Received August 12, 2019, accepted September 3, 2019, date of publication September 26, 2019, date of current version October 16, 2019.
Digital Object Identifier 10.1109/ACCESS.2019.2944089

Hybrid Water Cycle Optimization Algorithm With Simulated Annealing for Spam E-mail Detection

GHADA AL-RAWASHDEH, RABIEI MAMAT, AND NOOR HAFHIZAH BINTI ABD RAHIM
Department of Computer Science, University Malaysia Terengganu, Kuala Terengganu 21300, Malaysia
Corresponding author: Ghada Al-Rawashdeh (ghada_rwashdeh@yahoo.com)

ABSTRACT Spam is defined as junk and unwanted e-mail. The implementation of a reliable spam email filter becomes more and more important for e-mail users since they have to face with the growing amount of uninvited e-mails. The faults of spam classifiers are characterized by being more and more insufficient to handle huge volumes of relevant emails and to identify and detect the new spam email as example with high performance. The problem in spam classifiers is a huge number of features. Feature selection is an important task in keyword content classification for being among the most popular and effective methods for feature reduction. Accordingly, irrelevant and redundant features that can impede performance would be eliminated. Meta-heuristic optimization is to choose the optimal solution between possible multi-solutions, which respect the aim of this research that is the performance. The other problem is related to ambiguity of the effect of optimization feature selection on multiple classifiers algorithm which are popular used by previous work namely; K-nearest Neighbor, Naïve Bayesian and Support Vector Machine. Therefore, the aim of this research is to improve the accuracy of feature selection by applying hybrid Water Cycle and Simulated Annealing to optimize results and to evaluate the proposed Spam Detection. The methodology used in this study which consists of groundwork, induction, improvement, evaluation and comparison quality. The cross-validation was used for training and validation dataset and seven datasets were employed in testing the spam classification proposed. The results demonstrate that the meta-heuristic namely water cycle feature selection (WCFS) was employed and three ways of hybridization with Simulated Annealing as a feature selection employed. In comparison with other feature selection algorithms such as Harmony Search, Genetic Algorithm, and Particle Swarm, the hybridization interleaved hybridization outperformed other feature selection algorithms with accuracy 96.3%, on the other side the effect of using three classifier algorithms, the SVM was better than other of classifier algorithms with f-measurement 96.3%. The number of features using interleaved water cycle and Simulated Annealing the number of features has decreased to more than 50%.

INDEX TERMS Water cycle algorithm, classification algorithm, spam email, simulating annealing, hybridization., global search, local search.

I. INTRODUCTION

Spam is a major email services-associated problem globally, and it encompasses unsolicited/unwanted emails which do not have a specified/intended receiver but are pushed out for different purposes, ranging from marketing to scam and hoax. About 97% of emails sent or received in 2009 were classified as spam emails; thus, many recent studies have been focusing on emails classification in recent times. Presently, the struggle between spam detection tools and spammers is a continuous fight as each side strives to explore new ways of annulling the presence of the other [1], [35].

Classification is a supervised learning task that identifies a problem in a way that all the previous categories that are available in the existing problem are based on the observations, in a bid to create a new training-based capability set. This set is responsible for the collection of data and recording of the values that are not connected to the available set under consideration. Several real-world problems can be modelled as classification problems and can be identified with a particular solution. An example of such problems can be obtained from the current study on spam email

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Asif Naeem.

classifier, the category for spam or non-spam type classes. Since spam classification is a supervised kind of learning, many machine learning processes/algorithmic methods (such as K-Nearest neighbor), regression models (such as Naive Bayes classifier, the various decision trees, some inductive rule-based learning, the neural networks), or others (such as Support Vector Machines) can be employed [2]. However, most classification data are highly dimensional, and natural dimensionality reduction may be necessary for efficiency and accuracy. So, the major problem of content classification is its high dimensionality. High dimensionality or a large number of features can act as one of the features with space addresses (a very large collection of vocabulary that consists of all the special terms that occur at least once or more than once in the collection of emails). This problem influences and degrades the performance of the entire system, as well as the performance of most content classifiers. Additionally, it will increase the entire complexity of the system. To manage and deal with high dimensionality issue, as well as to avoid its effects, dimensionality reduction is strongly required [3]. This study is focused on the dimensionality of spam email classifiers. Hence, the mechanism of feature selection can be a curse to the dimensionality of the selection of relevant features and its classification. However, with the elimination and the reduction of the redundant features, several features can be reduced, and the training time can be increased. This will improve the classification performance. This study discussed the various limitations of the popular algorithms used in previous studies of feature selection. Three popular traditional classifiers were highlighted namely KNN, NB, and SVM, using seven datasets. This paper discussed the proposed WCFS, evaluated its hybridization with Simulated Annealing in terms of performance, and benchmarked it against three popular feature selection algorithms (GA, PSO, and HS).

The remaining part of this paper was arranged thus: the previous works related to this study were presented in section 2 while the proposed WCFS and SA and their parameters setting were discussed in section 3. Section 4 detailed the proposed hybridization approach while section 5 presented the results of the experiments and their analysis. The last section presented the conclusions.

II. RELATED WORK

The importance of automatic email classification technologies has increased recently due to the rapid development of the Internet and the increased demand for online information classification. However, the increased number of emails pushed out daily has made it almost impossible to classify them into real and spam categories. Therefore, several studies have been conducted on spam classifiers [4]–[8] and at present, more studies have been focusing on improving the classification performance of spam classifiers. Typically, text can be represented as a set of features using either Bag-Of Word (BOW) method (where single words or phrases are used as features) or n-gram method (where word sequences are used) [9]. However, the major problem of the TC system is on how to manage the huge number of features (usually in the orders of tens of thousands) [10].

Several IR techniques, such as Stemming, Stop-words Removal, and Feature Selection (FS), have been applied for feature space dimension reduction. Feature space dimension reduction by the FS techniques is achieved by eliminating the redundant or irrelevant features for a particular category [11]–[14]. Some of the existing FS techniques are Information Gain (IG), Mutual Information (MI), GSS Coefficient (GSS), Chi-Square Statistic (CHI), and Odds Ratio (OR). Therefore, the major problem in spam classification is the high number of features which has attracted the use of several techniques such as Chi-square to reduce them without causing a significant reduction in performance [34]

Feature selection algorithms are problem-dependent and do not rely on feature dependencies [18]. This has motivated researches of other FS techniques such as metaheuristics-based feature selection. The drawback of the metaheuristics-based feature selectors is their weakness in local search; they also have a slow convergence rate and more control parameters. Another problem of the metaheuristics-based feature selectors is the ambiguity of optimization feature selection on different classifiers [16], [17]. Therefore, the major gap in studies related to spam email analysis is the weak performance of optimization-based classifiers. Consequently, it is imperative to develop novel approaches for effective spam email identification, with the aim of supporting spam email detection.

Although the new Water Cycle Algorithm can address the issue of entrapment at local optima, its effectiveness as a spam classifier is still ambiguous. The WCA has only 3 control parameters while other optimization algorithms, such as the Harmony Search algorithm, have seven control parameters. WCA as a feature selector can address the issue of fast convergence of local optima entrapment using an evaporation technique [18].

