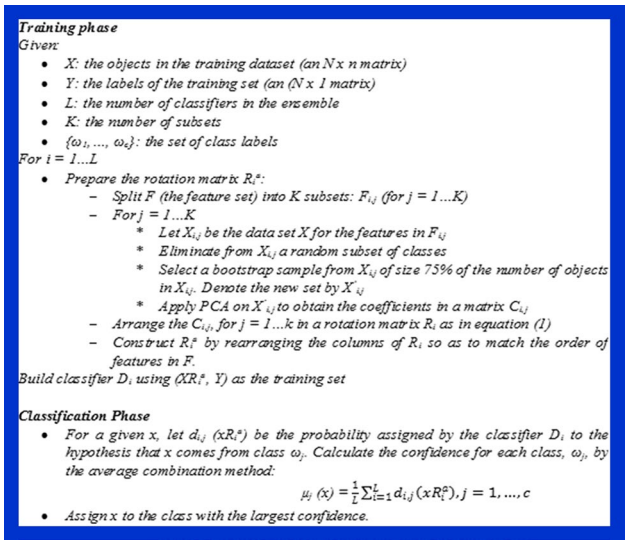

Fig. 2 Pseudocode of the rotation forest ensemble method

4.3.2 Classification phase

Rotation forest was then adopted as the classifier, and Waikato Environment for Knowledge Analysis (WEKA) tool was used as the interface. After loading the dataset, the attributes that were least relevant to the classification process were removed on WEKA before running the rotation forest algorithm for both spam corpus used.

5 Experimental set-up

The Matlab R2012a was used for feature selection. The WOA_toolbox was added into the Matlab library and ran with relevant parameters. The Waikato Environment for Knowledge Analysis (WEKA) toolkit was used for its ease of use when compared to other machine learning applications for the classification phase. The Rotation Forest classifier was run using tenfold cross-validation on the complete datasets before feature selection, and after feature selection, 10-fold, 20-fold and 66% split were used. The confusion matrix generated was then used for evaluation of the performance of the proposed techniques.

5.1 Performance metrics

The performance of the proposed system was evaluated using the performance metrics: accuracy, FP rate,

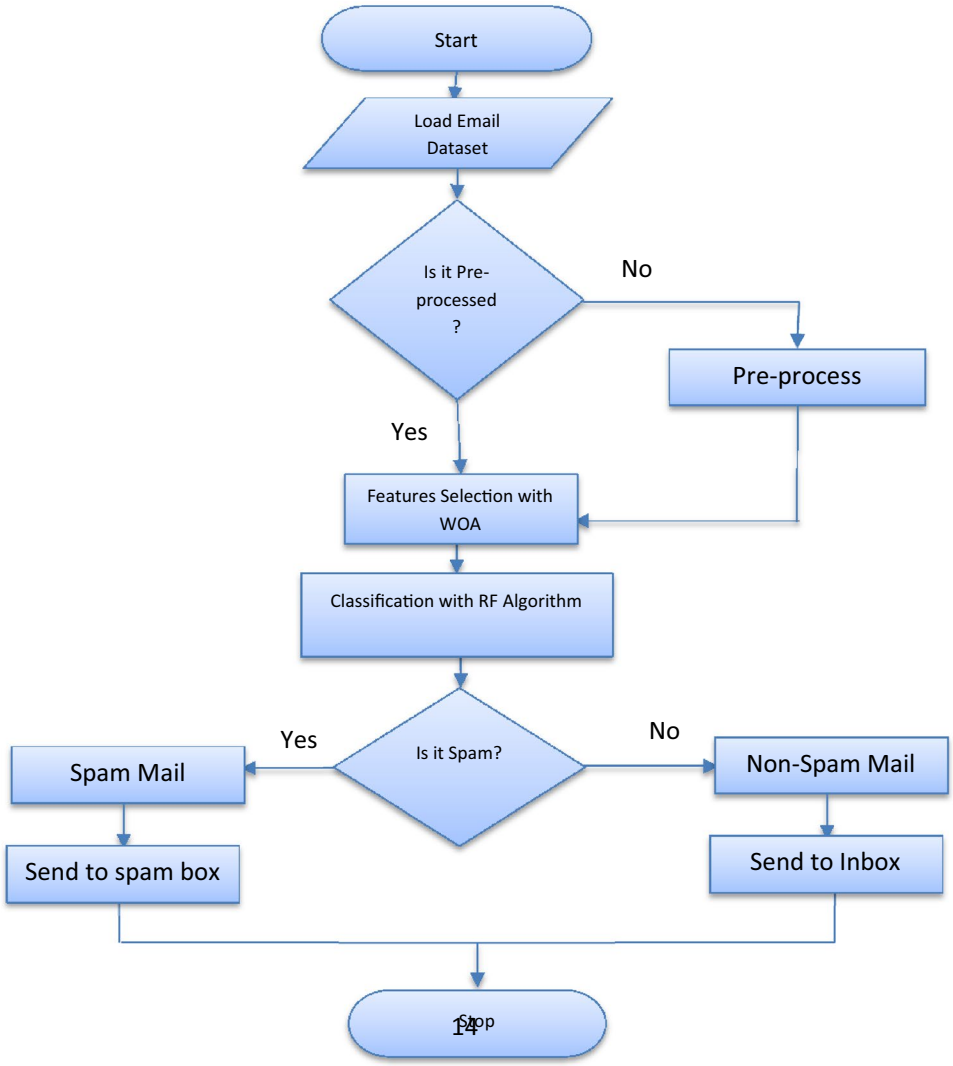

Fig. 3 Flow chart of the proposed system

precision, recall, *F*-measure, kappa statistics and root mean squared error (RMSE).

a. Accuracy

Accuracy of an algorithm is calculated as the percentage of the dataset correctly classified by the algorithm. It looks at positives or negatives dependently, and therefore, other measures for performance evaluation apart from the accuracy were used.

$$A = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \quad (9)$$

where TP = true positive, FP = false positive, TN = true negative and FN = false negative.

