sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **24**, No 2
Sofia • 2024 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2024-0021

Feature Selection Using Hybrid Metaheuristic Algorithm for Email Spam Detection

Ghada Hammad Al-Rawashdeh[*,1], *Osama A Khashan*[*,2], *Dr. Jawad Al-Rawashde*[3], *Jassim Ahmad Al-Gasawneh*[4], *Abdullah Alsokkar*[4], *Mohammad Alshinwa*[5, 6]

[1]*Amman Arab University* (*AAU*), *Amman, Jordan*
[2]*Research and Innovation Centers, Rabdan Academy, P.O. Box 114646, Abu Dhabi, United Arab Emirates*
[3]*Saudi Electronic University* (*SEU*), *Riyadh, Saudi Arabia*
[4]*Faculty of Business, Applied Science Private University, Amman 11931, Jordan*
[5]*Faculty of Information Technology, Applied Science Private University, Amman 11931, Jordan*
[6]*MEU Research Unit, Middle East University, Amman 11831, Jordan*
E-mails: g.alrawashdeh@aau.edu.jo okhashan@ra.ac.ae j.alrawashdeh@seu.edu.sa j_algasawneh@asu.edu.jo a_sokkar@asu.edu.jo m_shinwan@asu.edu.jo
** Corresponding authors: Ghada Hammad Al-Rawashdeh and Osama A Khashan*

Abstract: *In the present study, Krill Herd* (*KH*) *is proposed as a Feature Selection tool to detect spam email problems. This works by assessing the accuracy and performance of classifiers and minimizing the number of features. Krill Herd is a relatively new technique based on the herding behavior of small crustaceans called krill. This technique has been combined with a local search algorithm called Tabu Search* (*TS*) *and has been successfully employed to identify spam emails. This method has also generated much better results than other hybrid algorithm optimization systems such as the hybrid Water Cycle Algorithm with Simulated Annealing* (*WCASA*). *To assess the effectiveness of KH algorithms, SVM classifiers, and seven benchmark email datasets were used. The findings indicate that KHTS is much more accurate in detecting spam mail* (*97.8%*) *than WCASA.*

Keywords: *Krill Herd Algorithm, Tabu search, Metaheuristic Algorithm, Feature selection, Hybridization.*

1. Introduction

Accurately detecting spam emails has become a significant problem in everyday life due to the high volume of emails sent. For this reason, many researchers including [15, 36] have attempted to identify the main challenges associated with standard spam email classifiers. Moreover, researchers continue to investigate the performances of different text classifiers in detecting spam emails [8, 18, 20, 29, 34, 35].

The most significant challenge surrounding in-text classifiers is the vast number of features involved. Sometimes, it may be possible to reduce the number of features by extracting the key semantics from texts. However, this is often impossible because deleted items can contain important features [14, 17, 38, 41].

Thus, when it comes to spam classifiers, the vast number of FS is a significant problem. Several researchers have attempted to address this issue in the past using a variety of methods including chi-squares to reduce the number of features. However, such researchers failed to mention the effects that these reductions have had on the performance and efficiency of classifiers [8, 22, 23, 43].

Metaheuristic optimization is a process in which metaheuristics are employed to select the optimal solutions from a variety of potential solutions (in the present work, this is the classifiers' performance). Optimization plays an essential role in verifying important features of spam classifiers [1, 21]. Additionally, there is a great deal of ambiguity surrounding the impacts that FS optimizations have on different Support Vector Machines (SVM), which is another significant problem.

Given the points discussed above, several questions are put forward in the present study. Firstly, which level of optimization is needed to overcome the weakness of the vast number of features? Secondly, is the combined process of using KHFS with TSFS more effective than using WCSAFS? And finally, how is the performance of SVM as a spam detection tool impacted by a reduction in the number of features?

The present study will attempt to address all of the questions presented above, with a particular focus on using hybrid Krill Herb Optimizations (KHO) with Tabu Search for feature selection to address the issue of vast numbers of features. Moreover, the SVM and its hybrid use with KHO for feature selection will be assessed. The study will serve as an important contribution to research in the field of detecting and categorizing emails and will help enhance the efficiency and effectiveness of spam classifiers.

The remainder of this paper will thus be structured as follows: in Section 2, a literature review citing several previous studies on the topic will be presented. Subsequently, in Section 3 proposed KH and TS will be discussed, after which the experimental and analytical findings will be presented. Finally, conclusions will be made in the final section.

2. Relevant literature

No unified or standard definition of spam has been developed. In a majority of cases, it is defined as unwanted emails. However, it is important to note that not every undesired e-mail is spam. On the other hand, [4], and [8] suggest that spam could be defined as unsolicited commercial e-mails. However, spam is not just advertising material.

A text classifier is a tool used to identify spam emails. [24] Created a spam classifier using SVM combined with BOW to extract data. The findings revealed that combining SVM with a Gaussian Kernel was more effective than using SVM with a Polynomial Kernel or Linear Kernel. Similarly, [11] employed BOW to compare

SVM with AdaBoost and RF and revealed that SVM was more effective as a classifier.

Although there has been a continuous, significant increase in the amount of machine-readable information produced, the capabilities of interpreting and understanding this information have not been able to keep up. [10] Explain that machine learning tools enable large quantities of text to be automatically organized and feature selection plays a vital role in machine learning [5, 6] Feature selection can identify the most important features in learning by implementing a learning algorithm to predict the features that are most beneficial for analysis [18, 20, 29, 33-35, 40, 42].

Moreover, common machine-learning algorithms developed upon test theories have frequently been used for correlation-based feature selection. These tools can be used to address various natural and artificial problems. Feature selection is a quick and easy process to carry out. It removes irrelevant information and ultimately enhances the efficiency of learning algorithms. Additionally, this method has been found to produce results similar to those of advanced feature selectors whilst also requiring less computation. The present work will focus on identifying methods that can be used to reduce the vast number of features whilst simultaneously enhancing the efficiency of spam detection tools [10, 36].

