Ranking-based Feature Selection with Wrapper PSO Search in High-Dimensional Data Classification

Thinzar Saw and Win Mar Oo

***Abstract*—The use of feature selection approaches is frequently required for handling high-dimensional datasets. Applying the particle swarm optimization (PSO) algorithm to high-dimensional datasets with thousands of features is still a major challenge in data mining and machine learning. Considering the limitations of the feature model, the key difficulty is choosing usable features. To address this challenge, a two-phase approach is suggested to optimize the feature selection used for high-dimensional data classification. First, the feature ranking metric (FR) is used to select the relevant and non-redundant feature set suitable for classification. In the second phase, the geometric particle swarm optimization (GPSO) algorithm is used to find the most informative feature subset from the original attributes using the selected features acquired from the first phase. To further improve the classification accuracy and to make the proposed approach more accurate, a new fitness function is designed with classification accuracy, feature relevancy, and feature reduction rate to calculate the goodness of selected features. The efficiency of FR-GPSO is evaluated on five microarray datasets and six benchmark datasets from the UCI machine learning repository, evaluated by a k-nearest neighbor (kNN) classifier with 10-fold cross-validation and compared with the commonly used feature selection methods. The results show that the proposed method not only selects a smaller number of attributes but also increases the classification accuracy compared with other approaches. Furthermore, the statistical test shows that the proposed method is statistically significant over the other competitors' algorithms.**

***Index Terms*—Particle Swarm Optimization, Feature Ranking Metric, High Dimensional Data, Feature Selection, Classification**

I. Introduction

GRADUALLY, both the size of the data sample and the number of attributes have been emergent for the last few years in different areas. A serious problem is recognized as the issue of dimensionality reduction while applying data mining and learning algorithms to high-dimensional data. In addition, a large number of features will significantly increase the requirements for computational and memory storage. There are numerous issues with classification because of the massive size of the data. Therefore, it is difficult to learn good classifiers before deleting the unnecessary features.

Manuscript received April 7, 2022; revised September 28, 2022.

Thinzar Saw is a Ph.D. candidate in the Data Mining and Machine Learning Lab at the University of Computer Studies, Mandalay, Myanmar. (e-mail: thinzarsaw@ucsm.edu.mm).

Win Mar Oo is a Pro-Rector of the University of Computer Studies, Mandalay, Myanmar. (e-mail: winmaroo@ucsm.edu.mm).

Feature selection (FS) has been proven to be effective and efficient in handling these problems by removing non-relevant and duplicated features to improve learning and better understanding for building models and data. The general structure of feature selection procedures for many applications is shown in Figure 1.

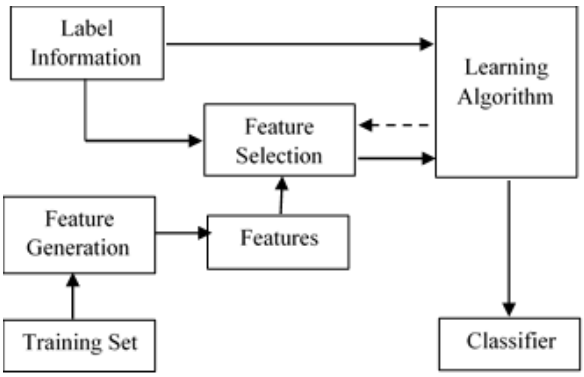

Fig. 1. A General Structure of Feature Selection

It is an essential role in data mining and pattern recognition. It can be generally categorized into three methods: the filter method, the wrapper method, and the embedded method, based on different aspects. Filter methods apply statistical measures and independent criteria. The wrapper methods predetermine the learning model to evaluate feature relevance. Embedded methods trade off the solution between filter and wrapper methods. Therefore, it uses both independent criteria and predetermined learning algorithms. All methods have their benefits and drawbacks in the feature selection process [2].

Although filter approaches are less time-consuming and suitable for high-dimensional data, they obtain lower classification accuracy than the wrapper methods. Wrapper methods take more time than filters due to the training model's being convoluted in objective function evaluation. But it provides a good result. The way of combing the processes of filter and wrapper with their strengths is still challenging. To achieve a good performance between the computational efficiency of the filter approach and the wrapper approach, different strategies have been proposed in [2], [3], and [9]. Therefore, this paper aims to hybridize the important points of these techniques on high-dimensional data by using particle swarm optimization (PSO) search with feature ranking.

In addition to this, meta-heuristic techniques are merged with classification techniques to solve feature selection problems. In general, the process of selecting the features can be considered a problem of global optimization.

According to its global search power, the swarm intelligence algorithm can effectively examine an optimized subset of features. Dealing with high-dimensional datasets typically takes a very long time. Redundant and unrelated features not only cause time consumption but also degrade the performance of classification accuracy, particularly in bioinformatics on high-dimensional data. In most heuristic methods, swarm intelligence algorithms have attained great success, like PSO for feature selection [4]. However, the solution space is too large with the growth of feature quantity, resulting in lower search efficiency [5].

In the last few years, bio-inspired optimization algorithms have become famous for solving combinatorial and complex problems. Many of these algorithms have been well adopted in the feature selection problem for high dimensional classification, for instance, Particle Swarm Optimization [6-8], Ant Colony Optimization [9-11], and Firefly Algorithm [12]. Among them, PSO is one of the known approaches, and it has been successfully applied [13], [14], and [15]. In this regard, a hybrid two-phase feature selection algorithm is proposed to attain higher classification accuracy by combining the characteristics of PSO with the feature ranking measure in this paper. The first phase of the proposed algorithm uses the feature ranking metric to select the relevant and non-redundant feature set suitable for classification. The geometric PSO search is conducted in the second phase, to find the most informative feature subset using the selected feature set obtained from the first phase. To enhance the accuracy, even more, the k-nearest neighbors (kNN) classifier is utilized to effectively classify the instances. The results of the proposed method pick out a better representative feature subset by improving classification accuracy.

To achieve better classification accuracy with fewer attributes from the original datasets, the contributions of this paper are:

- A two-phase dimensional reduction approach has been developed by integrating filter and wrapper methods to improve performance.
- A feature ranking metric is suggested as a pre-filtering phase to help in identifying highly discriminative features.
- A wrapper-based particle swarm optimization algorithm with a new fitness function is designed to generate the attribute candidate solutions.

The performance of the proposed approach is tested on five microarray datasets and six UCI datasets. For comparisons, several other feature selection techniques were selected and analyzed.

This section is an introduction to feature selection in high-dimensional data classification. Subsequently, the rest of the paper is described as follows. Several feature selection techniques in the literature that also serve as a background for the proposed method are reviewed in Section II, and the proposed method for selecting the informative feature in high-dimensional data is presented in Section III. Sections IV and V detail the experiments and present a comparison of the proposed approach with existing methods. The conclusion is reported in Section VI.