Positive and negative represent the classifier's prediction; true and false signify the classifier's expectation.

b. Precision

$$\text{Precision} = \frac{TP}{TP + FP} \quad (10)$$

It indicates the number of instances which are positively classified and are relevant. A high precision shows high relevance in detecting positives.

c. Recall

$$\text{Recall} = \frac{TP}{TP + FN} \quad (11)$$

It indicates how well a system can detect positives.

d. *F*-Measure

$$\text{F - Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

This is an approximate average of precision and recall when close, and generally referred to as the harmonic mean.

e. Root Mean Squared Error (RMSE)

After learning a model, the root mean square error shows how well the model has been learned.

$$\text{RMSE} = \frac{\sum (t_i - y_i)^2}{n}. \quad (13)$$

5.2 Testing and performance evaluation

Performance evaluation for email spam detection algorithms refers to the efficient and standard way of checking the functionality of the classification algorithm in terms of the performance of the algorithms and its ability to properly classify email as spam or non-spam on the basis of the performance metrics stated above. The proposed model was implemented and tested using Windows 8 with the following specification:

i. Processor: Intel Pentium (R) Core ™ i3-2350 M CPU @ 2.30 GHz 2.30 GHz
ii. Installed memory (RAM): 6.00 GB
iii. System type: 64-bit operating system.

5.3 Dataset description

Two datasets have been used in this work to carry out the experiments. They are Spambase dataset and the Enron-Spam corpus.

a. Spambase dataset

Spambase corpus obtained from the UCI machine learning repository was used. The dataset has 57 attributes with different variable types in 4601 instances, out of which 1813 (61%) are spam and 2788 (39%) are non-spam. The details of the dataset's attributes are presented in Table 1.

b. Enron-Spam corpus

The Enron: -C 53 -split-number 1135 a portion of the Enron Corpus with 1072 instances and 1054 attributes was also used for comparative reasons.

6 Results and discussion

The entire Spambase and Enron dataset consisting of both spam and non-spam emails were used for evaluating the RF classification algorithm before feature selection with WOA and after feature selection using the performance metrics discussed in Sect. 5.1. The experiment was carried out using MatLab and WEKA tool, and 10-fold cross-validation, 20-fold cross-validation and 66% split test options were used. The comparison of performance is discussed here.

6.1 Results analysis

In order to adequately classify the email spam corpus, ten-fold cross-validation was used because the larger the sample used for training, the better the performance of the classifier, but the returns start to decrease once a particular amount of training data are surpassed. Also the larger the testing sample, the higher the accuracy for the estimation

Table 1 Attribute description of the Spambase dataset

S/No	Attribute	Attribute type	No. of attributes	Description
1	word_freq_WORD	Continuous real [0, 100]	48	Gives the percentage of words present that corresponds to WORD, i.e. 100 * (number of times the WORD appears in the email)/total number of words in email
2	char_freq_CHAR	Continuous real [0, 100]	6	Gives the percentage of words present that corresponds to CHAR, i.e. 100 * (number of CHAR occurrences)/total characters in email
3	capital_run_length_average	Continuous real [1, …]	1	Shows the average length of continuous sequences of capital letters
4	capital_run_length_longest	Continuous integer [1, …]	1	Length of longest continuous sequence of capital letters
5	capital_run_length_total	Continuous integer [1, …]	1	Sum of length of continuous sequences of capital letters
6	spam	Nominal {0, 1}	1	Represents whether the email was seen as spam (1) or not (0), i.e. unrequested commercial email
	Total number of attributes		58	

of error, and because a number of tests on different datasets, with varied learning procedures, have shown that 10 is about the correct number of folds to get the best estimate of error [37], tenfold cross-validation test option was used. In using tenfold cross-validation, the dataset is partitioned randomly into 10 parts in which the class is represented in approximately the same ratios as in the full dataset. Each partition is then held out in turn, and the learning scheme trained on the remaining nine-tenths; then, its error rate is processed on the holdout set. Hence, the learning procedure is carried out a total of 10 times on various training sets (each of which have a lot in common). Finally, the average of the 10 error estimates is taken to give an overall error estimate.

In order to make the results of the classification more stable, the algorithms were run 10 times with different random number seeds (1, 2, 3,…, 10) and the classifiers' predictions were combined by averaging. For the purpose of comparisons, the datasets were also run using 20-fold cross-validation and percentage split which allows you to take out a certain percentage of the data for testing, 66% split were also employed for this research work.

6.1.1 Performance of the WOA–RF technique

The performance evaluation of the RF Algorithm after feature selection on both the Spambase dataset and Enron-Spam corpus with WOA using different test options is illustrated in Figs. 4 and 5. The minimum value, maximum value and the standard deviation for each of the performance metrics are also obtained for each test option and are presented in Tables 2, 3, 4, 5, 6 and 7.