As it can enhance the accuracy of ML and training efficiency feature selection is a vital process. Even though several FS filter approaches have already been developed, efforts are still being made to create new, innovative approaches to address FS issues. FS algorithms are both simple and time-efficient, which makes them ideal for addressing issues with feature selection [19]. Assert that several important features must be included to account for the shortcomings of available measures to create effective FS solutions.

However, it is important to note that different FS approaches have different impacts on learning frameworks. Many researchers are thus attempting to create new techniques to address FS and cruse dimensionality issues, with spam classifiers being a particular topic of focus. There are several reasons for which FS approaches may be used, including to simplify data, enhance performance and visualization, and reduce the dimensionality of features [22, 23, 39, 56]. To address the latter issue of dimensionality reduction, several different approaches have been developed including IG, Chi-Square Statistic, MI, TS, and DFT. [3, 21] Explained that feature selection algorithms are problem-dependent, with feature dependencies being largely ignored. The present research will thus investigate other FS approaches (i.e., meta-heuristic algorithms).

To search for a specific area for candidate solutions, a local search technique called an algorithmic method' can be used. In this process, a candidate solution is sought from the direct neighborhood (which is determined using local information). This process continues until a termination condition is fulfilled. This is a single-based method and commences with an initial candidate solution. If a solution cannot be identified in the direct neighborhood, then the process moves to the neighbors to identify a solution to the present issue and the expansion continues until the search condition is met. What's more, a solution from the neighborhood is only accepted if

it outperforms the other possible solutions in the candidate sets. In prior studies, researchers have used single-based approaches including SA, Hill climbing, and Tabu Searching [7, 9, 22].

Several studies have employed hybrid single-based methods to enhance their performance. Generally speaking, a local search is more attractive in terms of exploitation than a population search. Moreover, several existing studies have used hybrid methods combining both local and global searchers to identify "data classifiers" for data mining. An example of this is a study [9], in which a hybrid optimization algorithm was used. This algorithm simulated the annealing of spam classifiers. Additionally, the hybrid water algorithm is considered to be a global search, which is as powerful as a local search. It was found that the interleaved hybridization was much more effective in feature selection algorithms and demonstrated an accuracy of 96.3%. The SVM was found to be the most effective classifier algorithm, showing an f-measurement of 96.3%. When employing interleaved Water Cycle and Simulated Annealing in a hybrid approach, the number of features was successfully reduced by more than 50%.

On the other hand, [1] employed a new text-clustering approach. They developed an improved krill herd algorithm, which they collaborated with a hybrid function called MMKHA. The researchers put forward this approach as an effective clustering technique that achieved promising results. The findings indeed revealed that the new krill herd algorithm with hybrid functions successfully produced better results than other algorithms for all datasets.

The following observations were noted from the literature review: (i) even though several studies have explored how the efficiency and effectiveness of spam classifiers can be improved, there have been very few investigations into how FS can be optimized to enhance spam classification; (ii) it remains challenging to identify the key features involved in spam classification because existing FS techniques do not take into account the relationship between the features. The present research thus intends to investigate optimization algorithms; (iii) efforts are still being made by academics in this field to assess different FS methods, FS optimization, and spam classifiers.

As previously stated, there are several issues relating to spam classification. Moreover, the performance of the employed classifiers is impacted by the features used during the spam classification process. As far as feature-selection issues are concerned, it is crucial to identify the most significant features so that other processes are not affected [27]. The present work attempts to address these issues by developing a new feature selection method in which KHS is used to classify spam. The key FS problems are the dependency of features and the relationship between them. In prior research, several methods (including Chi-squares, information-gain FS, and term frequency probabilities) have been carried out to address these issues. However, these methods do not take into account FS comparisons. It is thus essential to use optimization algorithms (such as FS) to select the optimal features, and to subsequently assess their performance in spam classification processes.

3. Proposed method

In this research, an experimental methodology was employed to create a spam classification system with enhanced quality and performance. The researcher also wanted to implement KH to reduce the number of features during the feature selection process [49-51]. In Table 1, the seven spam email datasets used in the study are presented. The datasets were separated so that they could be cross-validated. K-fold cross-validation is a process involving the use of K-5 folds for training and validation purposes. On the other hand, the leftover folds may be used for testing. In the process, a maximum of 100 will be performed (MI=1000). Additionally, the total Number of Krill (NK) performed will be 30 and the total Number of Runs (NR) will be 1. The research model proposed for spam email detection is presented in the Fig. 1.