The investigation and creation of novel approaches for tackling the problems of FS and cruse dimensionality are still active areas of researches, particularly for spam classifiers. Hence, the FS approaches are considered for the following reasons: (a) improvement of performance (maybe the predictive accuracy or the learning speed); (b) simplification of data for model selection; and (c) redundant or irrelevant features elimination (dimensionality reduction) [19], [20].

Several studies have been conducted over the past 10 years based on feature selection techniques. This is because this technique has been a very eminent solution to many problems. Studies have also focused on improving the efficiency of this technique. The various related technologies and methods based on metaheuristic techniques have been discussed in this section, including the local search and population-based methods for both heuristics and hybrid metaheuristics. Metaheuristic-based techniques depend on the computational methods and have been proven as effective techniques since they can repeatedly optimize a problem by improving the employed candidate solution. The improvement of the employed candidate solution consequently improves a given measure of quality [21].

The core idea of Simulated Annealing (SA) developed by [22] was inspired by the hill climbing-based methods using the escape probability from a local optima problem. SA is initiated by randomly finding the number of any available best solution before generating the neighbouring best solution for the same general numbers. Having created all the values surrounding the neighbouring number, a check of the credentials is first conducted before the algorithm searches for the best suitable number in the next phase. Despite the acceptability of a worst-case solution with a certain degree of probability, this acceptance probability is dependent on its value from the Boltzmann probability, $\mathrm{P} = \mathrm{e} - \theta/\mathrm{T}$, where θ = difference in the fitness function evaluation between the trial solution (Soltrial) and the best solution (Solbest), whereby T denotes the temperature which decreases periodically during an active search based on some cooling schedule.

A simulated annealing concept (SimRSAR) has been used by [23] to solve various attribute reduction problems. In SimRSAR, all the states that represent a given set of attributes are duly considered. The selected attribute can find the neighbouring attributes in the algorithm. Furthermore, the mutation process will be implemented on the algorithm and the final three attributes will then be selected. The initial systems' temperature can be determined using the formula: $2^*|\mathrm{C}|$, where $|\mathrm{C}|$ is used to determine the number of the available attributes for a given data set, while the cooling schedule value is given as $\mathrm{T}(\mathrm{t}+1) = 0.93^*\mathrm{T}(\mathrm{t})$. This method is proposed as a compatible approach with all its outcomes because all the discussed techniques in the existing literature focused on finding the least number of attributes that will be comparable with this method.

A study by [24] used Wheal Optimization with SA as a feature selection algorithm. The results were compared with the other population-based techniques and found to perform better than population-based in terms of data classification.

Several studies have hybrid single-based approaches with other methods to enhance efficiency. For instance, HC has been hybrid with GA where HC was applied to search for the local optima is carried out immediately after each mutation or solution recombination (crossover) of the operators. Moreover, many metaheuristics, such as TS, SA, Ant systems, and Mas are based on HC. Belda-Lois *et al.* [25] used the HC algorithm to improve the phase of GRASP as the initial point to the solution. As such, it can be concluded that there are weaknesses in the existing text classifiers. Therefore, the current study adopts the local search hybridization with global search to ensure a balance between the exploration and exploitation capabilities during feature selection. Local search is better than population search in exploitation while population search is better than local search in exploration [24].

WCA optimization is comparable to other metaheuristics. In WC algorithm, the initial population called raindrops is first introduced and the algorithm assumes that there is rain or precipitation. The best raindrop is regarded as a sea while a collection of raindrops is considered as a river. The remaining raindrops are regarded as streams that flow into the sea and rivers. Each river draws water from the streams depending on the flow magnitude. However, the volume of stream water entering a river differs from that entering other streams. Additionally, river water flows into the sea down the hill.

In this algorithm, evaporation is the major influencing event because it can prevent the algorithm from fast convergence. Naturally, water is lost from rivers and lakes via evaporation but during photosynthesis, water is released by plants. The water that is lost through evaporation processes, upon entering the atmosphere, forms clouds. Later the clouds, upon contact with the colder atmosphere, condenses and returns as rain. The rain accumulates into streams, and streams flow into rivers, and rivers into seas [18]. As rivers/streams flow into the sea, water is also lost to evaporation processes. To avoid local optima entrapment in the WCA, this assumption is often adopted.

WCA was introduced to handle several engineering designs and optimization problems [18] because it can efficiently handle such problems. WCA has also been proven as an attractive approach based on the statistical results of the comparison of its efficiency with several optimization techniques, including GA, Stochastic Ranking (SR), Homomorphous Mappings (MH), Harmony Search, PSO, and Artificial Bee Colony (ABC). Furthermore, the findings showed that WCA achieved better solutions compared to other optimizers. This is in addition to the accuracy of the algorithm in terms of the number of evaluation functions for each problem. It is also empirically demonstrated that WCA can offer competitive solutions compared to most other metaheuristics. Nevertheless, the computational efficiency of WCA and the quality of its solution can be affected by the nature and complexity of the underlined problem. Furthermore, the accuracy and fluency of different metaheuristic-based approaches can also be affected by the nature and complexity of the problem at hand.

III. METHOD

A. THE PROPOSED WCA FOR FEATURE SELECTION

This study proposes the use of WC-based algorithms for feature selection. The proposed WCA, like other metaheuristics, initiates with an initial population referred to raindrops, where the best raindrop is designated as a sea. Then, a collection of good raindrops via a feature selection process forms a river while the rest of the raindrops are classified as streams that flow into the sea/rivers. The flow magnitude determines the movement of water from the streams into the rivers as described subsequently. The volume of water that enters the rivers from the streams varies from stream-wise. Rivers also flow downhill into the sea [18]. Notably, WCA is yet to be recognized as an efficient feature selection technique in spam classifiers despite its ability to prevent algorithmic rapid

1	2	3	4	5	6	7	8	9	10	11	12
1	1	0	0	1	0	0	0	1	0	0	1

FIGURE 1. Representation of the selected features.

convergence and local optima entrapment using evaporation technique [18].

HSFS technique can be considered weaker in terms of local searches but has a slow convergence rate. Owing to this factor, it can easily run into various local optimum conditions instead of striving for global optimum conditions [26]. In general, a huge drawback of HSFS is that it employs several control parameters which are not used by WC algorithm. There is only one control parameter in WCA while there are 3 in the HSFS, which are phmcr, p par, and bw [30]. WCA was developed to address some of the problems of the existing meta-heuristics [27].

In the proposed algorithm, the vector-space model is used to represent the emails such that every term represents one dimension of the multi-dimensional term spaces, and each email di = (wi1, wi2 … win) is regarded as the vector with n different terms in the term space. Moreover, each possible email detection solution is considered as the vector of features, hence, feature selection problems are projected as optimization tasks that mainly strive to locate the optimal features rather than considering all the features. Therefore, the quality of feature selection was considered as the fitness function while WCFS was used to optimize the objective. This approach is basically beneficial owing to the explicit nature of training and testing of the classifiers which, in turn, provides a better view of the classifiers' performance on certain data types, thereby allowing the utilization of task-specific classifier objectives. Another interesting aspect of this approach is that several objectives can be simultaneously considered [18]. When using a multi-purpose metaheuristic for feature selections, there is a need to make several design choices, predominantly the fitness function and the problem representation, as both have significant effects on the classification quality and optimization performance.

