2021 18th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE). Mexico City, Mexico. November 10-12, 2021

A Novel Hybrid Gene Selection Based on Random Forest Approach and Binary Dragonfly Algorithm

1st Sayed Pedram Haeri Boroujeni
Department of Artificial Intelligence Engineering
Istanbul Aydin University
Istanbul, Turkey
sayedpedramboroujeni@stu.aydin.edu.tr

2nd Elnaz Pashaei
Department of Software Engineering
Istanbul Aydin University
Istanbul, Turkey
elnazpashaei@aydin.edu.tr

***Abstract*—Microarrays dataset contains a huge number of genes and a few samples. This issue can lead to the curse of dimensionality in large datasets. To overcome this challenge, gene selection is a method used for identifying the independent genes and removing redundant or noisy ones from the dataset. This study proposes a novel hybrid approach based on the combination of Random Forest Ranking (RFR) and Binary Dragonfly Algorithm (BDA) to identify the significant genes. The proposed method comprises two steps. In the first step, RFR is employed to remove irrelevant genes and select the subsets of optimal genes. In the second step, BDA is applied to select the most informative genes that can lead to the accurate detection of cancer. The BDA optimizer is a recently proposed metaheuristic algorithm that utilizes Naïve Bayes (NB) classifier as an evaluator. In this paper, four microarray datasets are used to evaluate the performance of the proposed hybrid approach. Experimental results illustrate that the proposed work significantly outperforms existing meta-heuristic methods regarding classification accuracy and the optimal number of selected genes.**

Keywords—Gene selection, Random Forest Ranking, Binary Dragonfly Algorithm, Naïve Bayes classifier.

I. Introduction

Due to the advances in microarray technology, gene selection or feature selection is a pre-processing technique that has a significant effect on the performance of various machine learning tasks such as classification [1], bioinformatics [2], medical diagnosis [3], etc. Generally, the selection and classification of genes is a complicated task in microarray datasets, because of the fact that the datasets have thousands of genes and a limited number of instances. In high-dimensional data, irrelevant and redundant genes not only provide unnecessary information but also negatively affect the performance of learning algorithms [4]. To handle this issue, gene selection algorithms play an important role to select the most regulatory genes and eliminate the worthless genes from the original dataset. Gene selection techniques have several main advantages such as improving the classification accuracy, decreasing the number of selected genes, avoiding over-fitting, decreasing memory usage, and reducing the amount of time-consuming to analyze the whole dataset.

Feature selection methods are categorized into four sets including filter, wrapper, hybrid, and embedded methods [5]. Filter approaches are known as gene-ranking methods or open-loop schemes which evaluate each gene using its general statistical properties without using any learning algorithm [6]. Furthermore, the filter model chooses the genes that have high score values based on the ranking criterion function and they are not affected by the classifiers [7]. On the other side, wrapper approaches utilize learning algorithms (e.g. classification) to evaluate feature subsets to find the best optimal number of genes and enhance the classification performance [8]. In general, the performance of wrapper approaches is more efficient than filter approaches, and they provide better accuracy since they consider the relation between predictors and solutions [9]. However, wrapper methods have a more expensive computational level, so they are less efficient in comparison to the filter methods [10]. The next category is hybrid methods that combine the filter approaches with wrapper approaches to achieve great performance and solve the high-dimensional problems.

The main contribution of this paper is to propose a novel hybrid method in which the Binary Dragonfly Algorithm (BDA) and the Random Forest Ranking (RFR) are combined to select the appropriate set of genes. The wrapper gene selection approach utilizes the BDA as a search strategy and the Naïve Bayes (NB) classifier as an evaluator [11]. In all of the previous proposed works with the BDA optimizer, the K-Nearest Neighbors (KNN) classifier is used as an evaluator in a wrapper approach. So, this is a new idea to use the NB classifier in the wrapper approach part to reach high performance. Furthermore, RFR is used as a filter approach before using BDA optimization to choose the significant genes from the gene expression dataset and decrease the searching time of BDA [12]. At the beginning of this study, the efficiency of the RFR filter method is compared to the minimum-Redundancy-Maximum-Relevance (mRMR) and Relief-m methods. Next, the performance of the BDA optimizer compared to the obtained results from the filter approaches including mRMR, Relief-m, and RFR. Finally, the performance of the BDA is compared to the other state-of-the-art algorithms. To the best of our knowledge, this is the pioneer hybrid method in which the Random Forest (RF) is employed as gene ranking and after that, the BDA is applied as a gene selection for cancer classification.

The rest of this study has been structured as follows: Section II reviews the related work. In Section III, the details of the material and proposed methods are discussed. The experimental evaluation and analysis are illustrated in Section IV, Meanwhile, Section V outlines the conclusion.

II. Review of the Related Literature

In recent years, many studies focused on a combination of filter and wrapper methods. To this aim, some hybrid state-of-the-art methods are introduced: Fisher Score with novel Intelligent Dynamic Genetic Algorithm (IDGA) [13], mRMR

978-1-6654-0029-9/21/$31.00 ©2021 IEEE

with Artificial Bee Colony (ABC) [14], Information Gain (IG) with Genetic Algorithm (GA) [15], and RFR with IDGA [16]. The last study proposed a new hybrid model according to the IDGA and RF to diagnose valuable gene subsets for cancer classification. It used RF based on the IDGA algorithm not only in filtering the noisy and irrelevant genes but also in its fitness function.