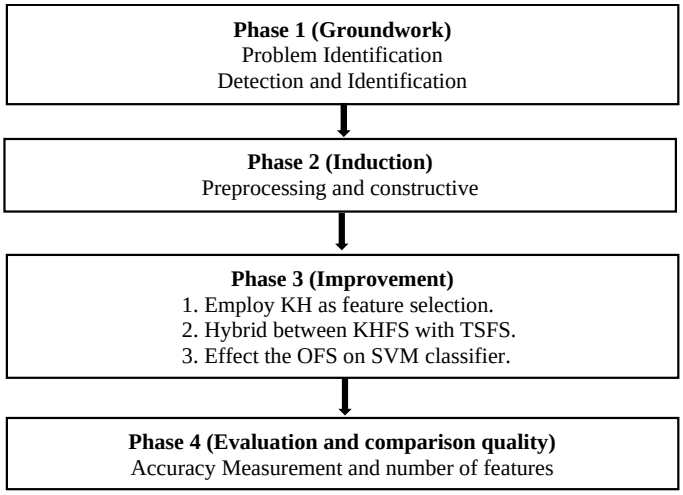

Fig. 1. Research model for spam email classification

3.1. Groundwork and induction phase

The groundwork performed in this research to identify spam classification problems is discussed in this section. Also, existing studies that have investigated the identified issues will be cited. Through the presentation of the literature review, the researcher aims to identify the shortcomings of existing algorithms and approaches commonly used in spam classification [52-55]. Designing the initial spam classification process for the model put forward in this work is the primary activity that will take place during the induction stage. During this stage, the improvement phase will be iterated. Additionally, a suitable programming language will be used to develop a KH feature-selection process. Pre-processing (i.e., the filtering out of irrelevant data from emails, word lists, term representations, and stop word lists) is the key activity in this phase. To achieve this, a TFIDF weighting approach will be employed. Subsequently, alternative spam classifiers will be constructed, which will ultimately determine how the initial solution to enhancing spam classifiers will be constructed.

3.2. Improvement phase

In this phase, the primary objective is to enhance the performance of spam classification in two key ways. Firstly, the most important terms must be extracted and non-relevant terms must be eliminated. Secondly, the number of features chosen in the preceding stage must be reduced to enhance the efficiency of the process. To achieve this, the similarity between accuracy values should be maximized.

3.2.1. The proposed krill herd feature selection method

In this research, we put forward a pure Krill-herd search-based algorithm for use in feature selection algorithms. Krill Herd (KH) is a new, innovative type of swarm-based metaheuristic optimization algorithm that is based on the herding behaviors of krill in the ocean. In the KH optimizations process, the objective function is calculated as the least distance between the food location and a krill's position. This technique is more effective than many advanced metaheuristic algorithms in many areas. In this section, standard KH is employed. The values of the solutions are continuously assessed and updated in the search space.

The algorithm proposed in this research is developed using a vector-space model. Thus, every term is representative of multi-dimensional term spaces, and each email ($d_i = (w_{i1}, w_{i2}, ..., w_{in})$) is considered to be a vector in the term space. Moreover, in the algorithm, there are different numbers (n) of terms. Additionally, every potential email detection solution is a vector of features. Therefore, the feature selection problem is rendered an optimization task that requires the identification and use of optimal features as opposed to all features. In this way, feature selection quality was chosen to be the objective function, and krill herd feature selection was employed as a means of optimizing the objective.

In the proposed algorithm, there are several representations in which F (Features) is codified as a set in a vector of length m; M represents the number of features, as presented in Fig. 2. Each element involved in this vector is labeled according to whether the features are selected or eliminated. Fig. 2 shows an example of a solution representation, in which 8 features {1, 2, 5, and 8} have been selected. All other features {3, 4, 6, 7} were eliminated.

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	1	0	0	1	0	0	1

Fig. 2. Representation of the chosen features

There are three key movements involved in the updating of each krill's position in KH, namely: movement affected by others; foraging action; and random physical diffusion.

The position of a krill can be idealized in the following way:

$$dx_i / dt = N_i + F_i + D_i. \quad (1)$$

In this expression, N_i, F_i, and D_i represent the three previously discussed movements.

The target, local, and repulsive effects are used to approximately calculate the induced direction (α_i,) of the first motion. Thus, Krill's first motion can be idealized as follows:

(2) $$N_i^{new} = N^{max}\,\alpha_i + \omega_n\,N_i^{old}.$$

Here, N^{max} represents the maximum speed, whilst ω_n represents the weight of the first motion, and N^{old} is the final motion.

Two factors impact the second action: the food position and the information relevant to it.

This motion for krill i can be expressed in the following way:

(3) $$F_i = V_f\,\beta_i + \omega_{fi}^{fold},$$

(4) $$\beta_i = \beta^{food}{}_i + \beta^{best}{}_i,$$

is the final foraging motion. In this equation, V_f represents the foraging speed, ω_{fold} represents the weight and fold random searching, physical diffusion is crucial. It allows a global search (exploration) to be conducted as part of the overall search process. This action can be expressed in the form presented below:

(5) $$D_i = D^{max}\,\delta.$$

In this equation, D^{max} represents the maximum diffusion speed, whilst δ is the random number.

A Krill's time-related position can be expressed using the following equation, which is based on the three actions discussed above:

(6) $$X_i(t + \Delta t) = X_i(t) + \Delta t\,dx_i/dt.$$

KRILL Herd Algorithm

Step 1. Initialization. (Initialize the generation counter ***G***, the population ***P*** of **NP** krill randomly, $\boldsymbol{V_f}$, $\boldsymbol{D^{max}}$ and $\boldsymbol{N^{max}}$).

Step 2. Fitness Evaluation. Calculate fitness for each krill according to its initial position.

Step 3. while G < MaxGeneration **do**

Sort the population according to their fitness.

For i=1:NP(all krill) **do**

Perform the following motion calculation .

Motion induced by other individuals

Foraging motion

Physical diffusion

Implement the genetic operators

Update the krill position in the search space.

Calculate fitness for each krill according to its new position.

End for i

$G = G + 1$.

Step 4. end while

End.

Fig. 3. Krill herd pseudo-code

The TS is a Metaheuristic Algorithm that enables flexible movements to be made, which deviates away from a local. In the TS, a new search movement is selected and the assessment of previous solutions is subsequently prohibited. The following elements are involved in basic TS:

Tabu List. This tool gives the algorithm a short-to-medium-sized memory. In other words, movements from prior searches are memorized and disabled. These moves are otherwise known as Tabu Moves.

Tenancy Period. This term refers to the length of time (i.e., number of iterations) for which Tabu Moves are banned by the Tabu List. A comprehensive outline of the TS Algorithm procedure is presented by [31] gives a comprehensive and general procedure for the TS Algorithm.