II. Related Works

High-dimensional data plays a significant role in data mining prediction models. Feature selection is necessary for better analysis due to its high dimensionality. Prior works on feature selection for classification have been focused, and research efforts have primarily been directed toward high-dimensional data.

In the aspect of feature selection in high-dimensional classification, PSO-based feature selection methods have achieved great success in the literature. Xian-fan Song et al. [14] explored the performance of PSO optimization based on a space division strategy. To categorize related features into the same subspace, the size of the swarm's adaptive adjustment mechanism is proposed for maintaining a suitable size for each sub-swarm. In feature selection issues, the approach of eliminating redundant particles and producing new particles is suggested to guarantee the quality of particles in sub-swarms. The experiments were conducted on 12 datasets from the UCI repository. The results have shown that the number of features is nearly reduced by 30% and most of the datasets have significantly better accuracy compared with other algorithms

Binh Tran et al. [15] developed an adaptive multi-swarm optimization (AMSO) method that selects a feature subset of high-dimensional data successfully. This study used symmetrical uncertainty (SU) to rank features (present individually relevant features). The authors also presented how to use the divide-and-conquer strategy flexibly and dynamically. PSO allows searching in tiny subspaces effectively and efficiently while still covering the entire huge search space. Using several sub-swarms focused on smaller subspaces in high-dimensionality issues, this strategy not only increased its performance but also reduced its running time.

In [16], a combination of GA and learning automata was used effectively for gene selection problems on different cancer datasets. In [17, 2], PSO combined with the kNN classifier was applied to evaluate the fitness of each particle in wrapper-based methods. While the method in [17] could select more relevant features in a shorter time than the other compared methods, the method in [2] obtained a higher classification accuracy compared with the state-of-the-art filter and wrapper FS method based on the size of the selected features and running time.

Moreover, the scoring criterion with an improved PSO method was applied to select highly relevant genes in [18], and improved PSO coupled with the ELM method was used to perform gene selection to obtain a compact set of effective genes. Lin et al. [19] implemented the classification of tumors for gene selection, which was demonstrated on the gene expression dataset. The authors also presented a hybrid gene selection method based on the Relief and Ant Colony Optimization (ACO) algorithms by combining the filter and wrapper method. The classification accuracy of this study was increased with the representative genes in all datasets, and the dimensionality of the genes was significantly reduced.

Chen and partners studied a cost-effective feature selection scheme using BPSO and feature confidence calculation using ReliefF, which is generally used in feature selection methods. The key inspiration of this work is that

most of the feature selection schemes only use accuracy as the evaluation criteria, ignoring the different achievement costs of different features. Another issue that was reported in this work was that the position of the particle was updated due to the swarm's overall fitness without taking into account the various effects of each feature in the particle [20].

The research works [21, 22, and 23] used a classification algorithm for evaluation criteria implemented as a wrapper approach using PSO. It examines how good the selected feature subset is. To design the fitness value balances between the accuracy of the selected subset of features, C4.5, kNN, and SVM classifiers were used with cross-validation on the training data in the above references, respectively.

A unique tunable swarm size approach is proposed by [24] to search for the best initial swarm size for the given data to overwhelm the local minimum problem. The feature discrimination score is designed as an objective function to get better classification accuracy and evaluate the effectiveness of the feature subsets. Experiments on 10 benchmark datasets obtained by the proposed approach improved the classification accuracy with the relevant feature subsets in comparison to other methods.

Furthermore, to improve the performance of classification algorithms and reduce the number of effective genes, Alekhya et al. [25] combined the correlation coefficient with PSO and an Extreme Machine Learning (EML) Classifier was applied in the fitness evaluation process for gene selection. Compared to the traditional tree-based classifiers like J48, decision stump, etc., the experimental results obtained by the proposed approach showed better classification accuracy and fewer gene subsets on the tested gene expression data.

In recent times, the improved versions of PSO with two stages for feature selection have become more widely developed in many domains. Qing Wu et al. [26] presented a hybrid Improved Binary Quantum PSO algorithm integrated with a filtering method (maximum information coefficient, MIC) to decrease the length of the dataset for feature selection. The design of the fitness strategy using the weighted average principle is applied to balance the number of chosen features and classification accuracy. The outcomes evaluated on 36 datasets from UCI proved that the proposed approach produced a more accurate classification level using three classifiers (SVM, MNB, and kNN) with the least selected features than the baseline methods.

Another improved binary PSO algorithm was proposed in [27], introducing two factors as Levy flight local search factor and a global search factor based on weighting inertia coefficient, and two mechanisms like mutation and binary, on 16 classical datasets for bio-inspired feature selection. In this study, kNN is applied as an evaluator to implement the wrapper method. The results demonstrated that the improved approach has better performance than other existing algorithms in terms of accuracy rate and the selected features.

Furthermore, based on PSO with learning memory, Bo Wei et al. [28] developed a new efficient feature selection algorithm to balance the local exploitation and the global exploration for higher fitness and faster progress in the feature selection problems. In this approach, each particle acquires from all individuals the best of the current generation and the prototype produced instead of obtaining from the global and local best position. The 10-fold cross-validation method with the kNN classifier is employed for candidate feature subset evaluation on some standard datasets. Compared with well-known wrapper-based approaches, the results verified that the outcome of the proposed algorithm is statistically significant and reasonable.

From the above-collected works, there are various feature selection approaches with PSO designed over the last decade. However, there is still a challenge for PSO that may perform differently in different feature selection applications. Therefore, this study was developed to address this issue by combining feature ranking followed by PSO search to pick the best features and to enhance the evaluation performance dealing with high-dimensional data classification.

III. Two-Phase Feature Selection Approach

A two-phase feature selection approach based on the PSO optimization process for high-dimensional classification is developed in this work. The proposed method includes two main parts, namely, pre-filtering and wrapper-based feature selection processes. Fig. 2 demonstrates the overview structure of the proposed method in this work.