In the available literature, the diversity of optimization algorithms for the optimal subset are numerous. Various kinds of gene selections are introduced such as Grey Wolf Optimizer (GWO) which is an algorithm [17] that has been applied for solving gene selection problems in [18]. The other method is an Ant Lion Optimizer (ALO) which is used as a wrapper gene selection approach that works based on the biological behavior of antlions in hunting and migration mechanisms [19]. The Particle Swarm Optimization (PSO) [20] algorithm is another method that is employed in gene selections purpose [21]. Finally, the recent population-based algorithm is a binary version of the Dragonfly Algorithm (DA) called BDA [12], which mimics the swarming behavior of idealized dragonflies.

In these decades, various kinds of filters such as Conditional Mutual Information Maximization (CMIM) [22], Joint Mutual Information (JMI) [23], and Relief-F [24] are introduced to rank the top genes from datasets. In recent works, the mRMR method which has the following characteristic including pre-processing, ranking, filtering, optimization, and classification is proposed. In this technique, each gene is estimated based on the information theory, and then they are ordered based on their ranking to reduce the number of data [25]. The mRMR filter has been used to eliminate the redundant and irrelevant genes from the dataset to increase the efficiency and the quality of optimization.

III. Material and Proposed Methods

A. Random Forest (RF)

RF is one of the most popular non-parametric tree-based ensemble classifiers because of various reasons. First and foremost, it is highly variable adaptable and suitable for various microarray classifications. Secondly, it is a useful method for analyzing "large p, small n" problems. The third advantage is that it can be applied in both simple and complicated classifications. Moreover, it can be also used for selecting and ranking the genes based on the important measures of genes. Besides, it can be employed for both binary and multi-class classifications. Finally, the necessity of parameter tuning for the RF is not mandatory and frequently the default parameters play an efficient role to achieve a high-performance result [26]. Consequently, these outstanding properties of RF make it a suitable method for prediction, classification, gene selection, cancer, and bioinformatics research [27]. According to the previous works, one of the major aspects of the RF method is gene ranking. RF works based on the following steps:

1) The tree is drawn using a bootstrap sample of the original data, so RF has often a collection of thousands of trees.

2) The tree is grown for each bootstrap dataset and variables are selected randomly by each node of the tree.

3) Information is collected from the trees to predict new data and classification.

4) Calculate the variable's permutation importance by using an Out-Of-Bag (OOB) error rate.

Firstly, the OOB calculates the prediction error which is known as non-premutation OOB error. In the next stage, when all the other samples remain steady, the values of the given variable are permuted randomly in the OOB which is known as permutation OOB error. As a result, the difference between the OOB error with permutation and non-permutation OOB error is the variable's permutation importance. If the variable has large permutation importance, it shows this variable is more important and predictive.

B. Binary Dragonfly Algorithm (BDA)

In this paper, BDA is used to solve the gene selection problems in gene expression datasets. The algorithm's outputs are restricted to binary forms (zero and one). A vector of zeros and ones is used to indicate the solution of the gene selection problem, where the zero elements represent that the corresponding gene is not selected and the one element means that the gene is selected. Furthermore, the DA is used for continuous optimization and the BDA is used for discrete optimization. The transfer function plays a significant role in the performance of BDA algorithm, which converts a continuous space to a discrete space. So, the transfer function is only used in BDA and this is the main difference between BDA and DA.

On the other hand, BDA has two important parts which are exploration (diversification: explore the search space) and exploitation (intensification: exploit the optimal solution). An appropriate balance should be chosen between the exploration and exploitation parts to achieve the high performance of BDA optimizer. BDA optimization works based on the dragonflies' life mechanism which is the hunting mechanism and the migration mechanism. The first part is known as a static swarm in which the dragonflies fly in small groups over a small area to find food sources. The second part is called a dynamic swarm in which the dragonflies fly in large groups along one direction that should be opposite the enemy's location.

Preparing a model for binary dragonflies needs five individual behaviors as follows. In the following equations, X variable shows the position of the current search agent, Xj shows the jth neighbor of the X search agent, and N is the neighborhood size [28]:

- Separation indicates that an individual should stay away from neighbors. Equation (1) is the mathematical model of separation behavior:

$$S_i = -\sum_{j=1}^{N} X - X_i \tag{1}$$

- Alignment represents that an individual should match its velocity with the other neighbors. Equation (2) is the mathematical model of alignment behavior and V_j indicates the velocity of the jth neighbor.

$$A_i = \frac{\sum_{j=1}^{N} V_j}{N} \tag{2}$$

- Cohesion represents the tendency of individuals to fly towards the neighboring. Equation (3) is the mathematical model of cohesion behavior:

$$C_i = \frac{\sum_{j=1}^{N} xj}{N} - X \tag{3}$$

- Attraction represents the tendency of individuals to fly towards the food source. Equation (4) is the mathematical model of the attraction behavior and F_{loc} indicates the position of food source.

$$F_i = F_{loc} - X \tag{4}$$

- Distraction refers to the tendency of individuals to fly away from an enemy. Equation (5) is the mathematical model of cohesion behavior and E_{loc} represents the enemy's position.

$$E_i = E_{loc} + X \tag{5}$$

In the next step, a position vector which is defined in (6) is used to solve optimization problems. Parameters s, w, a, c, f, and e represent the weights of the separation (S_i), alignment (A_i), cohesion (C_i), attraction towards the food source (F_i), and distraction from the enemy (E_i), respectively.