Each possible solution for features selection. So, it can be considered that feature selection problems can be projected using optimization as the problem mainly deals with finding the local available optimum value from a set of findings. In this regard, it can be considered that finding the feature quality as a result of the fitness function and finding the value of the raindrops as a feature selection mechanism can be applied on the said problem to find the local optimum and the most appropriate value. One can consider this approach to be beneficial because of these factors. The available features are explicitly for the selection of the fitness function. This feature which is available for the selection of the objection function can be used to achieve a better understanding of the performance of the algorithm, using several available features in this algorithm.

The proposed algorithm uses several representations to code the whole F of the features in a vector of length m, where m = number of the features as shown in Figure 1. Each element of this vector encompasses a label showing whether the features are selected or dropped. An example of the representation of solutions is illustrated in Figure 1. In this case, 12 features {1, 2, 5, 9, and 12} were selected while others {3, 4, 6, 7, 8 10, 11} were dropped.

1) CREATE THE INITIAL FEATURES

The values of the problem variables are formed in arrays. Such arrays are called 'Chromosome' in GA and 'Particle Position' in PSO. Hence, the label in the proposed method is called 'Raindrop features' for a single feature. A raindrop in a Nvar dimensional features selection problem is an array of 1 × Nvar. This array is defined as follows:

$$Raindrop\ feature = [X1, X2, X3 \ldots XN] \quad (1)$$

The classification algorithm is initialized by generating a candidate which represents a matrix of raindrops of size Npop × Nvar. Thus, the randomly generated matrix X is given as (rows = number of features selection, columns = number of design variables):

$$Raindrops\ of\ feature = \begin{bmatrix} Raindrop_1 \\ Raindrop_2 \\ Raindrop_3 \\ \vdots \\ Raindrop_{Npop} \end{bmatrix} \times \begin{bmatrix} x_1^1 x_2^1 x_3^1 & \cdots & x_{Nvar}^1 \\ \vdots & \ddots & \vdots \\ x_1^{Npop} x_2^{Npop} x_3^{Npop} & \cdots & x_{Nvar}^{Npop} \end{bmatrix} \quad (2)$$

The value of each decision variable (X1, X2, X3…XNvar) can be represented as either 0 or 1, where Npop = number of raindrops, and Nvars = number of design variables. Npop raindrops are generated and the cost of a raindrop can be achieved by evaluating the following cost function (Cost):

$$Cost_i = f(x_1^i, x_2^i, \ldots x_{Nvar}^i) \quad i = 1, 2, 3, \ldots, Npop. \quad (3)$$

2) FITNESS FUNCTION

To maintain a proper balance in between all the particular selected features which are available as a part of the solution in each of the minimum solutions and to provide maximum accuracy for the particular feature selection, the fitness function in Eq4 is used in both WC and SA algorithms to evaluate search agents.

$$fitness = \alpha \gamma_R(D) + \beta \frac{|R|}{|N|} \quad (4)$$

where γ_R (D) represents the classification error rate of a given classier (the each of three classifiers is used here). Furthermore, |R| is the cardinality of the selected subset and |N| is the total number of features in the dataset, α and β are two parameters corresponding to the importance of classification quality and subset length, $\alpha \in [0, 1]$ and $\beta = (1-\alpha)$ adopted from [24], [31]

3) EVALUATION OF SOLUTIONS

A number of Nsr is chosen from the best individuals (minimum values) as sea and rivers, where the raindrop with the least value is considered as a sea. In fact, a summation of the number of rivers is designated as Nsr while a single sea is given in Equation5. Equation6 below is used to calculate the rest of the initial features.

$$Nsr = Number\ of\ Rivers + (Sea = 1) \tag{5}$$

$$NRaindrops = Npop - Nsr \tag{6}$$

The following equation is used to allocate a given raindrop of feature to the rivers and sea-based on the flow intensity, as follows:

$$\mathrm{NSn} = \mathrm{round}\left\{\left|\frac{Cost_n}{\sum_{i=1}^{N_{sr}} Cost_i}\right| \times N_{Raindrops}\right\}, \quad \mathrm{n} = 1, 2, \ldots, \mathrm{Nsr} \tag{7}$$

where: NSn is the number of streams flowing to the specific rivers or sea [18].

4) STREAM FLOW TO THE RIVERS OR SEA

A collection of raindrops forms the streams while a collection of streams forms new rivers. It should be noted that some streams may have a direct connection with the sea, but all rivers and streams ultimately flow into the sea (best optimal selected features). The flow of streams into the sea over a randomly chosen distance is given as:

$$\mathrm{X} \in (0, C \times d), \quad C > 1 \tag{8}$$

where: C is a value ranging between 1 and 2 (near to 2), but the optimal value of C can be set to 2. The current from the stream to the river is given as d. In Equation 8, X represents a uniformly or randomly distributed number between 0 and (C × d). When the value of C is greater than 1, streams will flow in different directions into the rivers. This concept is also applicable to rivers flowing into the sea. Hence, the new stream and river positions can be calculated as:

$$X^{i+1}_{Stream} = X^{i}_{Stream} + r \quad and \ \times C \times (X^{i}_{River} - X^{i}_{Stream}) \tag{9}$$

$$X^{i+1}_{River} = X^{i}_{River} + r \quad and \times C \times (X^{i}_{Sea} - X^{i}_{River}) \tag{10}$$

where: rand represents a uniformly distributed random number ranging from 0 to 1. If a stream provides a better solution than its connecting river, their positions will be exchanged, i.e., the stream will become river and river will become stream. This form of positional exchange can also happen between rivers and sea [18].

5) EVAPORATION CONDITION

Evaporation is a significant factor for preventing the algorithm from fast convergence [18]. Naturally, water is lost from rivers and lakes via evaporation but during photosynthesis, water is released by plants. The water that is lost through evaporation processes, upon entering the atmosphere, forms clouds. Later, the clouds condense in the colder atmosphere and release back the water in the form of rain. The rain accumulates into streams and streams flow into rivers, and rivers into seas [28]. In the proposed method, it is assumed that water is lost to evaporation processes as the rivers/streams flow into the sea. A demonstration of the process of determining whether or not the river flows into the sea is given by the following pseudo-code:

$$\text{If } \left|X^{i}_{Sea} - X^{i}_{River}\right| < d_{max} \quad i = 1, 2, 3, \ldots, N_{sr} - 1 \quad \textit{Evaporation and raning process} \text{ end} \tag{11}$$

where: dmax = a small number (almost 0). When a river has finally joined the sea, the distance between the river and the sea will become less than dmax. The evaporation process is applied here, and as obtainable in nature, rain usually follows an appreciable rate of evaporation. The search is reduced by a large value of dmax while the search intensity near the sea is facilitated by a small value of dmax. Thus, the search intensity near the sea is controlled by the value of dmax (the optimum solution). The value of dmax can be adaptively reduced as follows:

$$d^{i+1}_{max} = d^{i}_{max} - \frac{d^{i}_{max}}{\max iteration} \tag{12}$$

6) RAINING PROCESS

The raining process is applied after fulfilling the evaporation process. During this process, streams are formed by the new raindrops in different locations (in a similar manner as the mutation operator of GA). The following equation is used to specify the new locations of the newly formed streams:

$$X^{new}_{Stream} = \mathrm{LB} + \mathrm{rand} \times (\mathrm{UB} - \mathrm{LB}) \tag{13}$$

where: LB = lower bound and UB = upper bound; both are problem-specific. Again, the best newly formed raindrop is considered as a river that flows into the sea while the remaining new raindrops are regarded as some new streams that flow into the rivers or straight into the sea. Equation 13 is deployed only for streams flowing directly into the sea to enhance the convergence rate and the algorithmic computational performance for constrained problems. This equation is mainly for encouraging the generation of streams that flow directly into the sea. The aim is to improve the search near the sea (the optimum features) in the feasible region for constrained problems [18].

$$X^{new}_{Stream} = \mathrm{Xsea} + \sqrt{\mu} \times r\ and(1, N_{var}) \tag{14}$$

where: μ = a coefficient that depicts the search range near the sea, while Randn = a normally distributed random number.