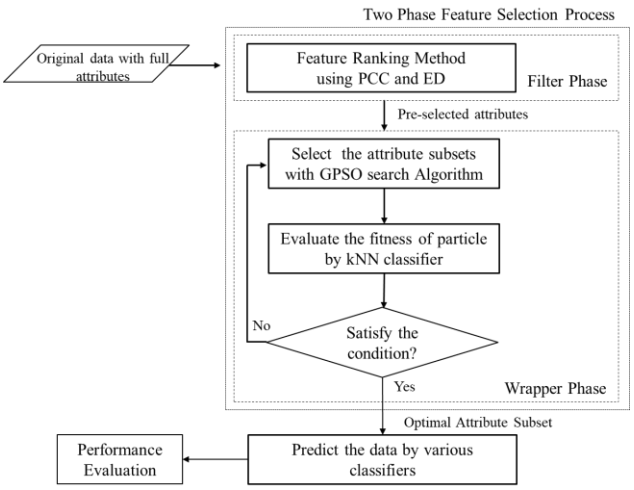

Fig. 2. Overview of the Proposed Approach

The feature ranking-based method (FR) is discussed for the pre-filtering process, and then geometric PSO (GPSO) search is used for the wrapper process; finally, the two-phase algorithm FR-GPSO is developed for high-dimensional classification. After transforming the features as informative data, the useful features are extracted from the original features according to Algorithm 1 and Algorithm 2, respectively. As soon as selecting the top-ranked features from the strongly relevant and un-redundant features using Algorithm 1 for different datasets, wrapper-based feature subset selection based on GPSO search is carried out to receive the optimal trained kNN model with the highest relevancy scores in Algorithm 2. An optimal feature subset is the output of this phase.

A. Phase 1: Feature Ranking (FR)

FR plays the role of a pre-filtering process and makes time convenient to process without pre-processing such as discretization [29]. It has two aspects: correlation and distance measures. Some mathematical symbols are noted to easily understand the ranking approach. Given the input dataset $D = \{A, C\}$tabled as N samples, M attributes A = $\{a_i, i = 1, \dots, M\}$ and the target class C. Finding a subset of m attributes, A^m, which is selected from the M attributes in the original set A^M, and classifying the target class $C$are the main aids.

Correlation Measure

To select an attribute set $A^m$with the highest relevance to the class label C, the Pearson correlation coefficient (PCC) is used due to its implementation being easy and is appropriate for computing continuous variables. For two random variables (attribute A and class label C), Pearson's correlation is given as:

$$PCC\,(A_i, C_i) = \frac{Cov\,(A,\ C)}{\sqrt{var\,(A).var(C)}} \quad (1)$$

$$Cov\,(A, C) = \sum(A_i - \bar{A}).(C_i - \bar{C}) \quad (2)$$

$$var(A) = \sum(A_i - \bar{A})^2, where\ \bar{A} = \frac{1}{n}.\sum_{i=1}^{n} A_i \quad (3)$$

$$var(C) = \sum(C_i - \bar{C})^2, where\ \bar{C} = \frac{1}{n}.\sum_{i=1}^{n} C_i \quad (4)$$

where, A_i and C_i are the i^{th} element of A, and C. The maximum relevancy score ($maxRel$) for the attribute i is defined as:

$$maxRel_i = |PCC\ (A_i, C_i)|(1 \le i \le M) \quad (5)$$

where $A_i$is the i^{th}attribute from each instance and C_i is every element that comes from the class label C of each instance. The greater the PCC value, the higher the significance between the attribute A and target class C.

Distance Measure

The Euclidean distance (ED) is used to measure the redundancy between two attributes A_i and A_j. This measure is the best proximity measure when data is dense or continuous. The distance between two attributes is easily calculated by using Eq. (6).

$$ED(A_i, A_j) = \sqrt{\sum_{i=1}^{n}(A_i - A_j)^2} \quad (6)$$

where, n is the number of instances, A_i and A_j are the two feature vectors. The maximum distance ($maxDis$) score of the attribute $i\ (1 \le i \le M)$ can be defined as:

$$maxDis_i = \frac{1}{M-1}\sum ED\ (A_i, A_j), (1 \le j \le M, i \ne j) \quad (7)$$

The attribute subset obtained from the maximum distance has minimal redundancy. The greater the distance, the more independence there is.

After calculating these two measures, the relevancy and redundancy scores are combined by optimizing the following situation:

$$FR_{Score\ (i)} = Rel_i + Dis_i \quad (8)$$

To obtain the normalized value, calculate the maximum value from the above score:

$$\max = FR_{Score\ (i)} \quad (9)$$

Each combined value FR_{Score} is divided by this maximum value to scale the data and sort it in descending order. In this approach, the learning model is applied to determine the selected attribute subset A^m with the highest accuracy rate. Algorithm 1 describes the procedure of feature ranking as a pre-filtering process.

Algorithm 1: Feature Ranking (FR)

Input: Dataset $D = \{A, c\}$with a set of attributes A = $\{a_1, a_2 \dots a_n\}$ and the target class *c*.

Output: pre-selected attributes $A_{pre-sele}$

1. Calculate PCC values for each attribute a_i and class label *c* using $Eq.(1)$.
2. Calculate ED values for every attribute $a_i, a_j \in D, i \ne j$ using $Eq.(6)$.
3. Add the PCC values and ED values for each attribute a_i using *Eq.* (8) as ranking values.
4. Calculate the maximum value using *Eq.* (9) and then each ranking value is divided by this maximum value to scale the data.
5. Sort the normalized scores in descending order.
6. For each attribute a_i, do the following:
   a) Randomly split the training and testing data using ten-fold cross-validation.
   b) Evaluate the accuracy rate of the ten-fold cross-validation with the kNN classifiers for each fold.
7. Determine the best attribute subset with the highest accuracy rate.
8. To construct $A_{pre-sele}$, keep the above-selected attributes set with ranking scores.

B. Phase 2: Geometric PSO

The Geometric PSO plays the role of the wrapper and differs from the standard PSO. It has no velocity and mutation [30]. The convex combination described in Eq. (10) is used for a position update. Algorithm 2 shows the feature subset generation using a wrapper-based GPSO search.

Algorithm 2: Wrapper-based GPSO Search

Input: Training data, w_1, w_2, w_2, m_ **popSize**
Output: the best attributes subset A_{best}

1. Randomly initialize the population of particles
2. Initialize the position x_i of each particle *i* in the search space
3. **Do**
4. **For** *i* to **m_popSize do**
5. Calculate the fitness value of each particle using 10-fold cross-validation with kNN classifier using *Eq. (11)*
6. Set the personal best $\hat{x}_i$ as the best position of each particle
7. Set the global best $\hat{g}$ as the best position with the best fitness value of the particle swarm
8. **End for**
9. **For** *i* to **m_popSize do**
10. Update position x_i using a randomized convex combination equation:
$$x_i = CX\,((x_i, w_1), (\hat{g}, w_2), (\hat{x}_i, w_3))$$
11. Mutate the position x_i
12. **End for**
13. **While** (maximum iterations or stopping criteria are not met)
14. Return the best attributes subset A_{best} (informative attributes)

The parameters w_1, w_2 ,and w_3 are non-negative, w_1 represents inertia weight, w_2 represents social weight, w_3 represents individual weight, and add up to one. $\hat{x}_i$ and $\hat{g}$ are the current best position of the individual particle and the best position of the global particle, respectively. In GPSO, the mutation is a conventional bit-wise mutation.