The accuracy is used to show the level of correct predictions. The value 1 is largest, indicating the highest accuracy, when the rotation forest classification algorithm was run on the Spambase dataset after feature selection with

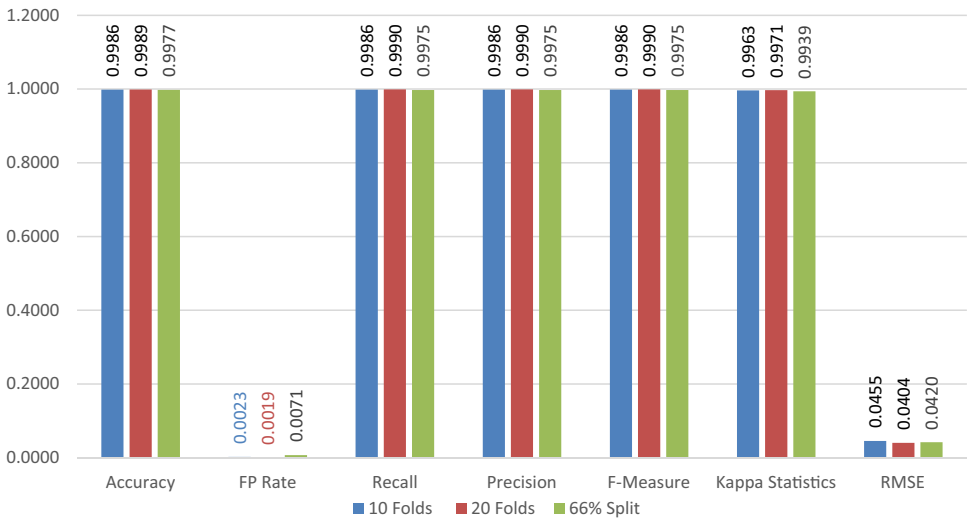

Fig. 4 Performance evaluation of the rotation forest algorithm after feature selection (AFS) with WOA using different test options using Spambase dataset

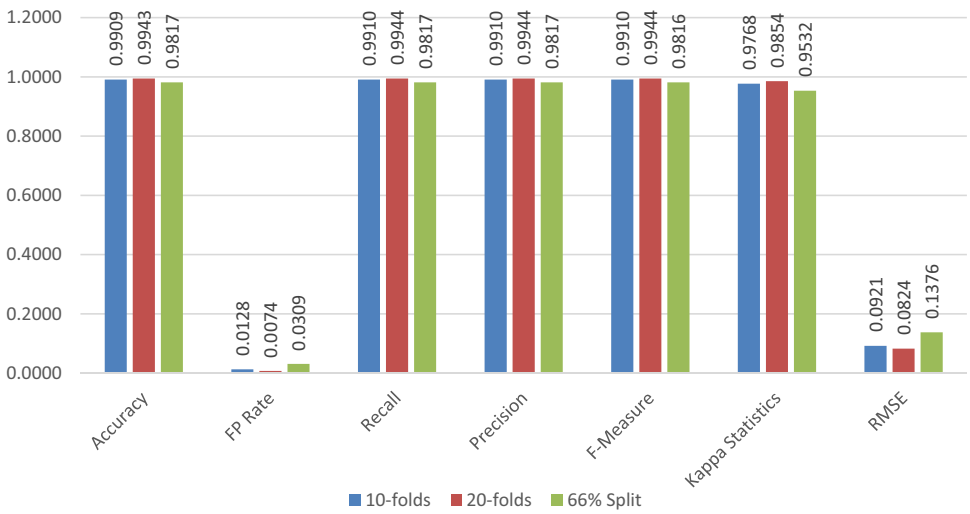

Fig. 5 Performance evaluation of the rotation forest algorithm after feature selection (AFS) with WOA using different test options using Enron-Spam corpus

Table 2 Summary of results after feature selection with WOA on Spambase dataset using 10-fold cross-validation

	Average	Min	Max	STD
Accuracy	99.86	99.7333	99.9333	0.05122
FP rate	0.0023	0.001	0.003	0.00078
Recall	0.9986	0.997	0.999	0.00066
Precision	0.9986	0.997	0.999	0.00066
F-Measure	0.9986	0.997	0.999	0.00066
Kappa statistics	0.99632	0.993	0.9982	0.00134
RMSE	0.04554	0.0402	0.0512	0.00291
Completion time	0.984913	0.04036	1.875	0.580547944

Table 3 Summary of results after feature selection with WOA on Spambase dataset using 20-fold cross-validation

	Average	Min	Max	STD
Accuracy	99.89	99.8667	99.9333	0.02132
FP rate	0.0019	0.001	0.003	0.00054
Recall	0.999	0.999	0.999	0
Precision	0.999	0.999	0.999	0
F-measure	0.999	0.999	0.999	0
Kappa statistics	0.99712	0.9965	0.9982	0.00056
RMSE	0.04036	0.0364	0.0432	0.00196
Completion time	1.875	1.62	0.23229531	0.403

Table 4 Summary of results after feature selection with WOA on Spambase dataset using 66% split

	Average	Min	Max	STD
Accuracy	99.7698	99.6164	99.8721	0.07672
FP Rate	0.0071	0.004	0.011	0.00221
Recall	0.9975	0.996	0.999	0.00103
Precision	0.9975	0.996	0.999	0.00103
F-Measure	0.9975	0.996	0.999	0.00103
Kappa statistics	0.99388	0.9898	0.9966	0.00204
RMSE	0.04197	0.0359	0.0518	0.00534
Completion time	3.091	2.6	4.29	0.474446

Table 5 Summary of results after feature selection with WOA on Enron dataset using 10-fold cross-validation

	Average	Min	Max	STD
Accuracy	99.0952	98.8249	99.3537	0.17665
FP rate	0.0128	0.008	0.02	0.00343
Recall	0.991	0.988	0.994	0.00179
Precision	0.991	0.988	0.994	0.00179
F-Measure	0.991	0.988	0.994	0.00179
Kappa statistics	0.97684	0.9699	0.9835	0.00453
RMSE	0.09213	0.0849	0.0972	0.00398
Completion time	1.03	0.94	1.1	0.05099