The possibility of exiting the feasible region is increased by a larger value of μ while a smaller value of μ limits the search space near the sea. Normally, a suitable value for μ is set to 0.1. The term $\sqrt{\mu}$ in Equation 14 mathematically represents the standard deviation while the concept of variance is defined by μ. With these concepts, generated individuals with variance μ are distributed around the best obtained optimum solution (sea) [18].

7) CONVERGENCE CRITERIA

The water cycle as feature selection stops when either there is no change in the average fitness by a predetermined value ε = dmax after several iterations or a predetermined number of generations has been achieved.

B. SIMULATED ANNEALING

Simulated annealing algorithm, as proposed by [22], can be considered a single heuristic solution which is available on the basis of Hill Climbing methodology. The simulated annealing approach can be used in order to overcome the problem of stagnation in the local optimal value. This algorithm makes use of the concept of a certain probability that can be used to accept the worst solution in any case. The proposed algorithm starts with a particular value which is random in nature and generated with the initial solution stage. However, as the iteration goes on, every neighbouring solution can be taken into consideration to find the best so far generated value according to a predefined value for the neighbouring structure which can be evaluated with the help of a fitness function. The improvement which considers the neighbours, as always fitted into the original solution space, can be always accepted; however, the worst solution, called the worst neighbour value, has the Boltzmann probability $P = e - \theta/T$, where θ is the difference between the fitness of the generated neighbour (TrialSol) and the best solution (BestSol). Moreover, T is the temperature which decreases periodically during an active search with respect to some cooling schedules. In this study, the initial temperature was set to 2*|N|, where |N| = number of attributes per dataset, while the cooling schedule was calculated as T = 0.93*T [23], [24]. The pseudo-code of SA is depicted in Figure 3.

1) TOURNAMENT SELECTION

Tournament selection is a very simple and very straightforward process that is adapted and applied for(Goldberg et al. 1989). It is regarded as one of the Simplest mechanism switches applied for the selection and can be used for very helpful That Is Proposed by[33]. In tournament selection, n different types of solutions can be taken into account and finally selected from a particular object. These solutions are based on the definition of the existing spaces, are compared to each other and will be determined by the winner."

This particular tournament will be helpful in include a specific random number among zero and one that used for the further processing of the algorithm. The tournament includes generating a random number between 0 and 1, after the selection of a specific element, which is implicated from the pool of all the prospective values available the application of the selection pressure is applied to the system algorithm (usually set to 0.5 [24]. the result that can be derived from the value of the so selected variable can be drawn in this format, if the value of the valuables required and which is created from the random number is more than the value of the highest with this value, then this particular value will be beholding to be selected with higher probability, or else if the value is lower than it will be rejected with a lower probability of acceptance [33].

TABLE 1. SCENARIOS for the analysis of WCFS convergence behaviour.

SCENARIO	NPOP	NSR
S1.1	8	2
S1.2	8	4
S1.3	8	7
S2.1	14	2
S2.2	14	4
S2.3	14	7
S3.1	28	2
S3.2	28	4
S3.3	**28**	**7**

C. PARAMETER SETTINGS

A better available approach can be applied based on the interaction of the three classifiers in the algorithm. For this approach, all the datasets were divided for cross-validation. Finally, the cross-validations were divided for evaluation in the same manner as in [29]. In K-fold cross-validation, K-5 folds are used for training and validation even though the remaining folds can be used for testing purposes. A total of M iterations can be applied for this process, and every single optimizer unit can be evaluated K*M times for each dataset. As a matter of resumption of the data used for training, the validation should be equal in size and all the parameters must be set as follows: The best results can be obtained whenever the maximum numbers of iterations are 50 as shown in APPENDIX A. The random drops must be used. All the SA parameters are similar to those previously discussed in the preceding subsection and can be considered the same way as they are created in the previous section. This section studied the algorithmic evolution over generations under two important parameters setting (Npop and Nsr). Here, Nsr is the sum of the number of rivers (a user parameter) and a single sea, while Npop is the number of raindrops (initial features population). Hence, the effects of single parameter changes will be highlighted in this section. The three scenarios presented in Table1 are tested in this section. Furthermore, it has been empirically demonstrated that Npop and Nsr linearly relate to the number of features, and this relationship can give the best outcomes. For each scenario, a fixed number of 50 iterations was tested for all runs, and the value of the fitness function is a measure of the cost value of each solution. The algorithm used for the evaluation was WCFS which was described in Section 3; dmax = 1E03. Each case design was executed twenty times, with the repetition numbers set

```
• Set user parameter of the WCA: Npop, Nsr, dmax, and Maximum_ Iteration.
• Determine the number of streams (individuals) which flow to the rivers and sea using Eq s.
(5) and (6).
• Create randomly initial population of feature selection.
• Define the intensity of flow (How many streams flow to their corresponding rivers and sea)
 using Eq. (7).
   while (t < Maximum_Iteration) or (any stopping condition)
      for i = 1 : Population Size ( Npop)
Stream flows to its corresponding rivers and sea using Eq (9) and (10)
Calculate the Fitness function of the generated stream using Eq. (4).
                    if F_New_Stream < F_river
                       River = New_ Stream;
                    if F_New_ Strea m < F_Sea
                        Sea = New_ Stream;
                                end if
                                end if
River flows to the sea using Eq. (11 )
Calculate the fitness function of the generated river
               if F_New_ River < F_Sea
                   Sea = New_ River;
                         end if
                        end for
 for i = 1 : number of rivers ( Nsr )
     if (distance (Sea and River) < dmax) or ( rand < 0.1 )
            New streams are created using Eq. (12)
                          end if
                         end for
Reduce the dmax using Eq. (13)
   end while
```

FIGURE 2. The pseudo-code of WCA.

```
Initialize (temperature T, random starting features selected)
While cool_iteration <= max_iterations
     Cool_iteration= cool_iteration +1
      Temp_iteration=0
       While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
         Calculate the fitness function of the generated
              current_cost (of this new point) using Eq. (4).
                   δ= current_cost-previous_cost
                       if δ<0, accept neighbor
              else, accept with probability exp(-δ/T)
                               end while
                         (0<α<1)      T=α*T
end while
```

FIGURE 3. The pseudo-code of simulated annealing approach.

to fifty for all runs. Based on the experiments, case S3.3 was chosen to carry out the tests in this section; the parameter was set to Npop = 28 and Nsr = 7. After the parameter setting this research used the same Npop and Number of Generation (number of iteration) value for all optimizations features selection as WCA and left the other parameter value as a previous works like (HRCR, Constant 2, Constant 1, weight;No of Selection, PAR max, PAR min, initial temp, temp Reduction rate). Table2 shows the parameter value for all optimizations features selection.