Convex Combination

A convex combination [30] is a linear set of points (which can be vectors, scalars, or more complex points) in a horizontal space with non-negative coefficients and the sum is one. In a real vector space, it is a convex combination of a finite number of points $x_1, x_2, \ldots, x_n$, which is of the form:

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n \quad (10)$$

where the real numbers α_i satisfy $\alpha_i \geq 0$ and $\alpha_1 + \alpha_2 + \ldots + \alpha_n = 1$. The complete set of particles then searches the search space for the best subset of features.

Using a randomized convex combination, Eq. (10), the position of each particle, the best experience of each particle, and the best experience of the swarm are restructured in each iteration. After many iterations, the particle with the best fitness value is selected as the best feature selection set.

Design of fitness function with feature relevancy

The aim of the fitness function is to improve the quality of each particle's evaluation. To select the optimal attribute subset with compromise accuracy, a new fitness function that represents the classification accuracy, feature relevancy, and feature reduction rate defined by a specific particle is shown in Eq. (11).

$$fitnessFun = acc + \frac{Rel(a)}{Rel(A)} \cdot \left(1 - \frac{\#a}{\#A}\right) \quad (11)$$

where a represents the selected attribute subset, A represents the full set of attributes consisting of all attributes in the dataset D, acc represents the accuracy obtained using the kNN classifier, $Rel(a)$ represents the sum of relevancy value of the attribute set a, $Rel(A)$ represents the sum of relevancy value of the attribute set A and therefore $(1 - \#a/\#A)$ indicates the feature reduction rate, # indicates the number of attributes. For feature relevancy of each feature, the PCC + ED value is used.

The kNN classifier [31] is employed to generate candidate feature subset solutions because it is a distance-based algorithm and the increased computation cost is slight. The training and test sets are transformed based on the feature subset and the parameter set into k = 1 to simplify performance evaluation. To estimate the testing accuracy of the classifier, 10-fold cross-validation is performed.

The next section will discuss the experimental details using the proposed approach. And then comparative analysis with other approaches will be performed.

IV. Experimental Design

In this section, the implementation of the proposed methodologies is presented. Various experiments have been conducted to compare the performance of the proposed FR-GPSO with all features (AF), the standard feature ranking FR (baseline) method, and W-PSO (i.e., the wrapper FS technique without a filter method) concerning the classification accuracy and the number of selected features for feature selection in high dimensional classification. The comparison is based upon the results of experiments evaluated on the best-known 5 microarray datasets and six benchmark datasets from the UCI Machine Learning Repository, as shown in Tables I and II.

Our methodology is implemented by the following technologies.

Software: Eclipse Oxygen, Windows 10 Pro 64-bit operating system, Microsoft Excel, Weka.

Hardware: Intel(R) Core(TM) i5-3210M CPU@2.50GHz machine, 4 GB RAM, Intel(R) HD Graphics 4000, AMD Radeon(TM) R9 M375.

In this paper, the implementation of various FS techniques was performed using Java programming and WEKA (3.8) [32] to run the experiments. Similar conditions are also set to compare the performance among comparative algorithms.

TABLE I
MICROARRAY DATASETS USED IN EXPERIMENTS

Dataset	#Sample	#Features	Class
Colon Tumor	62	2000	2
CNS	60	7129	2
Leukemia_2c	72	7129	2
Lymphoma	66	4027	3
SRBCT	83	2308	4

TABLE II
UCI DATASETS USED IN EXPERIMENTS

Dataset	#Sample	#Features	Class	Area
Madelon	2600	500	2	Artificial
Musk2	6,598	168	2	Physical
Movement	360	91	15	Brazilian sign language
Arrhythmia	452	279	16	Life
Soybean	307	35	19	Life
Isolet5	1559	617	26	Spoken Letter Recognition

A. Data collection and preprocessing

To verify the performance of the proposed approach, five binary and multi-class cancer microarray datasets were used for the gene selection problem [33, 34]. Furthermore, the six different areas of datasets from the UCI machine learning repositories [35] are used to evaluate the proposed algorithm.

In this study, a stratified k-fold cross-validation procedure with k = 10 is used and performed with 10 iterations for experiments. The input dataset is partitioned into ten subsets. The remaining nine folds (90%) are used as the training set for building the model and the remaining folds (10%) as the test set to justify the performance of the model. Each fold is randomly generated and the number of iterations is repetitive 10 times.

B. Parameters Setting

In all experiments, the population size (i.e., `m_popSize`) of all comparative methods was set to 30 and the maximum generations were set to 100. The parameter values of the

algorithms are listed in Table III.

TABLE III
PARAMETER SETUPS OF DIFFERENT ALGORITHMS

Algorithms	Value of key parameters	
GA	Mutation rate	0.02
	Crossover probability	0.6
EA	Mutation Probability	0.1
	Crossover Probability	0.6
ACO	Parameter ϕ	0.8
	Parameter α	1.0
	Parameter β	0.1
BPSO	Accelerating coefficients (c_1 and c_2)	1.4962
	Inertia weight (w)	0.7298
PSO(4-2)	Accelerating coefficients (c_1 and c_2)	2.0
HPSO_LS	Parameter ε	0.5
	Parameter α	0.65
FR-GPSO	Mutation Probability	0.01
	Inertia weight w_1	0.33
	Social weight w_2	0.33
	Individual weight w_3	0.34

V. RESULT ANALYSIS AND DISCUSSION

The core standard comparison examines the number of informative features and their impacts on the accuracy of the fitness function. The experiments have two portions. In the first portion, basic wrapper-based methods such as genetic algorithm (GA), evolutionary algorithm (EA), ant colony optimization (ACO), and particle swarm optimization (PSO), which are applied for feature subset selection problems, are selected for performance comparison.

In the other portion, the proposed method (FR-GPSO) is compared with existing feature selection methods containing: Binary PSO (BPSO) [36], a new initialization and updating mechanisms using PSO for feature selection (PSO (4-2)) [37], and a combination of PSO feature selection with a novel local search strategy (HPSO-LS) [2].

A. Performance of Two-Phase (FR-GPSO) Approach

In this section, a set of experiments has been conducted to analyze the performance of two-phase feature selection strategies. The experimental results consist of two parts: classification accuracy and the number of selected features. The best results are bolded in each row.

1) Microarray Data

Firstly, the performance of the proposed approach on five microarray datasets is tested in terms of two evaluation measures shown in Table IV and V. Table IV shows the number of selected attributes obtained in each phase of the proposed approach for microarray datasets. The second column represents the original number of attributes. The third column contains the number of selected attributes after applying feature ranking. The last column is the number of attributes after applying the GPSO algorithm. The results obtained using the proposed approach achieve the lowest number of selected attributes on all datasets. The smaller the number of selected attributes, the lesser the amount of computation time for processing.