Step 0. Choose an initial solution $s_0 \in S$. Initialize the Tabu List $L_0 = \emptyset$ and chose a list tabu size. Establish $k = 0$.

Step 1. Assess the viability of the neighbourhood ($N(S_k)$) and eliminate inferior items on the tabu list (L_k).

Step 2. Choose the subsequent movement ($S_k + 1$). This will be either $N(S_k)$ or L_k, depending on which option is best. After this, L_k+1 must be updated.

Step 3. If a termination criterion is fulfilled, then the process can be stopped her. If not, k=k+1 must be performed, after which one must return to Step 1.

Fig. 4. Tabu search pseudo-code

3.2.3. The proposed approach (hybridization krill herd and tabu search) and improved SVM

Feature selection is a tool that is frequently used to solve binary optimization issues. It limits solutions to specific binary numbers (i.e., 0, 1). When using a KH Algorithm, a binary value for the version must be created first. Subsequently, one must consider the solution needed in this situation, which should be employed as a single-dimensional vector to calculate the total length of the vector. This calculation is made based on how many attributes are present in the original dataset. Each value in the vector is represented using the values 1 or 0, with the former indicating that the corresponding attribute has been chosen. If this is not the case, the value will be presented as 0.

When creating and defining a metaheuristic tool, there are two somewhat contradictory criteria to take into account, namely the diversification and exploitation of search space. These factors are crucial in determining the best possible solutions [32]. Metaheuristic algorithms can be broadly categorized into two key types, depending on the criteria used. The first type is population-based (which includes techniques such as swarm intelligence and evolutionary algorithms) algorithms. Such algorithms are exploration-based. The second type is single solution-based algorithms (such as local searches and Tabu Searches). These algorithms are exploitation-based. It is crucial to establish an effective balance between the two to ensure that an algorithm will perform well in searching procedures.

It is thus widely accepted that TS can enhance the use of KH algorithms. In other studies, different formats have been employed to test different feature-selection algorithms. However, as far as we are aware, our research is the first of its kind to combine KH and TS algorithms in hybrid form for FS spam email detection. Both algorithms have powerful properties, and combining them can achieve significantly

improved results. This hybridization improves the performance and efficiency of the KH Algorithm.

In this section of the paper, the efficiency of the hybrid algorithm in the newly proposed feature selection process will be examined. The key objective here is to determine whether the combination of the two algorithms enhances the results of the KH Algorithm. To ensure that the optimal results were produced using the algorithm in the newly proposed model, a TS approach developed based on pipeline methodology is used when KH finishes. This also makes the algorithm more robust. Moreover, the application of tournament selection creates a stronger approach to identifying the algorithm's diversity. As the methodology gives all individuals a chance to be selected, his methodology can be used to identify all search agents present in a specific population dataset.

The new KH Algorithm enables more efficient optimization results to be achieved. In the primary algorithm, a blind operation is used. Thus, the process involves the operator, who exploits the algorithm. This occurs regardless of the solution's fitness value. At this stage in the investigation, the operator is replaced with a local search in which a simple solution is considered to be the initial state. The system continues to work on the solution until the original solution is finally replaced with an actual result. The system is incredibly robust due to the hybridization between these two actions (global search (KH) and local search algorithm (TS)).

In the new combined algorithm, the local method formed part of the KH Algorithm. Once each iteration round had been completed, the best N_p vector was selected by the TS to serve as the initial point. Subsequently, the N_p is updated if the locally-optimized vectors are found to have a better fitness value than the N_p. This continues until the stopping point is reached.

In essence, feature selection is a multi-objective optimization tool that can be used in situations where identifying the best solution is difficult. The key objective of applying a selective optimization method is to reduce the number of selected features and to achieve the highest level of classification accuracy. In simpler terms, the smaller the number of selected features, the higher the chances of achieving an effective classification solution.

Furthermore, to assess all of the potential solutions, the proposed fitness-selection model and final publisher are used. This evaluation process requires the use of an SVM classifier to determine the accuracy of the classification solution and to identify the number of selected features in the solution. The fitness function presented below in the next equation is used in the TS and KH algorithms to assess the search agents. This ensures that a balance is achieved between all selected features in the minimum solutions. It also ensures that the feature selection process is as accurate,

$$\text{fitness} = \beta\gamma_R(D) + \alpha \frac{|R|}{|N|}. \tag{7}$$

In this equation, $\gamma_R(D)$ represents the rate of classification error for a specific classifier; whilst $|R|$ represents the chosen subset's cardinality. Additionally, $|N|$ represents the total number of features present in the dataset, and α and β identify the relevance of the classification quality and subset length. The symbols $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$ are taken from the works conducted by [12, 23].

4. Evaluations and comparisons in the quality phase

4.1. Accuracy

The equation below can be used to calculate the relationship between the classifier and the data positive label:

(8) $$ACC = ((TP+TN)/ (TP+TN + FP+FN)) \times 100\%.$$

4.2. f-measurement

The equation below can be used to calculate a classifier's effectiveness in identifying positive labels:

(9) $$F(i, j) = \frac{2\text{Recall }(i,j)\text{Precision }(i,j)}{\text{Recall }(i,j)+\text{Precision }(i,j)}.$$

4.3. Experimental datasets

In this section, the datasets employed to explore different email classifications will be discussed. Several well-known spam email classification datasets were used in the research performed by [26], such as the Spam-Base dataset (eight studies) and, the Enron dataset (6/20 studies in multi-folder email categorization). Moreover, it is important to note that emails can contain a variety of phishing targets. This provides information about the materials involved. In most cases, researchers have employed Phishing Corpus to categorize spam and phishing emails and to combine different methods to create spam detection systems [26].