$$\Delta X_{t+1} = (sS_i + aA_i + cC_i + fF_i + eE_i) + wX_t \tag{6}$$

Generally, in a continuous search space (DA optimizer), the position of dragonflies is updated by adding the step vector to the previous position. However, in the binary search space (BDA optimizer) the following equations are used to solve binary optimization problems such as gene selection. The transfer function in (7) is employed to convert a continuous space into a binary one for the BDA. So, the transfer function generated the probability of changing the continuous positions to binary. Then, the result of the $T(v_d^i(t))$ is used to convert ith element of the position vector to 0 or 1 based on (9). Parameter r is a random number in [0,1].

$$T(v_d^i(t)) = \left| \frac{v_d^i(t)}{\sqrt{(v_d^i(t))^2 + 1}} \right| \tag{7}$$

$$X_{t+1} = \begin{cases} -X_t & r < T(v_d^i(t)) \\ X_t & r \geq T(v_d^i(t)) \end{cases} \tag{8}$$

As previously noted, feature selection is a multi-objective task where the classification accuracy should be maximized and the number of selected genes should be minimized. Thus, the represented fitness function in (9), calculates the classification accuracy and the number of selected genes. Parameter $\gamma_R(D)$ represents the classification error rate, |C| is the number of selected genes, and |N| is the total number of genes. Also, α and β are two parameters which represent the importance of classification quality and subset length, that α is in the [0,1] interval and β = (1 − α).

$$\downarrow Fitness = \alpha\gamma_R(D) + \beta\frac{|C|}{|N|} \tag{9}$$

The pseudo-code of BDA is shown in Algorithm I. The algorithm repeats the following steps in each iteration until the best solution is obtained. First of all, each dragonfly is evaluated using the specified objective function. After that, the algorithm updates the main coefficients. Thirdly, the parameters S_i , A_i , C_i , F_i , and E_i are computed using Equations (1) – (5). Eventually, the step vectors, position vectors, and transfer function are updated using Equations (6), (7), and (8).

Algorithm I. Pseudo-code of BDA

Set the initial parameters of the BDA algorithm
Initialize the population X_i(i = 1, 2, …, n)
Initialize ΔX_i(i = 1, 2, …, n)
While (termination criteria are not met)
    Evaluate the fitness of dragonflies
    Update (F) and (E)
    Update the swarming factors (w, s, a, c, f, and e)
    Calculate S, A, C, F, and E using Eqs. (1) to (5)
    Update step vectors using Eq. (6)
    Calculate $T(v_d^i(t)$ using Eq. (7)
    Update X_{t+1} using Eq. (8)
end while
Return the optimum search agent

C. Proposed Method (RFR-BDA)

In this article, a novel hybrid method is introduced where the Random Forest Ranking (RFR) with the Binary Dragonfly Algorithm (BDA) are combined to enhance the classification accuracy and also decrease the number of selected genes. The BDA is used as a wrapper gene selection approach and RFR is used as a filter approach. In the RFR-BDA technique, there are two main parts which are outlined as follows. In the first part of the work, the RFR is employed to eliminate irrelevant and redundant genes from the microarray dataset. So, the RFR ranks all of the genes and the best N-top genes are selected. The output of the first part is a new gene subset which is produced by the RFR for the next part. Meanwhile, in a new subset, the huge number of genes are numerously reduced and the most informative genes are found. In the second part of the study, BDA is applied to the reduced data subset as a wrapper approach in which the search strategy is BDA and the evaluator for the selected genes is the NB classifier. Hence, the best genes are found by the BDA optimizer and the classification accuracy is evaluated by using NB Classifier. Consequently, after two feature selection parts, the significant and relevant genes are selected from the microarray dataset meanwhile, the classification performance is improved as a result of the proposed method. The flow chart of the proposed approach (RFR-BDA) is illustrated in Figure I.

IV. Experimental Results and Discussion

In the proposed work, each experiment is repeated 20 independent times with random seed which leads to obtaining more reliable and confident results. Moreover, all of the algorithms are implemented using MATLAB 2018a with an Intel Core i3 processor, 2.2 GHz CPU, and 4 GB of RAM.

A. Datasets and Parameters Setting

Table I summarizes the details about four various kinds of popular gene microarray datasets namely Leukaemia-1, Leukaemia-2, SRBCT, and Lung-cancer. These datasets are different from each other in the number of instances, features, and classes. Out of these four datasets, two datasets are binary-class and two datasets are multi-class. In addition, we employed the K-Fold cross-validation technique (with K=10) during the performance evaluation of the RFR-BDA approach to obtain more confident results, avoid bias problems, and assess the desired prediction accuracy. In 10-fold cross-

validation, the testing sub-data is only used for assessing the final results and the training sub-data is used for making the model. Table II demonstrates the details of the optimal parameters for the proposed approach (BDA) and all the considered algorithms (GA, PSO, ACO, and DE).