TABLE IV
NO. OF SELECTED ATTRIBUTES OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASES

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Colon	2000	110	**32**
CNS	7130	266	**75**
Leukemia_2c	7129	1820	**554**
Lymphoma_66	4027	52	**14**
SRBCT	2308	145	**69**

Table V describes the classification accuracy with the kNN classifier obtained by the proposed approach for each microarray dataset. The results of the classifier on all datasets with original features are reported in the "All Features" column. The accuracy rates of Colon, CNS, Leukemia_2c, Lymphoma_66 and SRBCT are 96.77%, 80%, 97.22%, 100% and 97.59% respectively. According to Table V, the classification accuracy is significantly improved in the second phase compared to the first phase and the initial number of attributes.

TABLE V
CLASSIFICATION ACCURACY OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASES (KNN)

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Colon	70.97	80.65	**96.77**
CNS	56.67	70.00	**80.00**
Leukemia_2c	87.50	95.83	**97.22**
Lymphoma_66	98.48	100.00	**100.00**
SRBCT	84.34	92.77	**97.59**

In order to further study, the performance of FR-GPSO, C4.5 is employed as the classifier instead of k-NN. Table VI shows the classification accuracy of the proposed method with C4.5. It can be seen that all five datasets achieve the best performance with C4.5 classifiers. For microarray datasets, k-NN in Table V motivates the proposed approach to obtain higher accuracy compared with that of the C4.5 in Table VI.

TABLE VI
CLASSIFICATION ACCURACY OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASES (C4.5)

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Colon	77.42	83.87	**90.32**
CNS	58.33	75.00	**76.67**
Leukemia_2c	83.33	91.67	**94.44**
Lymphoma_66	92.42	92.42	**98.48**
SRBCT	84.34	89.16	**90.36**

2) UCI Benchmark Data

Second, to examine the effectiveness of the proposed approach, the experiments conducted on six UCI datasets are recorded in Table VII and Table VIII. The results of the size of the best attribute subset produced by the proposed

approach are shown in Table VII. FR-GPSO produced subsets with significantly fewer attributes in all datasets.

TABLE VII
NO. OF SELECTED ATTRIBUTES OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASE

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Madelon	500	13	**7**
Musk2	167	145	**61**
Movement	90	67	**32**
Arrhythmia	279	201	**60**
Soybean	35	30	**19**
Isolet5	617	523	**263**

Table VIII describes the accuracy results as evaluated by the kNN classifiers for each UCI dataset. It can be observed that the classification accuracy is significantly improved in phase 2 compared to phase 1. As seen in Tables VII and VIII, the performance of all UCI datasets was very satisfying, with maximum accuracy, and reduced the number of attributes significantly. Phase 1 removes over 70% of attributes that are considered irrelevant and redundant, except for the Madelon dataset. As well, phase 2 reduces under 50 % of the attributes and finds the optimal subset of attributes to improve the classification accuracy.

TABLE VIII
CLASSIFICATION ACCURACY OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASE (KNN)

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Madelon	54.27	87.77	**88.65**
Musk2	95.80	95.98	**97.68**
Movement	85.83	87.78	**88.33**
Arrhythmia	53.76	55.53	**65.49**
Soybean	87.62	88.27	**93.81**
Isolet5	85.63	86.47	**87.81**

Table IX shows the results as evaluated by the C4.5 classifiers. FR-GPSO achieves the highest accuracy with 5 of the 6 datasets. On Musk2 and Soybean datasets, FR-GPSO's accuracy was slightly higher and fewer on the Madelon datasets, than in the pre-filtering process. In summary, FR-GPSO has the highest accuracy with fewer selected attribute subsets on six UCI datasets by two classifiers.

TABLE IX
CLASSIFICATION ACCURACY OBTAINED BY THE PROPOSED APPROACH IN THE FIRST AND SECOND PHASE (C4.5)

Datasets	All Features	Phase-1: FR	Phase-2: FR-GPSO
Madelon	69.04	**81.69**	81.65
Musk2	96.88	97.35	**97.65**
Movement	69.72	69.44	**70.56**
Arrhythmia	64.82	67.04	**69.91**
Soybean	86.97	89.58	**90.55**
Isolet5	78.00	78.58	**79.09**

B. Comparison with basic wrapper methods

To demonstrate the performance of the proposed approach, several experiments were performed to compare with well-known wrapper-based methods such as GA, EA, ACO, and PSO. The results have been reported in terms of classification accuracy and the number of selected attributes tested on five microarray and six UCI datasets in Table X and Table XI, respectively.

It can be seen in Table X that different feature subsets have been produced by the proposed approach and those of GA, EA, PSO, and ACO algorithms for each microarray dataset. The proposed approach produces a small number of attributes for all microarray datasets and has the highest classification accuracy among those four competitors on all datasets. This is because the suggested method is a wrapper-based method that evaluates attribute subsets using a learning model; thus, both target class and feature relevancies are taken into account while selecting appropriate feature subsets.

TABLE X
PERFORMANCE COMPARISON WITH KNN CLASSIFIER ON SELECTED ATTRIBUTES FOR FIVE MICROARRAY DATASETS

Dataset	Methods	Size	Accuracy
Colon	GA	466	80.65
	EA	443	82.26
	PSO	484	80.65
	ACO	639	83.87
	FR-GPSO	**32**	**96.77**
CNS	GA	1149	66.67
	EA	3128	63.33
	PSO	1395	73.33
	ACO	2000	68.33
	FR-GPSO	**75**	**80.00**
Leukemia_2c	GA	3371	91.67
	EA	3383	91.67
	PSO	3126	93.06
	ACO	2810	91.67
	FR-GPSO	**554**	**97.22**
Lymphoma_66	GA	1146	98.48
	EA	759	100.00
	PSO	1376	100.00
	ACO	738	100.00
	FR-PSO	**14**	**100.00**
SRBCT	GA	1094	91.57
	EA	1086	90.36
	PSO	850	92.77
	ACO	850	92.77
	FR-PSO	**69**	**97.59**

As seen in Table XI, the optimal attribute subsets for all these algorithms on six UCI datasets were compared in terms of accuracy. The FR-GPSO approach outperformed the other algorithms on three of the six UCI datasets. The result of the size of the attribute subset for each approach performed significantly well on all datasets with even a third fewer attributes than the other algorithms. From these experiment results, it can be concluded that the proposed approach FR-GPSO produced satisfying results compared to other wrapper-based algorithms on the two evaluation criteria.

C. Comparison with existing feature selection methods

Three existing feature selection algorithms, which are BPSO, PSO (4-2), and HPSO-LS, are compared to the suggested approach (FR-GPSO). They are metaheuristic-based methods designed for feature selection problems.