Low-Level hybridization	The interleaved hybridization	High-Level hybridization

Datasets →WCC→ optimal features
 	
(WCC) optimal features → SA → optimal features
 	
F-measurement, accuracy, precision and recall → Final results
	Datasets →WCC→ features selection
	If (WCC) optimal features → SA by (fitness function) > WCC by (fitness function)
	Return step 2
	Else
	F-measurement accuracy, precision and recall → Final results
		Datasets →SA features selection → optimal features
		
		SA optimal features → WCC
		
		F-measurement accuracy, precision and recall → Final results

FIGURE 4. The flow chart of the main difference between three hybridization.

TABLE 2. The parameter value for all OFS.

OPTIMIZATIONS	POP SIZE	MUTATION RATE	NUMBER OF GENERATION	NO OF SELECTION	WEIGHT	CONSTANT 1	CONSTANT 2	HRCR	PAR MAX	PAR MIN	INITIAL TEMP	TEMP REDUCTION RATE	NSR
GA	**28**	0.5	**50**										
PSO	**28**		**50**	3	0.2	2	2						
HS	**28**		**50**					0.6	0.9	0.45			
SA	**1**		**50**								0.1	0.99	
WCA	**28**		**50**										**7**

IV. HYBRID WC-SA METHOD

WC algorithm is a new concept which has produced varying to several optimization problems, and the main algorithm uses the blind operation to operate each operator. This operator takes the place of exploitation notwithstanding the solution and operations' fitness value. The study depicted in this step replaces the idea of this operator with a local search which considers a simple solution at the initial state; work on the solution to find the next stage; finally, replace the original solution with the actual result. The hybridization of these two algorithms (WC and SA) produces a more sophisticated hybridization model, and as such, was considered in this study. The difference between the three hybridization is highlighted in Figure 4.

A. THE LOW LEVEL OF WCS

A hybrid approach which uses the SA to replace the refining stage in WC was presented in this section. The explorative capability of WC and the speed of SA were explored in the hybrid algorithm in refining solutions. There are two modules in the hybrid WC algorithm- WC module and SA module. The optimum region is found by WC while SA finds the optimum features. With this hybridization, the right balance can be achieved between global exploration and local exploitation. The WC module is engaged in the global searching stage while the SA module is involved in refining the local stage. The WC module was earlier executed for a short period of 1–50 iterations to determine the vicinity of the optimal solution via a global search and to minimize the utilization of computation resources. The outcome of the WC module will serve as the initially selected features for the SA module when refining and generating the final result. APPENDIX B referred Pseudo-code of the WC with SA low-level hybridization.

TABLE 3. Confusion matrix.

Predicted Class		
True Class	**Positive**	**Negative**
Positive	**TP**	**FN**
Negative	**FP**	**TN**

B. THE INTERLEAVED HYBRIDIZATIO

The local method in the hybrid algorithm was embedded in WC. After each round of iteration, the best vector from the Npop is adopted by SA as the initial point. If the fitness value of the locally optimized vectors is better than those in Npop, the Npop will be updated. This process is continued until the stopping point is reached. APPENDIX C referred Pseudo-code of the WC with SA Interleaved hybridization.

C. THE HIGH LEVEL OF WCSA

A one-step SA algorithm was introduced to enhance the algorithm. The WC operations are implemented to generate new features selection solution and the subsequent process is implemented on the new solution. The explorative

TABLE 4. Summary of description of spam datasets.

Document set	Source	# of emails
DS1	Spam Base http://archive.ics.uci.edu/ml/datasets/Spambase	Total 4601 emails (spam = 1813 and ham = 2788)
DS2	Enron Spam Corpus http://www.aueb.gr/users/ion/data/enron-spam/	Total 30041 emails (spam = 13496 and ham = 16545)

TABLE 5. The result of low-level (WCA-SA).

DS		SVM			NB			KNN		
	Original # of features	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement
Enron1	16383	8142	0.922	0.923	7829	0.689	0.703	7721	0.90	0.899
Enron2	11514	5681	0.948	0.955	7125	0.652	0.716	5420	0.858	0.86
Enron3	16382	7146	0.931	0.948	5125	0.668	0.682	7824	0.865	0.798
Enron4	15456	6582	0.865	0.868	6144	0.605	0.677	7270	0.734	0.807
Enron5	14696	6783	**0.959**	0.95	7086	0.619	0.716	7007	0.939	0.946
Enron6	16380	7214	0.888	0.86	5155	0.68	0.689	8102	0.847	0.829
SPAMBASE	57	38	0.943	0.919	**19**	0.85	0.82	24	0.87	0.85

TABLE 6. The result of the interleaved hybridization.

DS		SVM			NB			KNN		
	Original # of features	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement
Enron1	16383	7810	0.932	0.93	7007	0.692	0.719	7408	0.911	0.902
Enron2	11514	5699	0.951	0.963	7010	0.655	0.72	5341	0.865	0.871
Enron3	16382	7244	0.94	0.951	5482	0.674	0.703	7611	0.869	0.802
Enron4	15456	6410	0.874	0.878	6240	0.656	0.678	7423	0.74	0.816
Enron5	14696	6501	**0.963**	0.953	7189	0.643	0.72	6972	0.948	0.95
Enron6	16380	7011	0.895	0.869	5027	0.691	0.70	7482	0.853	0.841
SPAMBASE	57	26	0.949	0.931	**17**	0.856	0.85	20	0.882	0.858

TABLE 7. The result of high-level (WCA-SA).

DS		SVM			NB			KNN		
	Original # of features	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement	# of features	Accuracy	f-measurement
Enron1	16383	8019	0.92	0.92	8002	0.685	0.696	7935	0.899	0.899
Enron2	11514	5722	0.945	**0.952**	8019	0.649	0.71	5622	0.855	0.858
Enron3	16382	7256	0.932	0.951	5028	0.664	0.678	8031	0.861	0.791
Enron4	15456	6786	0.86	0.862	6075	0.599	0.672	7482	0.73	0.801
Enron5	14696	6635	**0.956**	0.95	7135	0.615	0.713	7106	0.931	0.943
Enron6	16380	7811	0.882	0.858	5088	0.674	0.683	8008	0.847	0.828
SPAMBASE	57	41	0.94	0.916	**17**	0.848	0.812	25	0.878	0.854

capability of WC and the fine-tuning power of SA algorithms are explored in this algorithm in every iteration to ensure the achievement of a high-quality feature selection. APPENDIX D referred Pseudo-code of the WC with SA high-level hybridization.