Table 6 Summary of results after feature selection with WOA on Enron dataset using 20-fold cross-validation

	Average	Min	Max	STD
Accuracy	99.4301	99.2362	99.5887	0.09492
FP rate	0.0074	0.005	0.014	0.00254
Recall	0.9944	0.992	0.996	0.00102
Precision	0.9944	0.992	0.996	0.00102
F-Measure	0.9944	0.992	0.996	0.00102
Kappa statistics	0.98543	0.9804	0.9895	0.00245
RMSE	0.08235	0.0733	0.0884	0.00414
Completion time	1.07	0.98	1.24	0.075719

Table 7 Summary of results after feature selection with WOA on Enron dataset using 66% Split

	Average	Min	Max	STD
Accuracy	98.1693	97.2366	98.9637	0.57903
FP rate	0.0309	0.02	0.041	0.0078
Recall	0.9817	0.972	0.99	0.00592
Precision	0.9817	0.973	0.99	0.00583
F-Measure	0.9816	0.972	0.99	0.00599
Kappa statistics	0.95322	0.9298	0.9735	0.01471
RMSE	0.13755	0.1237	0.1483	0.00767
Completion time	1.046	1	1.11	0.04274

WOA, 99.89% accuracy was achieved with 10-fold and 20-fold cross-validation and a slight drop to 99.77% when 66% spilt was used. The Enron dataset also displayed a high accuracy of 99.3% with 20-fold cross-validation and 98.2% when run with 66% split test option. Figure 6 shows the accuracy for the different test options.

The false positive (FP) rate is the proportion of the dataset which were classified as spam, but are not spam, among all examples which are not spam. A low FP rate assures the credibility of the classification system. Figure 7 shows a significant drop in FP rate to 0.0019 using 20-fold cross-validation for the Spambase dataset and 0.0074 using 20-fold cross-validation for the Enron dataset.

Precision is the fraction of relevant recollected instances, while recall is the fraction of relevant instances that are recollected. Precision and recall depend on an understanding and measure of relevance. When discussing precision and recall scores, either values for one measure are likened for a specific level at the other measure or both are combined as a single measure. In this research, the *F*-measure is used. A high *F*-measure is required since both precision and recall are desired to be high, and

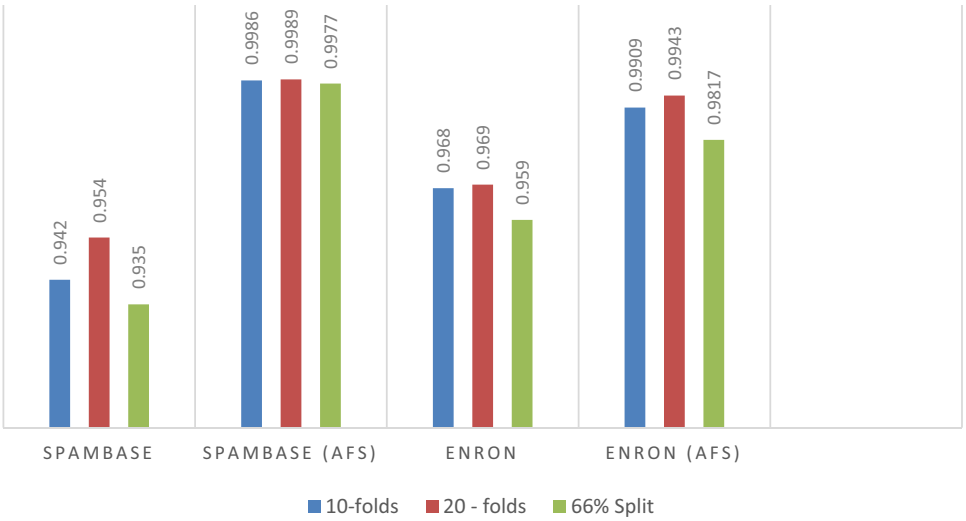

Fig. 6 Accuracy achieved before and after feature selection (AFS) with WOA using different test options on the Spambase and Enron datasets

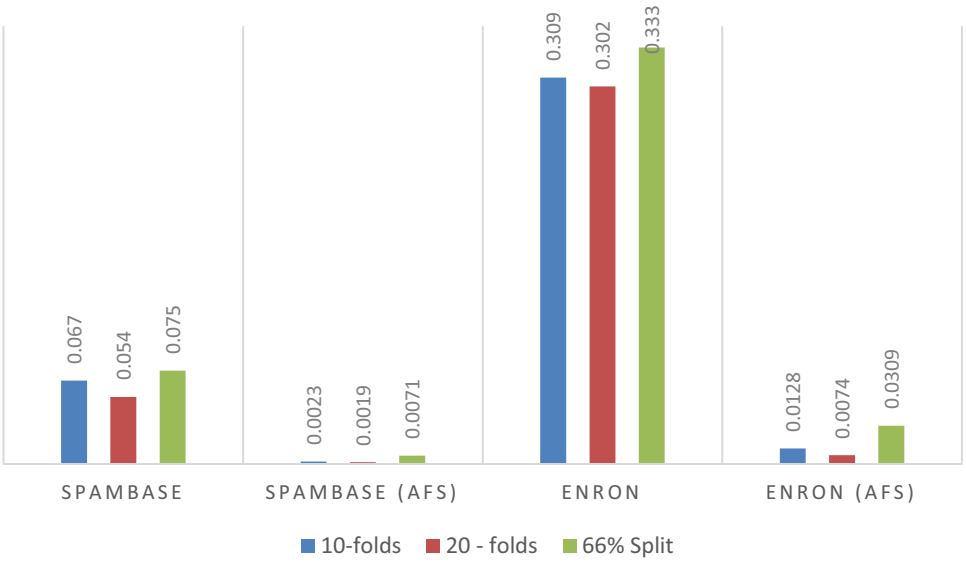

Fig. 7 FP rate before and after feature selection with WOA using different test options on the Spambase and Enron datasets

WOA–rotation forest recorded a high *F*-measure of 0.9990 and 0.9944 for Spambase and Enron datasets, respectively, and this is seen in Fig. 8.