TABLE XI
PERFORMANCE COMPARISON WITH KNN CLASSIFIER ON SELECTED ATTRIBUTES FOR SIX UCI DATASETS

Dataset	Methods	Size	Accuracy
Madelon	GA	225	58.69
	EA	262	59.73
	PSO	244	61.63
	ACO	218	61.31
	FR-PSO	**7**	**88.65**
Movement	GA	44	87.78
	EA	50	88.33
	PSO	39	**89.72**
	ACO	39	**89.72**
	FR-PSO	**32**	88.33
Musk2	GA	91	96.59
	EA	76	97.26
	PSO	80	**98.86**
	ACO	83	97.11
	FR-PSO	**61**	97.68
Arrhythmia	GA	106	65.27
	EA	126	63.05
	PSO	78	63.94
	ACO	104	64.38
	FR-PSO	**60**	**65.49**
Soybean	GA	26	92.18
	EA	19	92.83
	PSO	23	91.86
	ACO	27	92.18
	FR-PSO	**19**	**93.81**
Isolet5	GA	305	87.88
	EA	306	87.43
	PSO	294	**88.97**
	ACO	308	88.84
	FR-PSO	**263**	87.81

First, the comparison results of the kNN classifier on five microarray datasets are shown in Figs. 3 and 4 in terms of accuracy and the number of selected attributes, respectively.

From Fig. 3 and 4, it can be seen that the proposed approach attained the highest accuracy and the minimum number of selected attributes on all of the datasets. This is caused by the fact that it is useful to remove unnecessary and redundant features in order to select a subset of features. For instance, for the data set Lymphoma_66, FR-GPSO acquires a subset of the features involving the 14 lowest features and reaches the peak classification accuracy (100.00%) compared to the other methods. Therefore, it can be concluded that the proposed approach was able to overtake the state-of-the-art feature selection methods.

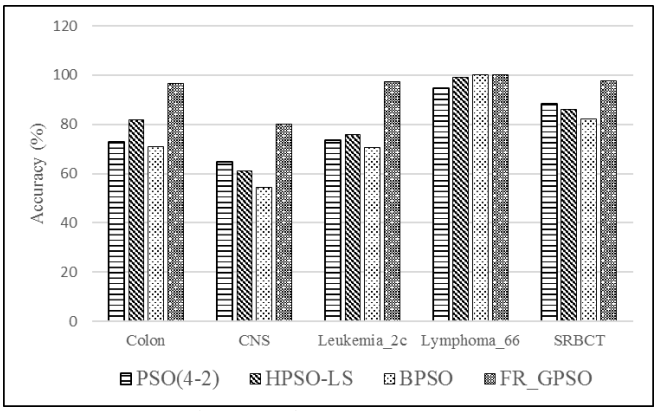

Fig. 3. Comparison of accuracy from the proposed approach with existing state-of-the-art algorithms on five microarray dataset

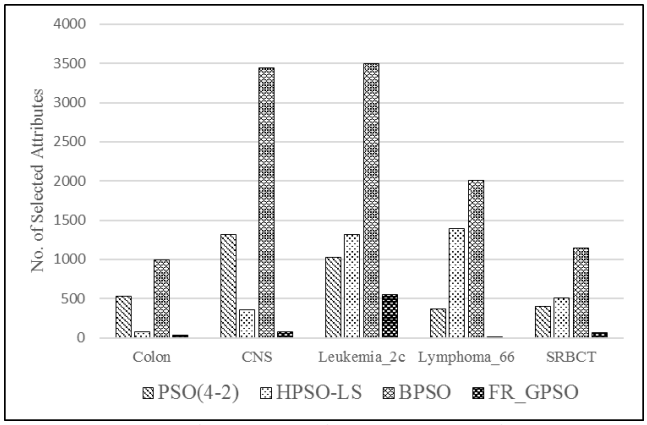

Fig. 4. Comparison of the number of selected attributes from the proposed approach with existing state-of-the-art algorithms on five microarray datasets

In addition, Figs. 5 and 6 illustrate the experimental results of six UCI datasets in terms of accuracy with the kNN classifier and the number of selected attributes for the proposed approach. By analyzing the experimental data, the FR-GPSO has the highest accuracy of all four methods for six UCI datasets. The significant test findings in Fig. 5 show that FR-GPSO performed better than other comparative methods in terms of accuracy in all cases.

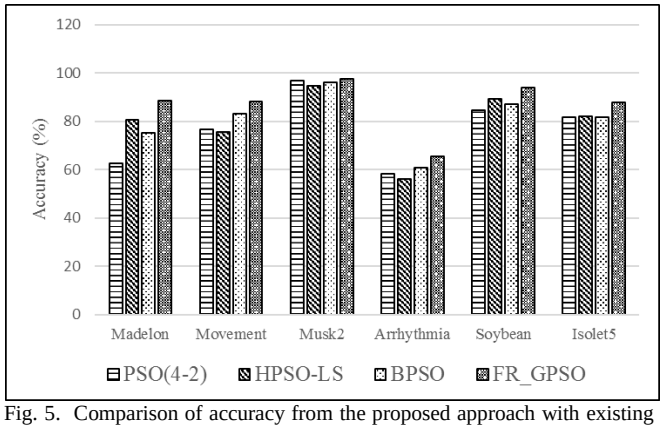

Fig. 5. Comparison of accuracy from the proposed approach with existing state-of-the-art algorithms on six UCI dataset

Moreover, it can be seen in Fig. 6 that, for each dataset, different attribute subsets have been selected by the proposed approach and those of PSO (4-2), HPSO-LS, and BPSO, and the proposed approach attains the smallest attribute subsets for all six UCI datasets. This is because it is useful to eliminate inappropriate and redundant attributes for attribute subset selection. For example, for the Madelon dataset, FR-GPSO acquires a subset of attributes containing the number of attributes (7). It reaches the highest classification accuracy (88.65%) compared with the other three methods.

According to the results of all figures, it can be indicated that FR-GPSO picks out the lowest number of attributes as well as attains the highest classification accuracy among those three methods. In contrast, the goal of feature selection, which is to eliminate irrelevant and redundant attributes without compromising the classification accuracy, has been validated by the experiments mentioned above. There is a problem with whether there are significant differences between the proposed approach and those methods. Therefore, statistical analysis is done in the next section.