The final dataset in the present study was taken from the Spam Assassin public mail corpus. Altogether, 9346 records with 90 features were included in this dataset. Each piece of data in the set has been labeled as either "Ham" or "Spam", with 6951 being labeled as ham emails and 2395 as spam emails. In other words, around 25.6% of the emails were classified as "spam". However, this causes an evident imbalance in the data, thus rendering the analysis process more difficult [13] Table 1 below shows all of the available email classification datasets and information about them.

Table 1. Summary of the Spam Datasets

Document set	Source	Number of emails	Number of features
DS1	Spam Base: **http://archive.ics.uci.edu/ml/data sets/Spambase**	Total 4601 emails (spam = 1813 and ham = 2788)	After run code
DS2	Enron Spam Corpus **http://www.aueb.gr/users/ion/dat a/enron-spam/**	Total 30,041 emails (spam = 13,496 and ham = 16,545)	After run code
DS3	Spam Assassin public mail corpus **http://spamassassin.org/ publiccorpus/**.	Total 9346 emails (spam = 2395 and ham = 6951)	After run code

4.4. Assessing the effectiveness of KH algorithms and other optimization techniques in feature selection

The results of the KH algorithm and other optimization algorithms are discussed in this section. It will also be noted whether the algorithm is a local or global search algorithm and whether any hybridization was involved (WCA PSO, HS, SA).

Table 2. Accuracy results for KH and other optimization techniques when SVM is used

Dataset	PSO	HS	WCA	SA	TS	KH
Enron1	0.905	0.855	0.919	0.875	0.773	0.917
Enron2	0.938	0.904	0.944	0.93	0.89	0.952
Enron3	0.936	0.811	0.93	0.929	0.902	0.939
Enron4	0.810	0.796	0.859	0.827	0.783	0.896
Enron5	0.943	0.887	0.951	0.891	0.880	0.965
Enron6	0.872	0.765	0.879	0.852	0.841	0.901
Spam base	0.922	0.808	0.938	0.880	0.855	0.950
Spam assassin	0.932	0.823	0.939	0.882	0.849	0.958

Table 3. F-measurement results when SVM is used

Dataset	PSO	HS	WCA	SA	TS	KH
Enron1	0.9025	0.9040	0.9164	0.874	0.863	0.922
Enron2	0.9383	0.9073	0.9470	0.871	0.859	0.953
Enron3	0.9328	0.8242	0.9486	0.905	0.899	0.951
Enron4	0.8000	0.7903	0.859	0.7886	0.755	0.905
Enron5	0.9449	0.8871	0.948	0.9213	0.912	0.968
Enron6	0.8668	0.8016	0.851	0.8288	0.802	0.905
Spam base	0.9056	0.7441	0.910	0.8486	0.835	0.959
Spam assassin	0.9078	0.7589	0.914	0.8546	0.8325	0.962

Table 4. Number of features results when SVM is used

Dataset	Number before select	PSO	HS	WCA	SA	TS	KH
Enron1	16,383	8313	8293	8189	8253	8500	8177
Enron2	11,514	10,113	3509	5810	3477	4001	4425
Enron3	16,382	14,357	4916	7394	3182	3658	4865
Enron4	15,456	7660	4630	6813	4411	4700	5200
Enron5	14,696	11,060	4380	6738	4229	4651	4521
Enron6	16,380	12,360	8160	7948	7682	7700	8258
Spam base	57	50	28	44	27	32	35
Spam assassin	90	68	57	48	47	50	51

4.5. Comparing the KHTS and WCASA with other methods

The results about the efficiency and effectiveness of the KHTS and WCASA algorithms will be discussed in this section. Moreover, the two algorithms will be compared to determine which one is most accurate and effective.

In Tables 5, 6, and 7, the results of the accuracy, f-measurements, and feature numbers of the SVM classifiers when KH-TS is used are presented. These results are compared with those of the interleaved WCSA optimization feature selection algorithm. The findings reveal that, when the interleaved KH-TS was used, a minimum number of features was selected. Additionally, the KH-TS was also found to be the most accurate (97.8%).

Table 5. Accuracy results for (KH-TS) and (WCA-SA) when SVM is used

Dataset	KH	WCA	Best	Accuracy
Enron1	0.930	0.932	WCSA	0.932
Enron2	0.962	0.951	KHTS	0.962
Enron3	0.958	0.94	KHTS	0.958
Enron4	0.901	0.874	KHTS	0.901
Enron5	0.978	0.963	KHTS	0.978
Enron6	0.913	0.895	KHTS	0.913
Spam base	0.945	0.949	WCSA	0.949
Spam Assassin	0.951	.0.9486	KHTS	0.951

Table 6. F-measurement results for (KH-TS) and (WCA-SA) when SVM is used

Dataset	KH	WCA	Best	f-measurement
Enron1	0.925	0.930	WCSA	0.930
Enron2	0.972	0.963	KHTS	0.972
Enron3	0.968	0.951	KHTS	0.968
Enron4	0.907	0.878	KHTS	0.907
Enron5	0.972	0.953	KHTS	0.962
Enron6	0.929	0.869	KHTS	0.929
Spam base	0.925	0.931	WCSA	0.925
Spam Assassin	0.959	0.942	KHTS	0.959

Table 7. Feature-number results for (KH-TS) and (WCA-SA) when SVM is used

Dataset	KH	WCA	Best	Number of features
Enron1	8010	7810	WCSA	7810
Enron2	4102	5699	KHTS	4102
Enron3	4589	7244	KHTS	4589
Enron4	5200	6410	KHTS	5200
Enron5	4005	6501	KHTS	4005
Enron6	5896	7011	KHTS	5896
Spam base	30	26	WCSA	26
Spam Assassin	42	40	WCSA	40