The kappa characteristic gives the level of agreements between the true classes and the classifications. The value 1 is highest showing total agreement; in this study, Spambase dataset showed a high kappa characteristics of 0.9971 which was got when the test was carried out with 20-fold cross-validation, while Enron dataset also with 20-fold cross-validation had 0.9854; Fig. 9 shows the respective kappa characteristics for the three test options used.

According to root mean squared error, a low value is an indication of an excellent classifier. A low value for the root mean square error was achieved using 20-fold cross-validation with 0.0404 on the Spambase dataset and 0.0824 for the Enron dataset. Figure 10 illustrates the root mean squared error.

6.1.2 Performance comparison before and after feature selection

The test was carried out on the datasets before feature selection with the proposed WOA and after feature selection. There is a significant increase in accuracy from 94.2 to 99.89% for the Spambase dataset with a drop in FP rate from 0.067 to 0.0019. Enron dataset recorded improved accuracy from 96.9 to 99.43% and a fall in FP rate from 0.302 to 0.007. The results of the experiment which shows a performance improvement in the metrics are shown in Fig. 11 and Fig. 12 in that other.

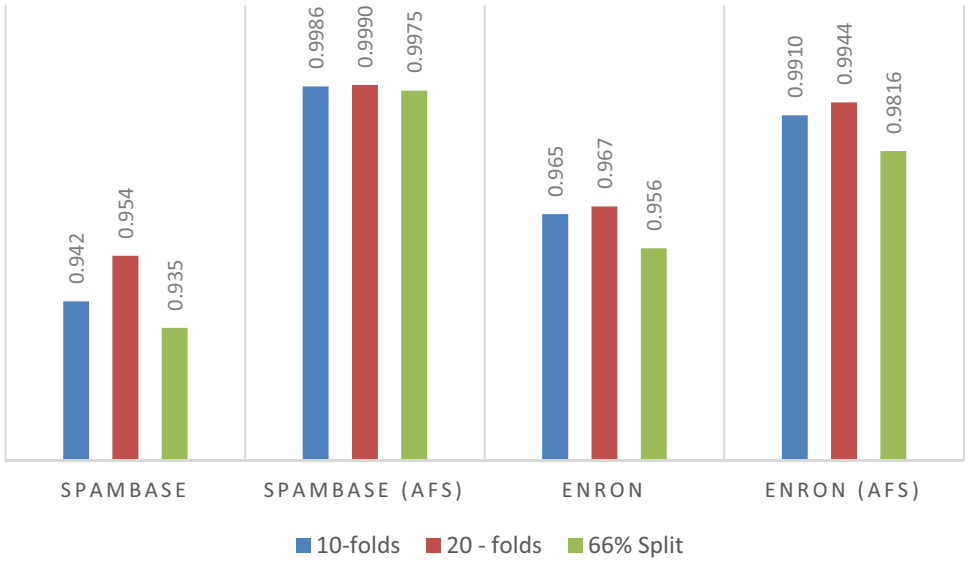

Fig. 8 *F*-Measure before and after feature selection with WOA using different test options on the Spambase and Enron datasets

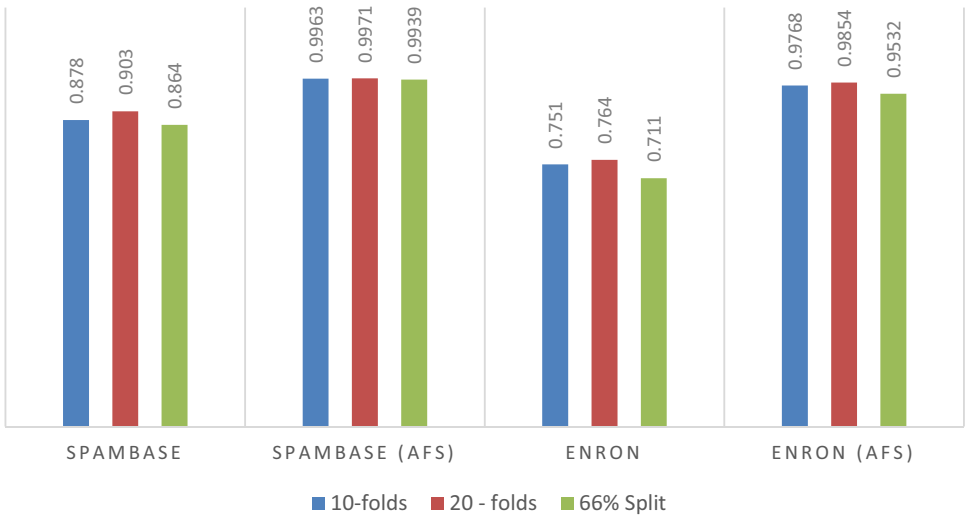

Fig. 9 Kappa statistics before and after feature selection with WOA using different test options on the Spambase and Enron datasets