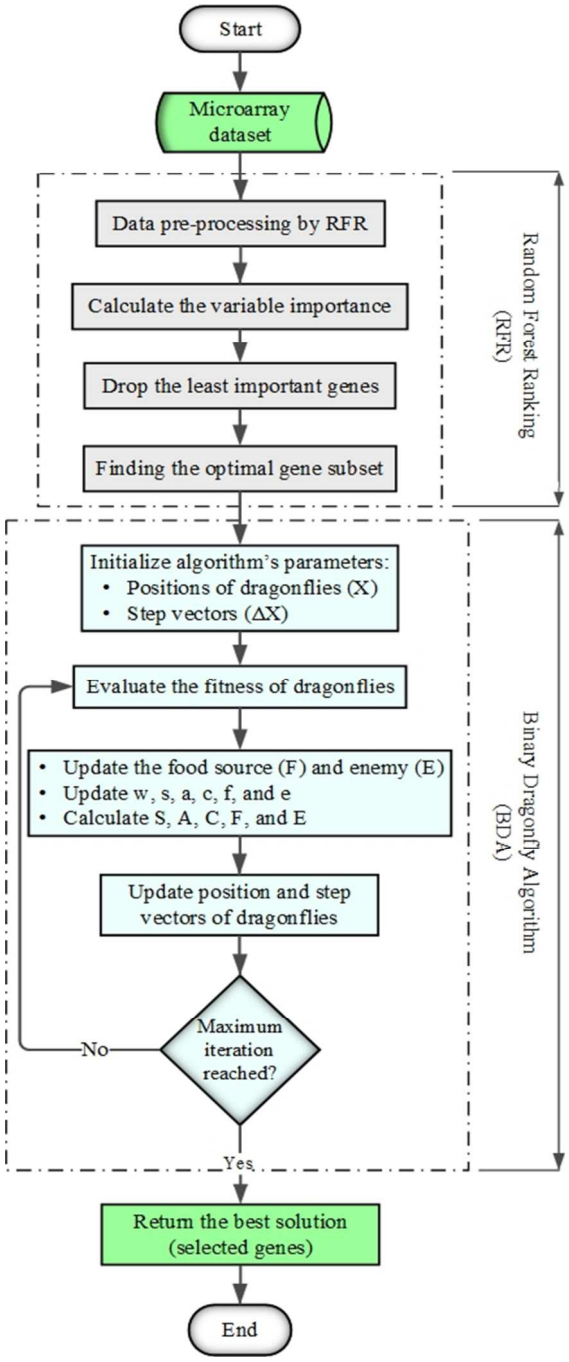

FIGURE I. THE FLOW CHART OF THE RFR-BDA FOR FEATURE SELECTION

TABLE I. MICROARRAY DATASETS DESCRIPTION

No.	Datasets Detail			
	Dataset	*Instances*	*Genes*	*Classes*
1	Leukaemia-1	72	7129	2
2	Leukaemia-2	72	11225	3
3	SRBCT	83	2038	4
4	Lung-Cancer	203	12600	2

TABLE II. PARAMETERS SETTING FOR THE PROPOSED METHOD

Algorithm	Parameters Detail		
	Abbreviation	*Parameter*	*Value*
BDA	α β K	In the fitness function In the fitness function In the KNN classifier	0.99 0.01 5
GA	Pc Pm Mu	Crossover percentage Mutation percentage Mutation rate	0.8 0.2 0.02
PSO	C_1 C_2 W	Constant value Constant value Intra weight	2.1 2.1 0.7
ACO	α β rho	Pheromone weight Heuristic Weight Evaporation rate	1 1 0.05
DE	B_{Min} B_{Max} PCr	Lower bound Upper bound Crossover probability	0.2 0.8 0.2
Common Parameters	N Iter Run	Population size No. of iterations No. of runs	15 50 20

It should be mentioned that the α and β parameters represent the weight of the classification accuracy and genes selection rate, respectively. In this study, various values of α and β are examined to obtain the optimal classification accuracy with the minimum number of genes. To this aim, we have been chosen the Leukaemia-1 dataset for the following experiment because of its sensitivity to the initial parameters more than the other datasets. Table III represents the classification performances of the BDA algorithm after selecting different values of α and β. From Table III, it can be claimed that the α=0.99 and β=0.01 achieve the highest classification accuracy with the minimum number of selected genes. Therefore, we have been selected the most suitable value of α and β (0.99 and 0.01, respectively) for our fitness function.

TABLE III. THE ACCURACY AND SELECTION RATIO OF THE BDA ALGORITHM FOR THE LEUKAEMIA-1 DATASET

No.	The Influence of the α and β Parameters			
	α	*β*	*Accuracy (%)*	*Selection Ratio*
1	0.5	0.5	95.85	0.465
2	0.6	0.4	96.33	0.461
3	0.7	0.3	97.51	0.459
4	0.8	0.2	94.70	0.466
5	0.9	0.1	95.22	0.463
6	0.99	0.01	100	0.457

B. Evaluation Criteria

In the proposed paper, four different classifiers including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Naïve Bayes (NB) are used to evaluate the microarray datasets. These classifiers are evaluated by five kinds of measurements: Accuracy, Sensitivity, Specificity, Matthews Correlation Coefficient (MCC), and F-measure [29]. Equation (10) to Equation (14) defined the formula for each performance measure.