V. EVALUATION PROCESS

The quality of the classifiers was evaluated using three quality measures, namely f-measure, accuracy. Majorly, the external quality measure depends on the labelled test of the email corpora. It involves comparing the resulting classifiers and

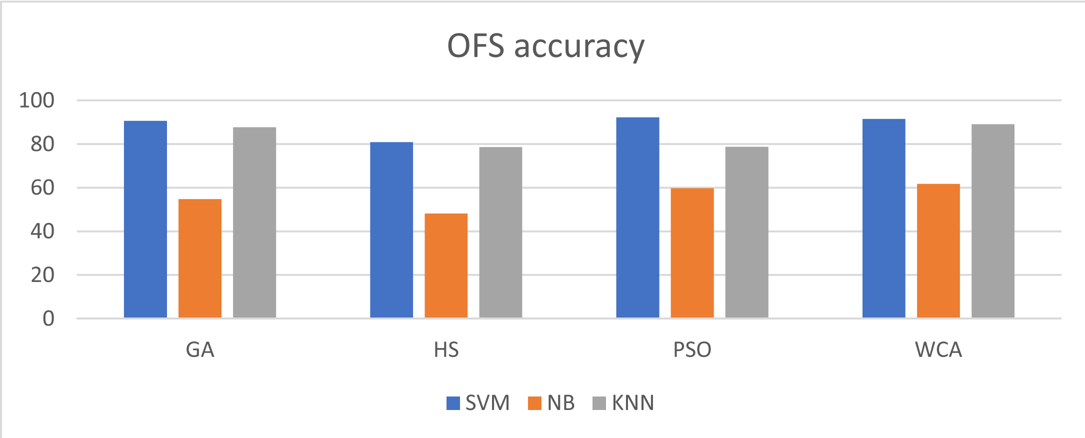

FIGURE 5. The comparison of OFS and WCA accuracies.

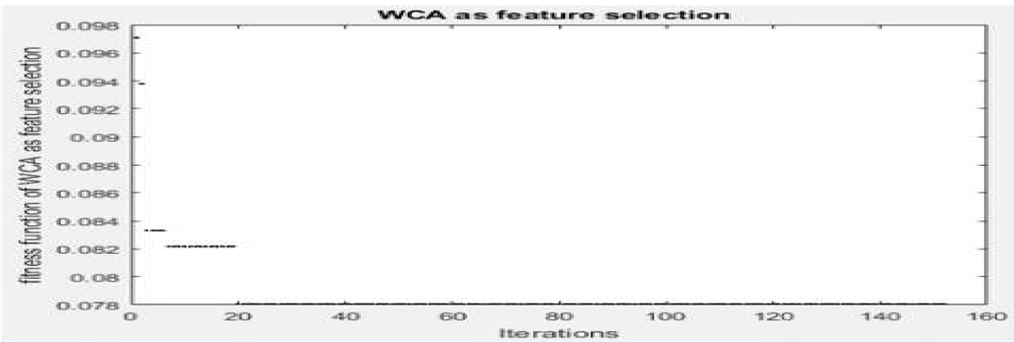

FIGURE 6. JUSTIFICATION FOR USING WCA ITERATION = 50.

the labelled classes, then, measure the extent that emails from the same class are allocated/assigned to the same class.

A. ACCURACY

In classification problems, the evaluation measures are generally defined from a matrix with the numbers of examples correctly and incorrectly classified for each class (also known as the confusion matrix (CM)). Table 3 showed the CM for a binary classification task with only positive and negative classes.

The accuracy rate (ACC) is the most common evaluation measure used in practice; it evaluates the effectiveness of a classifier based on the percentage of its correct predictions. The ACC equation is computed thus:

$$\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}))^{*}100 \quad (5.1.1)$$

B. F-MEASUREMENT

This metric merge both the recall and precision ideas gained from information retrieval. With this measure, each class is taken as the results of emails and perceived as the ideal set of emails or spam. The calculation of the recall and precision for each email j and class i is done thus:

$$\mathrm{Recall(i, j)} = \frac{n_{ij}}{n_i}. \quad (5.2.1)$$

$$\mathrm{Precision(i, j)} = \frac{n_{ij}}{n_j}. \quad (5.2.2)$$

where n_{ij} is the number of available mails having the class label i in class j, n_i is the number of emails with the class label I, and n_j is the number of emails in class j. The calculation of the F-measure of email j and class i is done thus:

$$\mathrm{F(i, j)} = \frac{\mathrm{2Recall(i, j)Precision(i, j)}}{\mathrm{Recall(i, j)} + \mathrm{Precision(i, j)}}. \quad (5.2.3)$$

The calculation of the cumulative F-measure measure is done by considering the weighted average value of the component F-measures as follows:

$$\mathrm{F} = \sum\nolimits_i \frac{n_i}{N} max\ F(i, j). \quad (5.2.4)$$

Therefore, the F-measure values are observed to be in the range of (0,1); larger values = better classifier quality.

C. DATASETS USED IN THE EXPERIMENTS

This section presents a detailed analysis of the datasets that were utilized in various application areas of email classification. As mentioned by [31], some of the most popular datasets in spam email classification are Spam-Base dataset (eight studies), Enron spam email corpus (five studies).Table 4shows all the public datasets used in the application areas in email classification and used in this research.

VI. EXPERIMENTAL RESULTS

A. EVALUATION OF THE WC ALGORITHM AS FEATURE SELECTION

In this section, the experiments were performed using one standard benchmark dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Spambase). The dataset comprised 57 attributes and 4601 emails, where 1813 of the emails were spam, leaving the remaining 2788 as regular emails. The employed dataset is multivariate, containing actual integer attributes. The evaluation of each approach was carried out using WCA. Accordingly, two criteria were referred as follows: classification accuracy and f-Measurement; the number of features in the dataset that will be used in the experiments = 57.

Figure 5 shows a summary of the results of the accuracy and the number of features after testing WCA, as well as the comparative results with other optimization feature selection algorithms like PSO, HS, GA using three popular classifiers (KNN, SVM, NB). The results show that the accuracy of WCA optimization feature selection method was slightly better than the accuracy of other methods using one dataset (SPAME BASE). This motivated the enhancement of the performance of WCA by using a hybridization. of local search and global search in three levels (low-level, high-level, interleaved) and using more datasets called Enron dataset.

B. RESULT OF THE HYBRID WCA-SA MODEL

Table 5 shows a summary of the results of the three classifiers' accuracies and the number of features' results using low-level of WCA-SA. The results show that the best accuracy of 95.9% was achieved with SVM while the minimum number of features (19) was selected with NB.

Table 6shows a summary of the results of the three classifiers' accuracies and the number of features' results using an interleaved hybridization of optimization feature selection methods. The results show that the minimum number of feature (17) was selected with NB while the best accuracy (96.3%) was achieved with SVM.

Table 7 shows a summary of the results of the three classifiers' accuracies and the number of features' results using a high-level of WCA-SA optimization feature selection method. The results show that the minimum number of features (17) was selected using high-level WCA-SA with NB while the best accuracy (95.6%) was achieved with SVM.