Fig. 10 RMSE before and after feature selection with WOA using different test options on the Spambase and Enron datasets

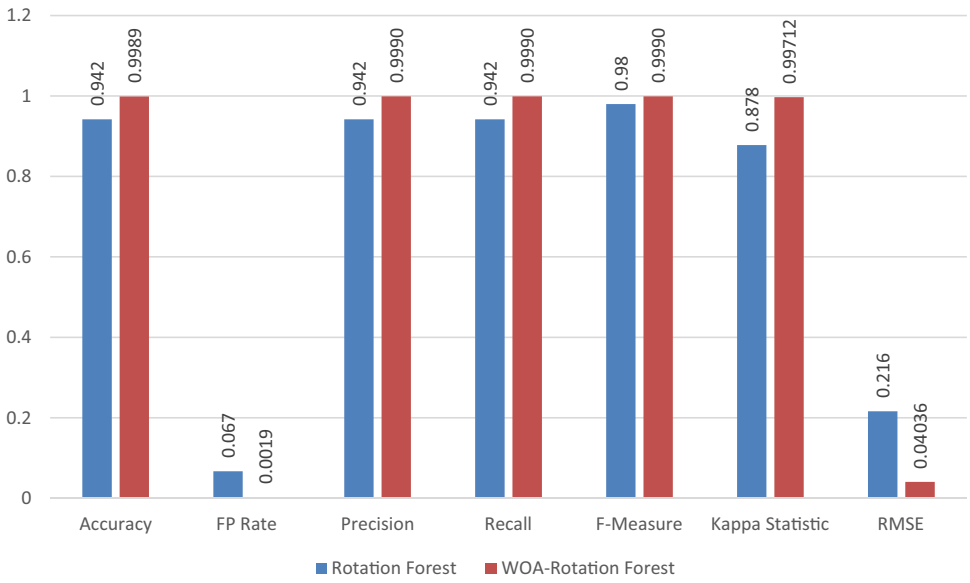

Fig. 11 Performance comparison using Spambase dataset before and after feature selection with WOA

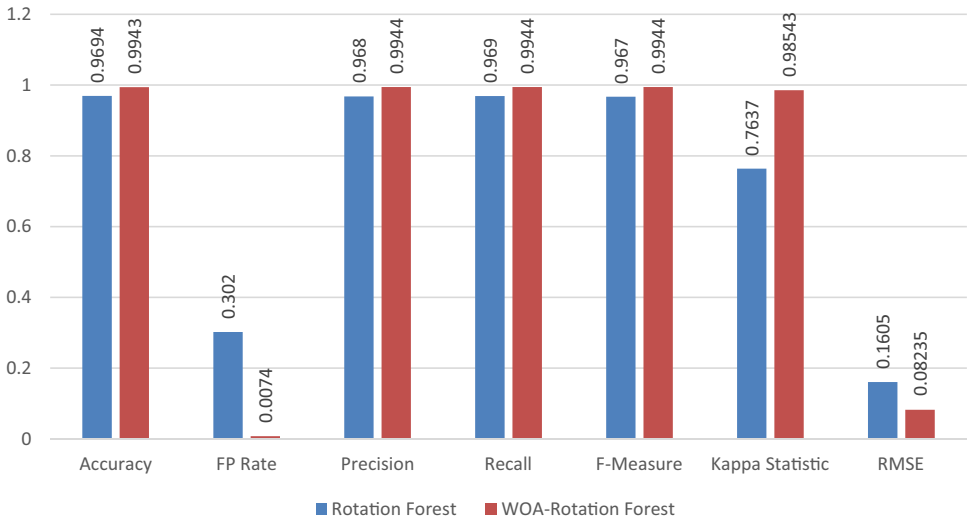

Fig. 12 Performance comparison using Enron-Spam corpus before and after feature selection with WOA

6.2 Comparison of the WOA–RF with related works

Figures 13 and 14 show a comparative analysis in terms of accuracy and false positive (FP) rate of the proposed WOA–RF with other related works.

7 Conclusion and recommendations

Email has continued to be an integral part of our lives and a requirement for any successful communication on the internet. The problem of spam mails occupying a huge amount of space and bandwidth, and the weaknesses of spam filtering techniques which includes misclassification of genuine emails as spam (false positives) are a growing challenge to the internet world. This research work proposed the use of WOA to select salient features in the email corpus and rotation forest algorithm for classifying emails as spam and non-spam. In achieving the aim and objectives of this research, the Spambase dataset from the UCI repository with 58 attributes and 4601 instances (spam and non-spam emails) and the Enron-Spam corpus were used.

The entire datasets were used, and the evaluation of the rotation forest algorithm was done before and after feature selection with WOA using 10-fold cross-validation, 20-fold cross-validation and 66% split test options. The rotation forest algorithm after feature selection with WOA was able to classify the emails into spam and non-spam with a performance accuracy of 99.89% and a low FP rate of 0.0019. The result obtained hence shows clearly that after feature selection with WOA, the rotation