Parameters T_P, T_N, F_P and F_N show the True Positive, True Negative, False Positive, and False Negative respectively. Finally, the performance of each dataset is assessed based on the confusion matrix.

$$Accuracy = \frac{T_P+T_N}{T_P+T_N+F_P+F_N} \quad (10)$$

$$Sensitivity = \frac{T_P}{T_P+F_N} \quad (11)$$

$$Specificity = \frac{T_N}{T_N+F_P} \quad (12)$$

$$F-measure = \frac{2*T_N}{2*T_P+F_N+F_P} \quad (13)$$

$$MCC = \frac{(T_P*T_N-F_N*F_P)}{\sqrt{(T_P+F_P)+(T_P+F_N)+(T_N+F_P)+(T_N+F_N)}} \quad (14)$$

Table IV shows the numerical results of four classifiers on every single microarray dataset. According to the mentioned table, the average classification performance is achieved before applying the proposed filter approach (RFR) and optimization algorithm (BDA). To be more comprehensible and comparable, the maximum accuracy for each dataset is bolded in the provided table.

TABLE IV. AVERAGE CLASSIFICATION RESULTS (%) OF FOUR WELL-KNOWN CLASSIFIERS ON FOUR MICROARRAY DATASETS

Dataset	Measures	Classifiers			
		SVM	***RF***	***KNN***	***NB***
Leukaemia-1	Accuracy	**95.83**	91.66	87.50	94.53
	Sensitivity	95.80	88.67	88.00	93.82
	Specificity	95.80	90.32	86.36	93.60
	F-Measure	95.80	91.30	87.30	93.12
	MCC	90.80	82.10	72.00	94.17
Leukaemia-2	Accuracy	89.80	**93.22**	84.72	92.36
	Sensitivity	87.63	92.20	84.70	91.25
	Specificity	88.18	92.63	86.40	92.07
	F-Measure	86.93	93.30	85.10	93.47
	MCC	91.32	93.60	87.35	94.32
SRBCT	Accuracy	97.59	**98.79**	84.33	**98.79**
	Sensitivity	93.54	96.66	86.20	98.80
	Specificity	94.62	97.52	90.00	96.15
	F-Measure	97.60	98.80	84.20	98.80
	MCC	96.70	98.30	78.60	98.20
Lung-Cancer	Accuracy	88.47	89.65	87.43	**90.20**
	Sensitivity	87.35	89.70	86.29	87.47
	Specificity	88.63	88.85	85.72	88.31
	F-Measure	89.05	90.13	86.03	91.33
	MCC	91.60	87.60	92.50	91.89

C. Results and Analysis

In this experimental part of the study, the performance of the proposed method is calculated and presented. The RFR-BDA technique is used to decrease the amount of redundant and irrelevant genes from datasets. In general, RFR ranks all of the genes from microarray datasets, and the best n-top genes with the highest score are selected. After the ranking part, the BDA applies to the n-top selected genes to find the most informative and relevant subset of genes. To be more specific, n-top significant genes are chosen from four datasets by using the RFR method, and the number of n-top genes is adjusted to 5, 50, 100, 250, and 500. After that, all of the selected n-genes are evaluated with four different classifiers in order to become more comparable with the previous results.

Table V demonstrates the average performance of the RFR on four gene expression datasets. Meanwhile, the parameters of the classification algorithms are adjusted as same as the previous parts of the experiment. To provide more confident results, each technique is run twenty times. Finally, the maximum accuracy achieved by the classifiers is bolded in Table V. It is clear that the NB classifier has achieved higher classification accuracy compared to the other classifiers including SVM, RF, and KNN in all of the datasets. The information in Table VI illustrates the comparison between the performance of the proposed method (RFR) and the other three approaches including Relief-f, mRMR, and Conditional Mutual Information Maximization (CMIM) for the top 50-genes. As mentioned before, due to the good performance of the NB classifier in the previous parts, only the results of the NB are compared to the other approaches.

TABLE V. AVERAGE CLASSIFICATION ACCURACY (%) WITH N-TOP SELECTED GENES USING RANDOM FOREST RANKING FILTER

Dataset	Random Forest Ranking Accuracy				
	#Top	***SVM***	***RF***	***KNN***	***NB***
Leukaemia-1	5	94.44	95.83	93.05	97.22
	50	97.22	**98.61**	95.83	**98.61**
	100	97.22	97.22	97.22	97.22
	250	97.18	**98.61**	97.22	**98.61**
	500	97.13	97.22	**98.61**	**98.61**
Leukaemia-2	5	90.27	88.88	90.27	87.50
	50	97.19	95.83	94.44	**98.59**
	100	95.83	95.83	95.83	93.05
	250	97.19	94.44	**98.59**	95.83
	500	**98.59**	95.83	**98.59**	97.19
SRBCT	5	91.12	91.15	86.67	92.46
	50	96.13	**100**	96.66	**100**
	100	96.16	**100**	**100**	**100**
	250	97.15	98.10	98.75	**100**
	500	**100**	**100**	**100**	**100**
Lung-Cancer	5	89.15	93.12	89.18	86.21
	50	93.10	95.08	93.63	**95.61**
	100	92.80	93.09	94.61	95.16
	250	92.09	94.10	94.11	95.13
	500	93.11	94.95	93.23	95.46

The results prove the strength of our proposed filter approach in the ranking part. It is obvious that the RFR technique provides better accuracy in comparison to the other methods in all of the four microarray datasets. So, the RFR method is superior to the other gene selection approach on 50 gene subsets. It should be mentioned that we had to compare all of the filter methods with 50 genes because of the limitations of previous works.