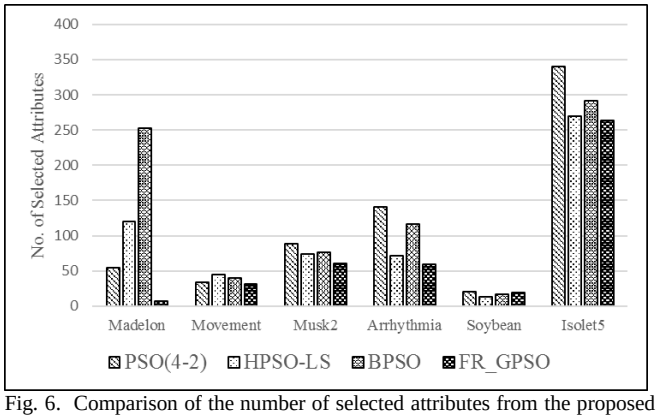

Fig. 6. Comparison of the number of selected attributes from the proposed approach with existing state-of-the-art algorithms on six UCI datasets

D. Statistical Analysis

Additionally, to compare the performance of different feature subset selection techniques statistically, the Friedman test [38] is used. The Friedman test is a non-parametric test that ranks each approach on each dataset to quantify statistical differences between methods over numerous datasets. It has been used in various studies to statistically examine feature selection methods [2, 26]. According to the number of methods, for each subset of features, the different accuracies are ranked. The higher the algorithm's ranking, the better the algorithm's performance.

The Friedman estimator is well-defined by:

$$F_F = \frac{(N_B - 1)\, X_f^2}{N_B\,(N_M - 1) - X_f^2} \quad (12)$$

where N_B represents the number of datasets and N_M is the number of methods.

$$X_f^2 = \frac{12 N_B}{N_M\,(N_M+1)} \left(\sum_{j=1}^{N_M} R_j - \frac{N_M\,(N_M+1)^2}{4} \right) \quad (13)$$

where R_j is the ranking of each method.

F_F follows a Fisher distribution X_f^2 with $N_M - 1$ and $(N_M - 1)\,(N_B - 1)$ degrees of freedom. The critical value of the Fisher distribution is set to $\propto = 0.05,\ 95\%$ a confidence interval in this experiment. The *p-value* which is less than 0.05 indicates that the classification accuracy of those algorithms differs significantly in all of the five microarray datasets.

TABLE XII
FRIEDMAN-TEST RESULTS AMONG THE OBTAINED RANK OF FR-GPSO AND COMPARISON METHODS ON MICROARRAY DATA

Average Rank	Algorithms	Ranking	Algorithms	Ranking
1	FR_GPSO	4.7	FR_GPSO	3.9
2	PSO	3.3	HPSO-LS	2.4
3	ACO	3.2	PSO(4-2)	2.2
4	EA	2.1	BPSO	1.5
5	GA	1.7		
P-value	0.00119		0.00677	
Statistic	11.04 (4, 16)		6.65 (3,12)	

Table XII demonstrates the compared results of the Friedman test for classification accuracies between the proposed approach and comparison methods. According to the *p-value* (0.00119 and 0.00677) in the Friedman test, all of those approaches have a significant difference in five microarray datasets, as shown in Table VII. Overall, FR-GPSO achieves the highest classification accuracy, and PSO (HPSO_LS) is the second best in feature subset selection for classification problems. In two sets of experiments, both the ACO and GA methods had poor performance. Based on the results of statistical analysis of microarray data, the FR-GPSO algorithm shows the performance efficiency of feature selection compared to basic (containing GA, EA, ACO, and PSO) and state-of-the-art (BPSO, PSO (4-2), and HPSO-LS) wrapper-based approaches.

VI. CONCLUSION

The proposed FR-GPSO aims to improve the accuracy of classification considerably under various dimensions of the different datasets from different areas. In terms of classification accuracy and the number of features selected on the given benchmark datasets, the proposed two-phase approach designed with a new fitness function implements well over those of the original datasets. Additionally, the performance metrics and the dimension metrics are calculated and compared with the basic wrapper algorithms, GA, EA, ACO, and PSO. The results demonstrate the effectiveness of the proposed method for handling high-dimensional data classification.

Furthermore, the proposed algorithm was compared with the three existing methods such as BPSP, PSO (4-2), and HPSO-LS on real datasets. The experimental results were reported for the two aspects of the classification accuracy and the number of selected attributes. The results indicate that the proposed algorithm is highly competitive with those existing methods with high dimensional data. Moreover, the performance of the FR-GPSO approach is significantly improved among the competitors' methods according to the *p-value* of the Friedman Test.

In our future work, more test datasets from different domains will be considered to further evaluate the proposed algorithm with different classifiers and apply the local search strategy according to the relevancy scores.

ACKNOWLEDGMENT

I would like to extend special thanks to Dr. Win Mar Oo, Pro-Rector of the University of Computer Studies, Mandalay (UCSM), and Dr. Si Si Mar Win, Professor of the Faculty of Computer Science at the University of Computer Studies, Yangon (UCSY), for both of their continuous guidance, support, and suggestions.

REFERENCES

[1] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis, vol. 1, no. 3, pp. 131–156, 1997.
[2] P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy," Applied Soft Computing, vol. 43, pp. 117–130, Jun. 2016, doi: 10.1016/j.asoc.2016.01.044.
[3] S. Li, K. Zhang, Q. Chen, S. Wang, and S. Zhang, "Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and

Genetic Algorithm," IEEE Access, vol. 8, pp. 139512–139528, 2020, doi: 10.1109/ACCESS.2020.3012768.

[4] W. Chen, Y. Xu, Z. Yu, W. Cao, C. L. P. Chen, and G. Han, "Hybrid Dimensionality Reduction Forest with Pruning for High-Dimensional Data Classification," IEEE Access, vol. 8, pp. 40138–40150, 2020, doi: 10.1109/ACCESS.2020.2975905.

[5] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), Anchorage, AK, USA, 1998, pp. 69–73. doi: 10.1109/ICEC.1998.699146.

[6] H. Wang, R. Ke, J. Li, Y. An, K. Wang, and L. Yu, "A correlation-based binary particle swarm optimization method for feature selection in human activity recognition," International Journal of Distributed Sensor Networks, vol. 14, no. 4, p. 155014771877278, Apr. 2018.

[7] A. H. Alsaeedi, A. L. Albukhnefis, D. Al-Shammary, and M. Al-Asfoor, "Extended Particle Swarm Optimization (EPSO) for Feature Selection of High Dimensional Biomedical Data," p. 20.

[8] B. Tran, B. Xue, M. Zhang, and S. Nguyen, "Investigation on particle swarm optimization for feature selection on high-dimensional data: local search and selection bias," Connection Science, vol. 28, no. 3, pp. 270–294, Jul. 2016, doi: 10.1080/09540091.2016.1185392.

[9] L. Sun, X. Kong, J. Xu, Z. A Xue, R. Zhai, and S. Zhang, "A hybrid gene selection method based on ReliefF and ant colony optimization algorithm for tumor classification. Scientific reports, 9(1), 1-14, 2019.

[10] Wenping Ma, Xiaobo Zhou, Hao Zhu, Longwei Li, Licheng Jiao, "A two-stage hybrid ant colony optimization for high-dimensional feature selection", Pattern Recognition, vol. 116, 2021, 107933, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2021.107933.