5. Conclusions

The key objective of the present work was to identify the optimal/near-optimal features of a given dataset. As far as the fitness function given in the criteria is concerned, the study aims to identify all available feature values in a specific classifier for different classes and categories. In this work, a hybrid system involving a local and global metaheuristic algorithm (KH-TS) has been proposed. The first step in this project was to employ KH as a feature selection method and to optimize the fitness function and the relevant features. Secondly, the project attempted to enhance the efficiency and effectiveness of the SVM classifier by implementing an interleaved hybrid KH- TS algorithm. The accuracy of the system was then compared with that of another hybrid optimization system (WC-SA), with the results indicating that the KHTS largely outperformed the WCSA (and other techniques such as HS, GA, and PSO) in terms of feature selection capacity. The KH-TS was found to give a 97.8% accuracy performance. In future studies, researchers should consider exploring the topic in different fields, such as deep learning. Moreover, they may wish to employ

a Bag-of-Narratives rather than a Bag-of-Words and to focus on the meanings of each word.

Acknowledgment: The authors gratefully acknowledge the Rabdan Academy for their invaluable support and resources provided throughout this research, which significantly contributed to its successful completion.

References

1. Abualigah, L. M., A. T. Khader, M. A. Al-Betar. Unsupervised Feature Selection Technique Based on Genetic Algorithm for Improving the Text Clustering. – In: Proc. of 7th International IEEE Conference on Computer Science and Information Technology (CSIT'16), 2016.
2. Abualigah, L. M., A. T. Khader, E. S. Hanandeh. Hybrid Clustering Analysis Using Improved Krill Herd Algorithm. – Applied Intelligence, Vol. **48**, 2018, No 11, pp. 4047-4071.
3. Al-Gasawneh, J. A., K. N. AlZubi, M. M. Anuar, S. F. Padlee, A. ul-Haque, J. Saputra. Marketing Performance Sustainability in the Jordanian Hospitality Industry: The Roles of Customer Relationship Management and Service Quality. – Sustainability, Vol. **14**, 2022, No 2, 803.
4. Alghoul, A., S. Al Ajrami, G. Al Jarousha, G. Harb, S. S. Abu-Naser. Email Classification Using Artificial Neural Network. – International Journal of Academic Engineering Research, Vol. **12**, 2018, No 6, pp. 25-33.
5. Aljanabi, M., H. M. Qutqut, M. Hijjawi. Machine Learning Classification Techniques for Heart Disease Prediction: A Review. – International Journal of Engineering & Technology, Vol. **7**, 2018, No 4, pp. 5373-5379.
6. Alkhalili, M., M. H. Qutqut, F. Almasalha. Investigation of Applying Machine Learning for Watch-List Filtering in Anti-Money Laundering. – IEEE Access, Vol. **9**, 2021, pp. 18481-18496.
7. Alnaser, A. S., M. S. Al-Shibly, M. Alghizzawi, M. Habes, J. A. Al-Gasawneh. Impacts of Social Media and Demographical Characteristics on University Admissions: The Case of Jordanian Private Universities. – PalArch's Journal of Archaeology of Egypt/Egyptology, Vol. **17**, 2020, No 7, pp. 6433-6454.
8. AL-Rawashdeh, G. H., R. B. Mamat. Comparison of four email classification algorithms Using WEKA. – International Journal of Computer Science and Information Security (IJCSIS), Vol. **17**, 2019, No 2, pp. 42-54.
9. Al-Rawashdeh, G., R. Mamat, N. H. B. A. Rahim. Hybrid Water Cycle Optimization Algorithm with Simulated Annealing for Spam E-mail Detection. – IEEE Access, Vol. **7**, 2019, pp. 143721-143734.
10. Dada, E. G., J. S. Bassi, H. Chiroma, A. O. Adetunmbi, O. E. Ajibuwa. Machine Learning for Email Spam Filtering: Review, Approaches and Open Research Problems. – Heliyon, Vol. **5**, 2019, No 6, pp. 20-30.
11. Diale, M., C. Van Der Walt, T. Celik, A. Modupe. Feature Selection and Support Vector Machine Hyper-Parameter Optimization for Spam Detection. – In: Proc. of Pattern Recognition Association of South Africa and the Robotics and Mechatronics International Conference (PRASA-RobMech'16), IEEE, 2016.
12. Emary, E., H. M. Zawbaa, A. E. Hassanien. Binary Grey Wolf Optimization Approaches for Feature Selection. – Neurocomputing, Vol. **172**, 2016, pp. 371-381.
13. Faris, H., I. Aljarah, J. F. Alqatawna. Optimizing Feedforward Neural Networks Using Krill Herd Algorithm for e-Mail Spam Detection. – In: Proc. of Jordan IEEE Conference on Applied Electrical Engineering and Computing Technologies (AEECT'15), IEEE, November 2015, pp. 1-5.
14. Fodeh, S., B. Punch, P. N. Tan. On Ontology-Driven Document Clustering Using Core Semantic Features. – Knowledge and Information Systems, Vol. **28**, 2011, No 2, pp. 395-421.