TABLE VI. COMPARISON THE ACCURACY (%) OF THE RFR METHOD WITH OTHER FILTERS

Dataset	Different Methods			
	Relief-f Filter (50 genes)	*mRMR Filer (50 genes)*	*Ref. Paper [40] (50 genes)*	*Proposed RFR (50 genes)*
Leukaemia-1	73.39	83.12	96.01	**98.61**
Leukaemia-2	80.95	92.61	96.01	**98.59**
SRBCT	83.15	93.07	99.54	**100**
Lung-Cancer	85.73	91.13	95.01	**95.61**

In the next part of the experiment, BDA is employed to choose the best optimal genes from four datasets. It is clear that the KNN and NB classifiers show their good performances regarding the accuracy results on almost all datasets in the last part. Therefore, the final results of BDA are obtained by using KNN and NB classifiers, meanwhile, the SVM and RF classifiers are ignored for the next step to avoid repetitious results. BDA is used immediately after the ranking part to select the best genes from the dataset with the number of 5, 10, 25, 100, 250, and 500 genes. The information in Table VII shows the accuracy (Acc) of each dataset and the optimal number of selected genes. It should be noted that the BDA is executed 20 times, then the results are obtained.

TABLE VII. AVERAGE CLASSIFICATION ACCURACY (%) OF BDA WITH STANDARD DEVIATION (STD) AND THE NUMBER OF SELECTED GENES

Dataset	Results of The Proposed Wrapper Approach				
	#Top RFR Gene	*KNN*		*NB*	
		#Acc ± STD	*#No. Gene*	*#Acc ± STD*	*#No. Gene*
Leukaemia-1	5	98.75±0.09	3	97.32±0.10	2
	10	98.50±0.23	5	**100.00**	**4**
	25	100.00	11	100.00	12
	100	100.00	49	100.00	49
	250	100.00	102	100.00	79
	500	100.00	116	100.00	252
Leukaemia-2	5	93.15±0.13	4	90.53±0.14	2
	10	98.25±0.17	7	98.75±0.9	5
	25	98.57±0.09	13	**100**	**9**
	100	100.00	22	98.35±0.10	33
	250	97.50±0.07	127	98.75±0.08	93
	500	100.00	219	98.75±0.06	260
SRBCT	5	89.54±0.11	3	92.87±0.14	3
	10	98.70±0.10	7	99.20±0.18	6
	25	98.90±0.23	13	**99.88±0.14**	**11**
	100	100.00	28	100.00	25
	250	100.00	139	100.00	119
	500	100.00	245	100.00	248
Lung-Cancer	5	90.61±0.05	4	92.18±0.04	3
	10	96.70±0.12	6	**97.02±0.03**	**5**
	25	96.52±0.07	18	96.11±0.06	13
	100	96.61±0.06	22	97.04±0.07	21
	250	96.66±0.04	43	96.64±0.05	72
	500	97.52±0.08	135	97.54±0.06	92

D. Comparison of the proposed work with other methods

In recent years, various wrapper approaches have been introduced for feature selection tasks. This section compares the results of the proposed method with different wrapper approaches in terms of classification accuracy and the number of selected genes. In order to assess the performance of the proposed work, four gene selection methods including GA, PSO, ACO, and DE are used for comparison purposes. Table VIII demonstrates the performance of the proposed work with the other gene collection methods. According to the results, it is clear that the RFR-BDA achieved 100% accuracy in Leukaemia-1 and around 99.50% accuracy in the other three datasets. Meanwhile, the minimum number of genes are selected through the proposed method which is very better than the other algorithms.

E. Evaluate the proposed method with past literature

In recent years, several studies focusing on improving the performance of microarray datasets were proposed to increase the quality of gene subsets through the different wrapper approaches. However, existing wrapper approaches are faced with some limitations such as not having very high classification accuracy or a high number of selected genes. Accordingly, potential gene selection techniques are needed to select the optimal gene subsets with higher quality. The proposed method, not only tried to improve the performances of classification by proposing a new hybrid approach but also aimed to select the minimum number of genes from datasets.

The result of the proposed method is calculated in two phases. The first phase is RFR that the best genes are selected from microarray datasets. The second phase is where the BDA optimizer selected the best optimal genes to form the new subset. After these two phases, the classification accuracy is calculated. Table IX shows the comparison between the results of the proposed method and those of other literature. The first column represents the methods mentioned in the previous literature and compared them to the proposed method. The remaining columns show the accuracy percentages and the number of selected genes for each microarray datasets. Notice that the symbol "∗" means that no information is available. The results prove that the strength of the proposed method is higher than all the other methods in all of the datasets.

Statistical results prove that the strength of the proposed paper is very competitive. The performance of the proposed method is better than that of other methods in all of the microarray datasets except the Lung-cancer dataset in which the accuracy is two percent lower than the TLBOGSA method. However, our method selected 5 genes in comparison to the TLBOGGSA which selected 13 genes, so this is a noticeable matter that although the classification accuracy a little declined, a small number of significant genes are selected.

F. Statistical Analysis

There are several statistical methods for comparing the performance of various feature selection approaches. In this study, the Friedman test [36] has been used to detect the notable differences between the results of different methods. It works according to the principle of the null hypothesis (H_0), where the rejection of H_0 is regarded as the algorithm significantly outperforms other existing algorithms.