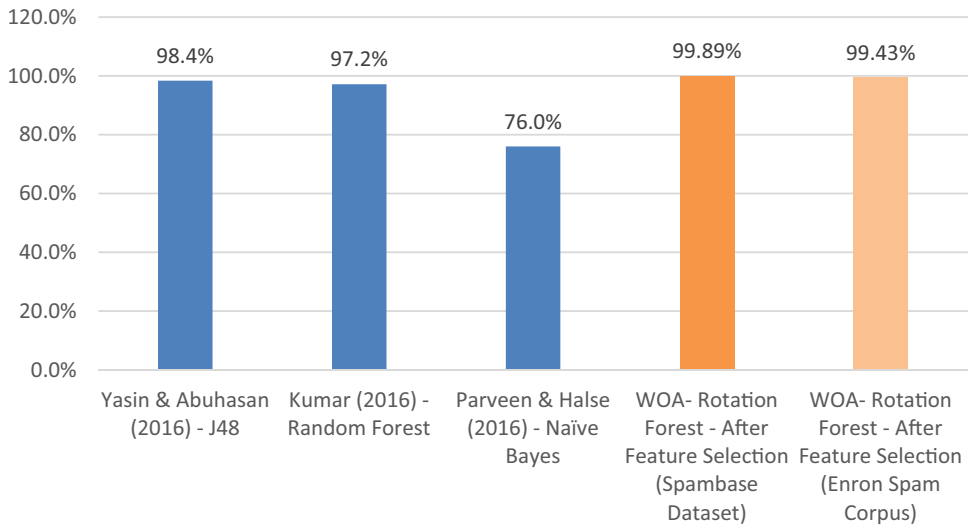

Fig. 13 Comparison of accuracy

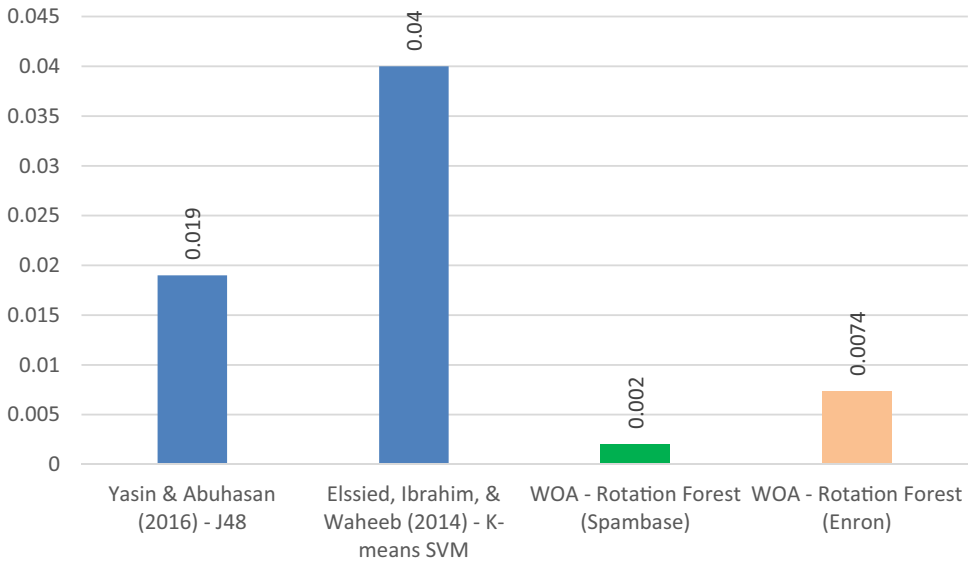

Fig. 14 Comparison of FP rate

forest algorithm outperformed 98.4% [38], 97.2% [13] and 76% [21].

Based on the findings of the study, we make the following recommendations; the WOA toolbox should be added to the WEKA platform and other machine learning tools as an enhancement. The WOA should be tested for feature selection with more datasets.

Compliance with ethical standards

Conflict of interest The authors declared that they have no conflict of interests to this work.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Alguliev, R. M., Aliguliyev, R. M., & Nazirova, S. A. (2011). Classification of Textual E-Mail Spam Using Data Mining Techniques. Appl Comput Intell Soft Comput 1–8. https://doi.org/10.1155/2011/416308
2. Aljarah I, Faris H, Mirjalili S (2016) Optimizing connection weights in neural networks using the whale optimization algorithm. Soft Comput. https://doi.org/10.1007/s00500-016-2442-1
3. Carmona-cejudo JM, Castillo G, Baena-garcía M, Morales-bueno R (2013) A comparative study on feature selection and adaptive strategies for email foldering using the ABC-DynF framework. Knowl-Based Syst 46:81–94. https://doi.org/10.1016/j.knosys.2013.03.006
4. Dondi R, El-Mabrouk N, Swenson KM (2014) Gene tree correction for reconciliation and species tree inference: complexity and algorithms. J Discrete Algorithms 25:51–65