15. F o r s a t i, R., M. M a h d a v i, M. S h a m s f a r d, M. R. M e y b o d i. Efficient Stochastic Algorithms for Document Clustering. – Information Sciences, Vol. **220**, 2013, pp. 269-291.
16. G a n d o m i, A. H., A. H. A l a v i. Krill Herd: A New Bio-Inspired Optimization Algorithm. – Communications in Nonlinear Science and Numerical Simulation, Vol. **17**, 2012, No 12, pp. 4831-4845.
17. G h a d a, A. R., R. B. M a m a t, J. H. R a w a s h d e h. Evaluation of the Performance for Popular Three Classifiers on Spam Email without Using FS Methods. – WSEAS Transactions on Systems and Control, Vol. **16**, 2021, pp. 121-132.
18. G u p t a, H., M. S. J a m a l, S. M a d i s e t t y, M. S. D e s a r k a r. A Framework for Real-Time Spam Detection in Twitter. – In: Proc. of 10th International Conference on Communication Systems & Networks (COMSNETS'18), IEEE, 2018.
19. H u a n g, Y., C. Z h a o, H. Y a n g, X. S o n g, J. C h e n, Z. L i. Feature Selection Solution with High Dimensionality and Low-Sample Size for Land Cover Classification in Object-Based Image Analysis. – Remote Sensing, Vol. **9**, 2017, No 9, 939.
20. J a i n, G., M. S h a r m a, B. A g a r w a l. Spam Detection on Social Media Using Semantic Convolutional Neural Network. – International Journal of Knowledge Discovery in Bioinformatics (IJKDB), Vol. **8**, 2018, No 1, pp. 12-26.
21. L i u, Y., Y. W a n g, L. F e n g, X. Z h u. Term Frequency Combined Hybrid Feature Selection Method for Spam Filtering. – Pattern Analysis and Applications, Vol. **19**, 2016, No 2, pp. 369-383.
22. M a f a r j a, M. M., S. M i r j a l i l i. Hybrid Whale Optimization Algorithm with Simulated Annealing for Feature Selection. – Neurocomputing, Vol. **260**, 2017, No 5, pp. 302-312.
23. M a f a r j a, M., S. A b d u l l a h. A Fuzzy Record-to-Record Travel Algorithm for Solving Rough Set Attribute Reduction. – International Journal of Systems Science, Vol. **46**, 2015, No 3, pp. 503-512.
24. M a l d o n a d o, S., G. L' H u i l l i e r. SVM-Based Feature Selection and Classification for Email Filtering. – In: Pattern Recognition-Applications and Methods, Berlin, Heidelberg, Springer, 2013, pp. 135-148.
25. M c c o r d, M., M. C h u a h. Spam Detection on Twitter Using Traditional Classifiers. – In: Proc. of International Conference on Autonomic and Trusted Computing, Berlin, Heidelberg, Springer, 2011.
26. M u j t a b a, G., L. S h u i b, R. G. R a j, N. M a j e e d, M. A. A l-G a r a d i. E-mail Classification Research Trends: Review and Open Issues. – IEEE Access, Vol. **5**, 2017, No 5, pp. 9044-9064.
27. R a m a d a n, Q. H., M. M o h d. A Review of Retrospective News Event Detection. – In: Proc. of International Conference on Semantic Technology and Information Retrieval, IEEE, Vol. **95**, 2011, No 6, pp. 209-214.
28. R a w a s h d e h, G., R. B i n M a m a t, Z. B. A. B a k a r, N. H. A. R a h i m. Comparative between Optimization Feature Selection by Using Classifiers Algorithms on Spam E-mail. – International Journal of Electrical & Computer Engineering, Vol. **9**, 2019, pp. 2088-8708.
29. S h a h, F. P., V. P a t e l. A Review of Feature Selection and Feature Extraction for Text Classification. – In: Proc. of International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET'16). IEEE, 2016.
30. S t e i n b a c h, M., et al. A Comparison of Document Clustering Techniques. – In: Proc. of KDD Workshop on Text Mining, Vol. **400**, 2000, pp. 525-526.
31. T a h a, Z., S. R o s t a m. A Hybrid Fuzzy AHP-PROMETHEE Decision Support System for Machine Tool Selection in Flexible Manufacturing Cells. – Journal of Intelligent Manufacturing, Vol. **23**, 2012, No 6, pp. 2137-2149.
32. T a l b i, E. G. Metaheuristics: From Design to Implementation. – Scientific Research, Vol. **74**, John Wiley & Sons, 2009.
33. T r i v e d i, S. K., P. K. P a n i g r a h i. Spam Classification: A Comparative Analysis of Different Boosted Decision Tree Approaches. – Journal of Systems and Information Technology, Vol. **20**, 2018, No 3, pp. 298-105
34. T r i v e d i, S. K., P. K. P a n i g r a h i. Spam Classification: A Comparative Analysis of Different Boosted Decision Tree Approaches. – Journal of Systems and Information Technology, 2018.