TABLE VIII. THE COMPARATIVE RESULTS OF THE PROPOSED METHOD AND OTHER WRAPPER APPROACHES REAGRDING THE ACCURACY RESULTS (%) AND THE NUMBER OF SELECTED GENES

Dataset	Performance	Results of The Different Wrapper Approaches				
		GA	***PSO***	***ACO***	***DE***	***Proposed Method***
Leukaemia-1	#Acc (%) ± STD #No. Gene	97.84 ± 0.04 (19)	98.23 ± 1.81 (21)	90.87 ± 1.56 (27)	91.35 ± 1.87 (18)	100.00 (4)
Leukaemia-2	#Acc (%) ± STD #No. Gene	94.15 ± 2.51 (16)	87.23 ± 2.89 (21)	86.32 ± 2.74 (25)	87.35 ± 2.57 (19)	99.51 ± 0.07 (9)
SRBCT	#Acc (%) ± STD #No. Gene	95.17 ± 1.52 (17)	94.22 ± 2.07 (19)	90.27 ± 2.51 (21)	94.36 ± 2.21 (23)	99.88 ± 0.08 (11)
Lung-Cancer	#Acc (%) ± STD #No. Gene	95.61 ± 0.53 (16)	97.05 ± 0.87 (19)	88.46 ± 1.31 (23)	91.25 ± 0.92 (18)	99.52 ± 0.03 (5)

TABLE IX. COMPARISON BETWEEN THE PROPOSED APPROACH AND OTHER METHODS REGARDING THE AVERAGE CLASSIFICATION ACCURACY (%) AND THE AVERAGE NUMBER OF SELECTED GENES

Dataset	Performance	Results of The Different Methods							
		PSO dICA [31]	***BFO [32]***	***DRFO-CGS [33]***	***RFR-PSO [34]***	***IWSSr [35]***	***TLBOGSA [30]***	***Proposed Method RFR-BDA***	***Best of The Proposed Method***
Leukaemia-1	#Accuracy #No. Gene	97.00 (72)	96.19 (23)	91.18 (13)	* *	97.10 (7)	94.15 (16)	100.00 (4)	100.00 (2)
Leukaemia-2	#Accuracy #No. Gene	83.33 *	* *	94.12 *	96.28 (57)	97.30 (6)	98.84 (12)	99.51 (9)	99.59 (6)
SRBCT	#Accuracy #No. Gene	* *	97.50 (35)	* *	94.25 (30)	92.30 (13)	99.17 (11)	99.88 (11)	100.00 (9)
Lung-Cancer	#Accuracy #No. Gene	97.95 (25)	93.11 (39)	98.66 (17)	93.00 (65)	* *	99.61 (13)	99.52 (5)	99.57 (5)

In this section, different algorithms are ranked based on their accuracy obtained from microarray datasets (see Table VIII). After that, Equation (15) calculates the Average Rank (AR) of each method. The approach with the best performance receives the lowest rank, while the approach with the worst performance receives the highest rank. Table X represents the rank of each algorithm according to Friedman's statistic.

$$AR = \frac{\text{The sum of the algorithm's ranks for each dataset}}{\text{Total number of datasets}} \quad (15)$$

TABLE X. THE RANKING OF FIVE FEATURE SELECTION ALGORITHMS BASED ON THE ACCURACY RESULTS

Dataset	Ranking of Different Algorithms				
	GA	***PSO***	***ACO***	***DE***	***Proposed Method***
Leukaemia-1	3	2	5	4	1
Leukaemia-2	2	4	5	3	1
SRBCT	2	4	5	3	1
Lung-Cancer	3	2	5	4	1
Average Rank	**2.5**	**3**	**5**	**3.5**	**1**

The results in Table X illustrate that the proposed method is ranked first among all of the approaches. Friedman test is performed to compare the performance of RFR-BDA with other methods. Assume that the number of datasets is N and the number of algorithms is M. The level of significance in our case is alpha = 0.05 and the degree of freedom (df) is in the interval of [4, 12], where the lower bound is (M-1) and the upper bound is (M-1)×(N-1). Consequently, the calculated Friedman statistic is 13.6 and the corresponding p-value at 0.05 level of significance is 0.0032. Therefore, the null hypothesis is rejected which means there is a significant difference between all the considered algorithms.

V. CONCLUSION

Because of the high dimension of gene expression datasets and the limited number of samples, determining an essential subset of cancer classification is a challenging task. To address this issue, this paper proposed a new hybrid approach that the BDA as a wrapper method and the RFR as a filter method are combined for gene selection. Meanwhile, four different classifiers are used for prediction purposes based on the selected gene subsets. Furthermore, this is the first time that NB classifier is used as an evaluator in the BDA approach because it could overcome the other classifiers in all of the experimental parts. The main aim of this work was to identify the minimum number of genes that could achieve the highest accuracy instead of using all genes in the microarray dataset.

In this study, four various kinds of microarray datasets are used to evaluate the proposed method and the results are compared to the eight recent approaches. Eventually, the proposed method not only improves the performance of classification accuracy but also is able to select significant and informative genes from the datasets. The experimental results demonstrate the superior performance of this hybrid approach in all of the datasets. The accuracy was around 99.70% in three datasets including Leukaemia-2, SRBCT, and Lung-Cancer, whereas in the Leukaemia-1 dataset the accuracy was 100% with the least number of genes (4 genes) that ever have achieved by any approaches. Consequently, the minimum number of selected genes with maximum accuracy proved the high performances of our proposed approach.

REFERENCES

[1] U. Qamar and M.A. Reza, "Classification is the process of group the objects and entities on the basis of the available information," in Data Science Concepts and Techniques with Applications, pp. 978-981, 2020.