Algorithm 1 Pseudo Code Of the WC With SA Low-Level Hybridization

```
• Set user parameter of the WCA: Npop, Nsr, dmax, and Maximum_ Iteration.
• Determine the number of streams (individuals) which flow to the rivers and sea using Eq s. (5) and (6).
• Create randomly initial population of feature selection
• Define the intensity of flow (How many streams flow to their corresponding rivers and sea) using Eq. (7).
while (t < Maximum Iteration) or (any stopping condition)
   for i = 1 : Population Size ( Npop)
Stream flows to its corresponding rivers and sea using Eqs. (9) and (10)
Calculate the fitness function of the generated stream using Eq. (4).
if F_New_Stream < F_river
River = New_ Stream;
if F_New_ Strea m < F_Sea
Sea = New_ Stream;
end if
end if
River flows to the sea using Eq. (11 )
Calculate the fitness function of the generated river using Eq. (4).
if F_New_ River < F_Sea
Sea = New_ River;
end if
   end for
for i = 1 : number of rivers ( Nsr )
if (distance (Sea and River) < dmax) or ( rand < 0.1 )
New streams are created using Eq. (12)
end if
end for
Reduce the dmax using Eq. (13)
      end while
take the best solution "sea" to the SA.
While cool_iteration <= max_iterations
Cool_iteration= cool_iteration +1
Temp_iteration= 0
While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost "sea"
if δ < 0, accept neighbor
else, accept with probability exp(-δ/T)
end while
(0 < α < 1)    T = α*T
end while
```

VII. CONCLUSION

The major scope of this paper is finding all the near-to-optimal features in a given dataset. Regarding the fitness function given in the criteria, the aim is to find all the available feature values in a specific classifier for different classes and categories. Considering all the algorithms, it is clear that SVM has the best performance while KNN was better than NB. This paper proposed three hybridization of WCA with SA algorithm. At the beginning step, WCA was used

Algorithm 2 Pseudo-Code of the WC With SA Interleaved Hybridization

• Set user parameter of the WCA: Npop, Nsr, dmax, and Maximum_ Iterati on.
• Determine the number of streams (individuals) which flow to the rivers and sea using Eq s. (5) and (6).
• Create randomly initial population of feature selection using SA.
While cool_iteration <= max_iterations
Cool_iteration= cool_iteration +1
Temp_iteration=0
While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost "each solution in Npop"
if $\delta < 0$, accept neighbor
else, accept with probability exp(-δ/T)
end while
$(0 < \alpha < 1)$ T = α*T
end while
• The new Npop generated by SA will be used in WCA.
• Define the intensity of flow (How many streams flow to their corresponding rivers and sea) using Eq. (7).
while (t < Maximum_Iteration) or (any stopping condition)
for i = 1 : Population Size (Npop)

While cool_iteration <= max_iterations
Cool_iteration= cool_iteration +1
Temp_iteration=0
While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost " new Stream"
if $\delta < 0$, accept neighbor
else, accept with probability exp(-δ/T)
end while
T = α*T $(0 < \alpha < 1)$
end while
• The new Stream generated by SA will be used in WCA.
Calculate the fitness function of the generated stream using Eq. (4).
if F_New_Stream < F_river
River = New_ Stream;
if F_New_ Strea m < F_Sea
Sea = New_ Stream;
end if
end if
River flows to the sea using Eq. (11)
while (t < Maximum_Iteration) or (any stopping condition)
for i = 1 : Population Size (Npop)

Algorithm 2 *(Continued.)* Pseudo-Code of the WC With SA Interleaved Hybridization

While cool_iteration <= max_iterations
    Cool_iteration= cool_iteration +1
    Temp_iteration=0
    While temp_iteration <= nrep
    temp_iteration =temp_iteration+1
    Select a new point from the neighborhood
  Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost "new River"
if $\delta < 0$, accept neighbor
else, accept with probability exp(-δ/T)
end while
$(0 < \alpha < 1)$ T = α*T
end while

Calculate the fitness function of the generated river using Eq. (4).
if F_New_ River < F_Sea
Sea = New_ River;
end if
end for
for i = 1 : number of rivers (Nsr)
if (distance (Sea and River) < dmax) or (rand < 0.1)
New streams are created using Eq. (12)
end if
end for
Reduce the dmax using Eq. (13)
end while
take the best solution "sea" to the SA.
While cool_iteration <= max_iterations
Cool_iteration= cool_iteration +1
Temp_iteration=0
While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost "sea"
if $\delta < 0$, accept neighbor
else, accept with probability exp(-δ/T)
    end while
$(0 < \alpha < 1)$ T = α*T
end while

as a feature selection technique based on the traditionally employed WC algorithm with the aim of optimizing the fitness function used and their associated optimal features. The impact of Npop and Nsr (WCA parameters) was tested and the empirical studies demonstrated that the parameters must be set to Npop = 28, and Nsr = 7. The first finding was that the proposed WCA performed better than HS and GA, and PSO as the feature selection and the interleaved

Algorithm 3 Pseudo-Code of the WC With SA High Level Hybridization

• Set user parameter of the WCA: Npop, Nsr, dmax, and Maximum_ Iterati on.
• Determine the number of streams (individuals) which flow to the rivers and sea using Eqs. (5) and (6).
• Create randomly initial population of feature selection
• Define the intensity of flow (How many streams flow to their corresponding rivers and sea) using Eq. (7).
while (t < Maximum_Iteration) or (any stopping condition)
    for i = 1 : Population Size (Npop)
Stream flows to its corresponding rivers and sea using Eqs. (9) and (10)
Calculate the fitness function of the generated stream using Eq. (4).
if F_New_Stream < F_river
River = New_ Stream;
if F_New_ Strea m < F_Sea
Sea = New_ Stream;
end if
end if
River flows to the sea using Eq. (11)
Calculate the fitness function of the generated river using Eq. (4).
if F_New_ River < F_Sea
Sea = New_ River;
end if
    end for
for i = 1 : number of rivers (Nsr)
if (distance (Sea and River) < dmax) or (rand < 0.1)
New streams are created using Eq. (12)
end if
end for
Reduce the dmax using Eq. (13)
        end while
take the best solution "sea" to the SA.
While cool_iteration <= max_iterations
Cool_iteration= cool_iteration +1
Temp_iteration=0
While temp_iteration <= nrep
temp_iteration =temp_iteration+1
Select a new point from the neighborhood
Calculate the fitness function of the generated current_cost (of this new point) using Eq. (4).
δ = current_cost- previous_cost "sea"
if $\delta < 0$, accept neighbor
else, accept with probability exp(-δ/T)
end while
$(0 < \alpha < 1)$ $T = \alpha^*T$
end while
    Post-process results and visualization

hybridization of WCA with SA showed the best performance. Having considered these observations, it becomes necessary to maximize the best advantage of WCA by extending WC algorithm with SA algorithm. The same dataset as previously described was also used to design the new algorithm with this hybridization.The hybridization. of the two algorithms resulted in better exploitation of the advantages of SA and WCA, especially on finding the global optima features. The performance of the hybridization was best with interleaved WCA-SA which gave 96.3% then low-level hybridization which resulted in 95.9% and lastly high-level hybridization which resulted in 95.6%.