5. Harris A, Yates D (2015) Phishing attacks over time: a longitudinal study. In: Twenty-first Americas Conference on Information Systems, Puerto Rico, 2015, pp 1–6
6. Horng MF, Dao TK, Shieh CS, Nguyen TT (2017) A multi-objective optimal vehicle fuel consumption based on whale optimization algorithm. In: Pan JS, Tsai PW, Huang HC (eds) Advances in intelligent information hiding and multimedia signal processing Smart innovation, systems and technologies. Springer, Berlin, pp 371–380. https://doi.org/10.1007/978-3-319-50212-0
7. Hu H, Bai Y, Xu T (2016) A whale optimization algorithm with inertia weight. WSEAS Trans Comput 15:319–326
8. Ibrahim HO, Laporte G (1996) Metaheuristics: a bibliography. Ann Oper Res 63:513–623
9. Idris I, Selamat A (2014) Improved email spam detection model with negative selection algorithm and particle swarm optimization. Appl Soft Comput J 22:11–27. https://doi.org/10.1016/j.asoc.2014.05.002
10. Idris I, Selamat A, Omatu S (2014) Hybrid email spam detection model with negative selection algorithm and differential evolution. Eng Appl Artif Intell 28:97–110. https://doi.org/10.1016/j.engappai.2013.12.001
11. Kaur A, Kaur J (2015) Spam detection using data mining tool in Matlab. Int J Adv Res Comput Commun Eng 4(8):342–344. https://doi.org/10.17148/IJARCCE.2015.4873
12. Kaveh A (2017) Sizing optimization of skeletal structures using the enhanced whale optimization algorithm. In: Applications of metaheuristic optimization algorithms in civil engineering. Springer, Berlin, pp 47–69. https://doi.org/10.1007/978-3-319-48012-1
13. Kumar M (2016) Effective spam filtering using random forest. Int J Innov Res Comput Commun Eng. https://doi.org/10.15680/IJIRCCE.2016
14. Kumar RK, Poonkuzhali G, Sudhakar P (2012) Comparative study on email spam classifier using data mining techniques. In: Proceedings of the international multiConference of engineers and computer scientists, vol 1, pp 14–16
15. Lacosere B (2016) A hybrid whale algorithm and pattern search technique for optimal power flow problem. In: 8th international conference on modelling, identification and control (ICMIC-2016), pp 1048–1053
16. Latiff MSA, Madni SHH, Abdullahi M (2018) Fault tolerance aware scheduling technique for cloud computing environment using dynamic clustering algorithm. Neural Comput Appl 29(1):279–293
17. Laws S (2017) Spam statistics and facts. Retrieved 9 Nov 2017 from https://spamlaws.com/spam-stats.html
18. Malarvizhi R, Saraswathi K (2013) Content-based spam filtering and detection algorithms—an efficient analysis & comparison. Int J Eng Trends Technol 4(9):4237–4242
19. Mirjalili S, Lewis A (2016) The whale optimization algorithm. Adv Eng Softw 95:51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008
20. Moradpoor N, Clavie B, Buchanan B (2017) Employing machine learning techniques for detection and classification of phishing emails. In: Computing conference. IEEE, pp 149–156
21. Parveen P, Halse PG (2016) Spam mail detection using classification. Int J Adv Res Comput Commun Eng 5(6):347–349. https://doi.org/10.17148/IJARCCE.2016.5674
22. Puri S, Gosain D, Ahuja M, Kathuria I, Jatana N (2013) Comparison and analysis of spam detection algorithms. Int J Appl Innov Eng Manag 2(4):1–7
23. Radicati S (2017) Email statistics report. A technology market research Firm Palo Alto, CA, USA. 2017–2021. https://www.radicati.com/wp/wp-content/uploads/2017/01/Email-Statistics-Report-2017-2021-Executive-Summary.pdf
24. Rathi M, Pareek V (2013) Spam mail detection through data mining—a comparative performance analysis. Int J Mod Educ Comput Sci 5(December):31–39. https://doi.org/10.5815/ijmecs.2013.12.05
25. Reddy PDP, Reddy VCV, Manohar TG (2017) Whale optimization algorithm for optimal sizing of renewable resources for loss reduction in distribution systems. Renew Wind Water Sol. https://doi.org/10.1186/s40807-017-0040-1
26. Sarno DM, Lewis JE, Bohil CJ, Shoss MK, Neider MB (2017) Who are Phishers luring?: A demographic analysis of those susceptible to fake emails. In: Proceedings of the human factors and ergonomics society annual meeting, vol 61, No. 1. SAGE Publications, Sage CA, Los Angeles, CA, pp 1735–1739
27. Sayed GI, Darwish A, Hassanien AE, Pan JS (2017) Breast cancer diagnosis approach based on meta-heuristic optimization algorithm inspired by the bubble-net hunting strategy of whales. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-319-48490-7_36
28. Sharaff A, Nagwani N, Dhadse A (2016) Comparative study of classification algorithms for spam email detection. Springer, (January). https://doi.org/10.1007/978-81-322-2553-9
29. Sharma R, Kaur G (2016) E-mail spam detection using SVM and RBF. Int J Mod Educ Comput Sci 8(April):57–63. https://doi.org/10.5815/ijmecs.2016.04.07
30. Shuaib M, Osho O, Ismaila I, Alhassan JK (2018) Comparative analysis of classification algorithms for email spam detection. Int J Comput Netw Inf Secur 10(1):60
31. Statista (2017) Spam e-mail traffic share 2017_Statistic. Retrieved 9 Nov 2017 from https://www.statista.com/statistics/420391/spam-email-traffic-share/
32. Teli S, Biradar S (2014) Effective spam detection method for email. IOSR J Comput Sci 2014:68–72
33. Toke B, Puri D (2016) Review on spam detection in OSN using integrated approach. Int Res J Eng Technol 3:2539–2542
34. Touma HJ (2016) Study of the economic dispatch problem on IEEE 30-bus system using whale optimization algorithm. Int J Eng Technol Sci 5(June):11–18
35. Trivedi IN, Pradeep J, Narottam J, Arvind K, Dilip L (2016) A novel adaptive whale optimization algorithm for global optimization. Indian J Sci Technol 9(October):319–326. https://doi.org/10.17485/ijst/2016/v9i38/101939
36. Tuteja SK, Nagaraju B (2016) Email spam filtering using BPNN classification algorithm. In: International conference on automatic control and dynamic optimization techniques (ICACDOT), pp 915–919
37. Witten IH, Eibe F (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers, Burlington
38. Yasin A, Abuhasan A (2016) An intelligent classification model for phishing email detection. Int J Netw Secur Appl 8(4):55–72. https://doi.org/10.5121/ijnsa.2016.8405
39. Zavvar M, Rezaei M, Garavand S (2016) Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. Int J Mod Educ Comput Sci 7(July):68–74. https://doi.org/10.5815/ijmecs.2016.07.08