[2] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics," in Bioinformatics, pp. 2507-2517, 2007.

[3] S. Shilaskara, A. Ghatolb, "Feature selection for medical diagnosis: evaluation for cardiovascular diseases," Expert Systems with Applications 40, pp. 4146-4153, 2013.

[4] N. Almugren and H. Alshamlan, "A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification," IEEE Access, vol. 7, pp. 78533-78548, 2019.

[5] N.K. Verma and A. Salour, "Feature Selection," in book Intelligent Condition Based Monitoring, 2020.

[6] E. Pashaei and N. Aydin, "Markovian encoding models in human splice site recognition using SVM," Computational Biology and Chemistry, vol. 73, pp. 159-170, 2018.

[7] J. Wang, J.-M. Wei, Z. Yang, S.-Q. Wang, "Feature selection by maximizing independent classification information," IEEE Trans. Knowl. Data Eng, pp. 828–841, 2017.

[8] H. M. Zawbaa, E. Emary, C. Grosan, and V. Snasel, "Largedimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach," Swarm and Evolutionary Computation, vol. 42, pp. 29-42, 2018.

[9] N. D. Cilia, C. D. Stefano, F. Fontanella, S. Raimondo, and A. S. d. Freca, "An experimental comparison of feature-selection and classification methods for microarray datasets," Information, 2019.

[10] G.Wu, R. Mallipeddi and P.N. Suganthan, "Ensemble strategies for population-based optimization algorithms–a survey," Swarm Evol. Comput. 44, pp. 695–711, 2019.

[11] S. Mirjalili, "Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," Neural Comput. Appl. 27, pp. 1053–1073, 2016.

[12] M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, S. Mirjalili, "Binary dragonfly algorithm for feature selection," IEEE, New Trends in Computing Sciences (ICTCS), International Conference, pp. 12-17, 2017.

[13] M. H. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," Genomics, vol. 109 no. 2, pp. 91-107, 2017.

[14] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling," BioMed Research International, vol. 2015, pp. 15, 2015.

[15] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," Applied Soft Computing, vol. 50, pp. 124-134, 2017.

[16] E. Pashaei, "2Gene Selection using Intelligent Dynamic Genetic Algorithm and Random Forest", pp. 470-474, 2019.

[17] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," Advances in Engineering Software, vol. 69, pp. 46-61, 2014.

[18] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," Neurocomputing, vol. 172, pp. 371-381, 2016.

[19] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary ant lion approaches for feature selection," Neurocomputing, 2016.

[20] J. Kennedy and R. Eberhart, "Particle swarm optimization," presented at the Neural Networks, 1995. Proceedings., IEEE International Conference on, 1995.

[21] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature election in classification: Novel initialization and updating mechanisms," Applied Software Computing, vol. 18, pp. 261-276, 2014.

[22] S. Anbuchelian, Chitra S and B. Madhusudhanan, "Feature extraction using CMIM for sentiment analysis," International Journal of Advanced Intelligence Paradigms, 2019.

[23] M. Bennasar, Y. Hicks and R. Setchi, "Feature selection using Joint Mutual Information Maximisation," Expert Systems with Applications, Volume 42, Issue 22, pp. 8520-8532, 2015.

[24] A. Arauzo-Azofra, J.M. Benitez, J.L. Castro, "A feature set measure based on relief," in Proceedings of the Fifth International Conference on Recent Advances in Soft Computing, pp. 104–109, 2014.

[25] H.-C. Wu, X.-G. Wei, S.-C. Chan, "Novel consensus gene selection criteria for distributed GPU partial least squares-based gene microarray analysis in diffused large b cell lymphoma (dlbcl) and related findings," IEEE ACM Trans. Comput. Biol. Bioinf 15, pp. 2039–2052, 2017.

[26] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," Genomics, vol. 99, no. 6, pp. 323-329, 2012.

[27] E. Pashaei, M. Ozen, and N. Aydin, "Splice site identification in the human genome using random forest," Health and Technology, vol. 7, no. 1, pp. 141–152, 2017.

[28] C.W. Reynolds, Flocks, herds and schools, "a distributed behavioral model," ACM SIGGRAPH Comput. Graph. 21, pp. 25–34, 1987.

[29] A.K. Santra and C.J. Christy, "Genetic algorithm and confusion matrix for document clustering," IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, 2012.

[30] A.K.Shukla, P.Singh, and M.Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," Swarm and Evolutionary Computation, 2020.

[31] M. Mollaee, M.H. Moattar, "A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification," Biocybern. Biomed. Eng. 36, pp. 521–529, 2016.

[32] H. Wang, X. Jing, B. Niu, "A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data," Knowl. Base Syst. 126, pp. 8–19, 2017.

[33] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, "Distributed feature selection: an application to microarray data classification," Appl. Soft Comput. 30, pp. 136–15, 2015.

[34] E. Pashaei, M.Ozen, N.Aydin, "A novel gene selection algorithm for cancer identification based on random forest and particle swarm optimization," IEEE Conference on Bioinformatics, 2015.

[35] A. Wang, N. An, G. Chen, L. Li, G. Alterovitz, "Accelerating wrapper-based feature selection with K-nearest-neighbor," Knowledge-Based Systems, pp. 81-91, 2015.

[36] Conover WJ and Iman RL, "Rank transformations as a bridge between parametric and nonparametric statistics," The American Statistician, pp. 124-129, 1981.