[11] A. Bir-Jamel, S. M. Douiri, and S. Elbernoussi, "Gene selection via a new hybrid ant colony optimization algorithm for cancer classification in high-dimensional data". Computational and mathematical methods in medicine, 2019.

[12] B. Selvakumar, K. Muneeswaran, "Firefly algorithm based feature selection for network intrusion detection", Computers & Security, vol. 81, 2019, pp. 148-155, doi: 10.1016/j.cose.2018.11.005.

[13] H. Dong, Y. Pan, and J. Sun, "High Dimensional Feature Selection Method of Dual Gbest Based on PSO," in 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, United Kingdom, Jul. 2020, pp. 1–8. doi: 10.1109/CEC48606.2020.9185635.

[14] X. Song, Y. Zhang, Y. Guo, X. Sun, and Y. Wang, "Variable-size Cooperative Coevolutionary Particle Swarm Optimization for Feature Selection on High-dimensional Data," IEEE Trans. Evol. Computat., pp. 1–1, 2020, doi: 10.1109/TEVC.2020.2968743.

[15] B. Tran, B. Xue, and M. Zhang, "Adaptive multi-subswarm optimisation for feature selection on high-dimensional classification," in Proceedings of the Genetic and Evolutionary Computation Conference, Prague Czech Republic, Jul. 2019, pp. 481–489. doi: 10.1145/3321707.3321713.

[16] H. Motieghader, A. Najafi, B. Sadeghi, and A. Masoudi-Nejad, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata," Informatics in Medicine Unlocked, vol. 9, pp. 246–254, 2017, doi: 10.1016/j.imu.2017.10.004.

[17] B. Tran, M. Zhang, and B. Xue, "A PSO based hybrid feature selection algorithm for high-dimensional classification," in 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, Canada, Jul. 2016, pp. 3801–3808. doi: 10.1109/CEC.2016.7744271.

[18] F. Han, D. Tang, Y.-W.-T. Sun, Z. Cheng, J. Jiang, and Q.-W. Li, "A hybrid gene selection method based on gene scoring strategy and improved particle swarm optimization," BMC Bioinformatics, vol. 20, no. S8, p. 289, Jun. 2019, doi: 10.1186/s12859-019-2773-x.

[19] L. Sun, X. Kong, J. Xu, Z. Xue, R. Zhai, and S. Zhang, "A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification," Sci Rep, vol. 9, no. 1, p. 8978, Dec. 2019, doi: 10.1038/s41598-019-45223-x.

[20] Chen, Yiyuan, Wang, and Yufeng and Cao, "CCFS: A Confidence-based Cost-effective feature selection scheme for healthcare data classification," IEEE/ACM transactions on computational biology and bioinformatics, 2019, doi: DOI 10.1109/TCBB.2019.2903804.

[21] L. Brezo˘cnik, "Feature selection for classification using Particle Swarm Optimization," in IEEE EUROCON 2017-17th International Conference on Smart Technologies, Jul. 2017, pp. 966--971.

[22] M. Mafarja and N. R. Sabar, "Rank based binary particle swarm optimisation for feature selection in classification," in Proceedings of the 2nd International Conference on Future Networks and Distributed Systems - ICFNDS '18, Amman, Jordan, 2018, pp. 1–6. doi: 10.1145/3231053.3231072.

[23] Q.-L. Xiao, H. Zheng, and Q.-A. Yao, "Feature Selection for Cancer Classification Based on SRPSO Algorithm," dtetr, no. icicr, Aug. 2019, doi: 10.12783/dtetr/icicr2019/30576.

[24] N. Mallenahalli and T. H. Sarma, "A Tunable Particle Swarm Size Optimization Algorithm for Feature Selection," arXiv:1806.10551 [cs], Jun. 2018, Accessed: Jul. 04, 2020. [Online]. Available: http://arxiv.org/abs/1806.10551.

[25] G. C. Alekhya, D. Sruthi, J. Abhilasha, K. Vineetha, and V. L. Narayana, "Hybrid Feature Selection of Correlation Coefficient with PSO on Micro Array Gene Expression Data," vol. 11, no. 4, p. 15.

[26] Q. Wu, Z. Ma, J. Fan, G. Xu, and Y. Shen, "A Feature Selection Method Based on Hybrid Improved Binary Quantum Particle Swarm Optimization," IEEE Access, vol. 7, pp. 80588–80601, 2019, doi: 10.1109/ACCESS.2019.2919956.

[27] B. Ji, X. Lu, G. Sun, W. Zhang, J. Li, and Y. Xiao, "Bio-Inspired Feature Selection: An Improved Binary Particle Swarm Optimization Approach," IEEE Access, vol. 8, pp. 85989–86002, 2020, doi: 10.1109/ACCESS.2020.2992752.

[28] B. Wei, W. Zhang, X. Xia, Y. Zhang, F. Yu, and Z. Zhu, "Efficient Feature Selection Algorithm Based on Particle Swarm Optimization with Learning Memory," IEEE Access, vol. 7, pp. 166066–166078, 2019, doi: 10.1109/ACCESS.2019.2953298.

[29] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," Neurocomputing, vol. 173, pp. 346–354, Jan. 2016, doi: 10.1016/j.neucom.2014.12.123.

[30] A. Moraglio, C. D. Chio, and R. Poli. Geometric particle swarm optimization. In European Conference on Genetic Programming, pages 125–136, 2007.

[31] R. Kohavi and G. H. John, ``Wrappers for feature subset selection," Artif. Intell., vol. 97, nos. 12, pp. 273324, Dec. 1997

[32] *E. Frank, M. A. Hall, and I. H. Witten, The WEKA workbench. Morgan Kaufmann, 2016*.

[33] Zexuan Zhu, Y. S. Ong and M. Dash, "Markov Blanket-Embedded Genetic Algorithm for Gene Selection", Pattern Recognition, Vol. 49, No. 11, 3236-3248, 2007.

[34] T. Li, C. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression". Bioinformatics, 20 (2004) 2429-2437.

[35] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2017, [Online]. Available: http://archive.ics.uci.edu/ml.

[36] L. Y. Chuang, C. S. Yang, K. C. Wu, and C. H. Yang, "Gene selection and classification using taguchi chaotic binary particle swarm optimization," Expert Systems with Applications, vol. 38, no. 10, pp. 13 367–13 377, Sep. 2011.

[37] B. Xue, M. Zhang, and W. N. Browne, ``Particle swarm optimization for feature selection in classification: Novel initialisation and updating mechanisms," Appl. Soft Comput., vol. 18, pp. 261-276, May 2014.

[38] M. Friedman, "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance," Journal of the American Statistical Association, vol. 32, pp. 675-701, 1937.