Based on this finding, it can be concluded that the content classification performance will be improved with enhancements to WCA as a feature selection. The second finding is that the use of the interleaved hybridization generated better optimal features for the SVM classifier than using all the features From this observation, it can be stated that content classification can be better performed using all the optimal features generated by the interleaved hybridization of WCA with SA. Future work could further be used for further researches in several fields of a kind such Used deep learning and To used the Bag-of-Narratives instead of Bag-of-Words and the semantic of each words.

APPENDIX A

See Figure 6.

APPENDIX B

See Algorithm 1.

APPENDIX C

See Algorithm 2.

APPENDIX D

See Algorithm 3.

REFERENCES

[1] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 27, no. 1, pp. 46–57, Jan. 2015.
[2] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, Sep. 2004, pp. 39–50.
[3] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in *Proc. 11th Int. Conf. Inf. Knowl. Manage.*, Nov. 2002, pp. 659–661.
[4] F. Wang, T. Xu, T. Tang, M. Zhou, and H. Wang, "Bilevel feature extraction-based text mining for fault diagnosis of railway systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 49–58, Jan. 2017.
[5] S. Gupta, A. Khattar, A. Gogia, P. Kumaraguru, and T. Chakraborty, "Collective classification of spam campaigners on Twitter: A hierarchical meta-path based approach," in *Proc. World Wide Web Conf.*, Apr. 2018, pp. 529–538.
[6] S. K. Trivedi and P. K. Panigrahi, "Spam classification: A comparative analysis of different boosted decision tree approaches," *J. Syst. Inf. Technol.*, vol. 20, no. 3, pp. 105–298, Aug. 2018.
[7] G. Jain, M. Sharma, and B. Agarwal, "Spam detection on social media using semantic convolutional neural network," *Int. J. Knowl. Discovery Bioinf.*, vol. 8, no. 1, pp. 12–26, Jan. 2018.
[8] S. Saha, S. DasGupta, and S. K. Das, "Spam mail detection using data mining: A comparative analysis," in *Smart Intelligent Computing and Applications*. Singapore: Springer, 2019, pp. 571–580.
[9] M. El Kourdi, A. Bensaid, and T.-E. Rachidi, "Automatic Arabic document categorization based on the Naïve Bayes algorithm," in *Proc. Workshop Comput. Approaches Arabic Script-Based Lang.*, Aug. 2004, pp. 51–58.

[10] S. Eyheramendy, D. Lewis, and D. Madigan, "On the naive Bayes model for text categorization," in *Proc. Artif. Intell. Statist.*, 2003.
[11] S. Al-Harbi, A. Almuhareb, A. Al-Thubaity, M. S. Khorsheed, and A. Al-Rajeh, "Automatic Arabic text classification," in *Proc. 9th Int. Conf. Stat. Anal. Textual Data*, 2008.
[12] B. Al-Salemi and M. J. A. Aziz, "Statistical Bayesian learning for automatic Arabic text categorization," *J. Comput. Sci.*, vol. 7, no. 1, p. 39, 2011.
[13] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289–1305, Mar. 2003. doi: 10.1162/153244303322753670.
[14] D. Fragoudis, D. Meretakis, and S. Likothanassis, "Best terms: An efficient feature-selection algorithm for text categorization," *Knowl. Inf. Syst.*, vol. 8, no. 1, pp. 16–33, Jul. 2005. doi: 10.1007/s10115-004-0177-2.
[15] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Mach. Learn.*, vol. 3, no. 2, pp. 9–95, Oct. 1988.
[16] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, "An ant colony classifier system: Application to some process engineering problems," *Comput. Chem. Eng.*, vol. 28, no. 9, pp. 1577–1584, Aug. 2004.
[17] S.-M. Chen and P.-Y. Kao, "TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines," *Inf. Sci.*, vol. 247, pp. 62–71, Oct. 2013.
[18] H. Eskandar, A. Sadollah, A. Bahreininejad, and M. Hamdi, "Water cycle algorithm—A novel metaheuristic optimization method for solving constrained engineering optimization problems," *Comput. Struct.*, vols. 110–111, pp. 66–151, Nov. 2012.
[19] M. Mafarja and S. Abdullah, "Fuzzy modified great deluge algorithm for attribute reduction," in *Recent Advances in Soft Computing and Data Mining*. Cham, Switzerland: Springer, 2014, pp. 195–203.
[20] M. Mafarja and S. Abdullah, "A fuzzy record-to-record travel algorithm for solving rough set attribute reduction," *Int. J. Syst. Sci.*, vol. 46, no. 3, pp. 12–503, Feb. 2015.
[21] E. G. Talbi, *Metaheuristics: From Design to Implementation*. Hoboken, NJ, USA: Wiley, 2009.
[22] F. Y. Vincent, S.-W. Lin, W. Lee, and C.-J. Ting, "A simulated annealing heuristic for the capacitated location routing problem," *Comput. Ind. Eng.*, vol. 58, no. 2, pp. 288–299, Mar. 2010.
[23] R. Jensen and Q. Shen, "Fuzzy–rough attribute reduction with application to Web categorization," *Fuzzy Sets Syst.*, vol. 141, no. 3, pp. 469–485, Feb. 2004.
[24] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, Oct. 2017.
[25] J. M. Belda-Lois, S. M.-D. Horno, I. Bermejo-Bosch, J. C. Moreno, J. L. Pons, D. Farina, M. Iosa, M. Molinari, F. Tamburella, A. Ramos, A. Caria, T. Solis-Escalante, C. Brunner, and M. Rea, "Rehabilitation of gait after stroke: A review towards a top-down approach," *J. Neuro-Eng. Rehabil.*, vol. 8, no. 1, p. 66, Dec. 2011.
[26] D. Karaboga and B. Akay, "A comparative study of artificial bee colony algorithm," *Appl. Math. Comput.*, vol. 214, no. 1, pp. 32–108, Aug. 2009.
[27] A. H. Gandomi and A. H. Alavi, "Krill herd: A new bio-inspired optimization algorithm," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 17, no. 12, pp. 4831–4835, Dec. 2012.
[28] A. Sadollah, H. Eskandar, A. Bahreininejad, and J. H. Kim, "Water cycle algorithm with evaporation rate for solving constrained and unconstrained optimization problems," *Appl. Soft Comput.*, vol. 30, pp. 58–71, May 2015.
[29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Cham, Switzerland: Springer, 2001.
[30] D. Karaboga and B. Akay, "Artificial bee colony (ABC), harmony search and bees algorithms on numerical optimization," in *Proc. Innov. Prod. Mach. Syst. Virtual Conf.*, Mar. 2009, pp. 1–6.
[31] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email classification research trends: Review and open issues," *IEEE Access*, vol. 5, pp. 9044–9064, 2017.
[32] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, Jan. 2016.
[33] G. Sanchita and D. Anindita, "Evolutionary algorithm-based techniques to handle big data," in *Techniques and Environments for Big Data Analysis*. Cham, Switzerland: Springer, 2016, pp. 113–158.
[34] A. Onan and S. Koruko lu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *J. Inf. Sci.*, vol. 43, no. 1, pp. 25–38, Feb. 2017.
[35] N. K. Nagwani and A. Sharaff, "SMS spam filtering and thread identification using bi-level text classification and clustering techniques," *J. Inf. Sci.*, vol. 43, no. 1, pp. 75–87, Feb. 2017.