35. W a n g, F., T. X u, T. T a n g, M. Z h o u, H. W a n g. Bilevel Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems. – IEEE Transactions on Intelligent Transportation Systems, Vol. **18**, 2017, No 1, pp. 49-58.
36. W e i, T., Y. L u, H. C h a n g, Q. Z h o u, X. B a o. A Semantic Approach for Text Clustering Using WordNet and Lexical Chains. – Expert Systems with Applications, Vol. **42**, 2015, No 4, pp. 2264-2275.
37. A l h e n a w i, E. A., H. A l a z z a m, R. A l-S a y y e d, O. A b u A l g h a n a m, O. A d w a n. Hybrid Feature Selection Method for Intrusion Detection Systems Based on an Improved Intelligent Water Drop Algorithm. – Cybernetics and Information Technologies, Vol. **22**, 2022, No 4, pp. 73-90.
38. A l S o k k a r, A., M. O t a i r, H. E. A l f a r, A. Y. N a s e r e d d i n, K. A l d i a b a t. Original Research Article Sentiment Analysis for Arabic Call Center Notes Using Machine Learning Techniques. – Journal of Autonomous Intelligence, Vol. **7**, 2024, No 3, pp. 1-16.
39. A l S o k k a r, A. A., E. L. C. L a w, D. A. A l M a j a l i, J. A. A l-G a s a w n e h, M. A l s h i n w a n. An Indexed Approach for Expectation-Confirmation Theory: A Trust-Based Model. – Electronic Markets, Vol. **34**, 2024, No 1, 12.
40. O r e h o v a č k i, T., A. A l S o k k a r, J. D e r b o v e n, A. K h a n. Exploring the Hedonic Quality of Slow Technology. – In: Proc. of CHI2013, ACM, Paris, France, 05.01.2013-27.01.2013.
41. A l S o k k a r, A., A. A. M u s a. Multimodal Human-Computer Interaction for Enhancing Customers' Decision-Making and Experience on B2C e-Commerce Websites. Doctoral Dissertation, University of Leicester, 2014.
42. H i j j a w i, M., M. S h i n w a n, M. Q u t q u t, W. A l o m o u s h, O. K h a s h a n, M. A l s h d a i f a t, L. A b u a l i g a h. Improved Flat Mobile Core Network Architecture for 5G Mobile Communication Systems. – International Journal of Data and Network Science, Vol. **7**, 2023, No 3, pp. 1421-1434.
43. A l S o k k a r, A., E. L a w, D. A l m a j a l i, M. A l s h i n w a n. The Effect of Multimodality on Customers' Decision-Making and Experiencing: A Comparative Study. – International Journal of Data and Network Science, Vol. **7**, 2023, No 1, pp. 1-14.
44. A l s h i n w a n, M., A. S h d e f a t, N. M o s t a f a, A. A l S o k k a r, T. A l s a r h a n, D. A l m a j a l i. Integrated Cloud Computing and Blockchain Systems: A Review. – International Journal of Data and Network Science, Vol. **7**, 2023, No 2, pp. 941-956.
45. A l-G a s a w n e h, J. A., M. A l s o u d, A. A l S o k k a r, L. H. W a r r a d, J. S a p u t r a, M. K. D a o u d. Internet Advertisements and Brand Equity Amongst User-Generated Content and Purchase Intention. – Migration Letters, Vol. **20**, 2023, No S8, pp. 467-478.
46. A l-S o u s, N., A. A b d u l l a h, M. T h a' e r, M. A y m a n, A. A l a, M. R a' e d, Z. D a h a l i. Antecedents of e-Commerce on Intention to Use the International Trade Center: An Exploratory Study in Jordan. – International Journal of Data and Network Science, Vol. **6**, 2022, No 4, pp. 1531-1542.
47. A l-G a s a w n e h, J. A., K. N. A l Z u b i, M. M. A n u a r, S. F. P a d l e e, A. u l-H a q u e, J. S a p u t r a. Marketing Performance Sustainability in the Jordanian Hospitality Industry: The Roles of Customer Relationship Management and Service Quality. – Sustainability, Vol. **14**, 2022, No 2, 803.
48. A l s m a d i, A., A. A l f i t y a n i, L. A l h w a m d e h, A. A l_h a z i m e h, J. A l-G a s a w n e h. Intentions to Use FinTech in the Jordanian Banking Industry. – International Journal of Data and Network Science, Vol. **6**, 2022, No 4, pp. 1351-1358.
49. H a m m o u r i, Q., A. M. A l t a h e r, A. R a b a a' i, H. K h a t a y b e h, J. A. A l-G a s a w n e h. Influence of Psychological Contract Fulfillment on Job Outcomes: A Case of the Academic Sphere in Jordan. – Problems and Perspectives in Management, Vol. **20**, 2022, No 3, pp. 62-71.
50. R a b a a i, A., E. A l-l o c i, Q. H a m m o u r i, N. M u h a m m a d, A. A l s m a d i, J. A l-G a s a w n e h. Continuance Intention to Use Smartwatches: An Empirical Study. – International Journal of Data and Network Science, Vol. **6**, 2022, No 4, pp. 1643-165.
51. A l n a s e r, F., S. R a h i, M. A l g h i z z a w i, A. H. N g a h. Does Artificial Intelligence (AI) Boost Digital Baking User Satisfaction? Integration of Expectation Confirmation Model and Antecedents of Artificial Intelligence Enabled Digital Banking. – Integration of Expectation Confirmation Model and Antecedents of Artificial Intelligence Enabled Digital Banking, 2023.

52. H a b e s, M., M. A l g h i z z a w i, M. E l a r e s h i, A. Z i a n i, M. Q u d a h, M. M. A l H a m m a d i. E-Marketing and Customers' Bank Loyalty Enhancement: Jordanians' Perspectives. – In: The Implementation of Smart Technologies for Business Success and Sustainability, Springer, 2023, pp. 37-47.
53. R a h i, S., M. A l g h i z z a w i, A. H. N g a h. Factors Influencing User's Intention to Continue Use of e-Banking During COVID-19 Pandemic: The Nexus between Self-Determination and Expectation Confirmation Model. – EuroMed Journal of Business, Ahead-of-Print, 2022. **https://doi.org/10.1108/EMJB-12-2021-0194**
54. R a h i, S., M. A l g h i z z a w i, A. H. N g a h. Understanding Consumer Behavior toward Adoption of e-Wallet with the Moderating Role of Pandemic Risk: An Integrative Perspective. – Kybernetes, 2023.
55. A l g h i z z a w i, M., M. H a b e s, A. A l A s s u l i, A. A. R. E z m i g n a. Digital Marketing and Sustainable Businesses: As Mobile Apps in Tourism. – In: Artificial Intelligence and Transforming Digital Marketing. Springer, 2023, pp. 3-13.
56. I s t a t i e h, H., M. A l s o u d, J. A l-G a s a w n e h, A. S h a j r a w i, M. Z o u b i. The Impact of Digital Marketing on the Adoption of Building Information Modeling System in Jordanian Interior Design Companies: The Moderating Role of Credibility. – Uncertain Supply Chain Management, Vol. **12**, 2024, No 2, pp. 1267-1274.

Received: 13.11.2023; Second Version: 23.01.2024; Third Version: 20.03.2023; Accepted: 04.04.2024