RESEARCH ARTICLE

# Genetic Clustering Algorithm-Based Feature Selection and Divergent Random Forest for Multiclass Cancer Classification Using Gene Expression Data

L. Senbagamalar[1] · S. Logeswari[2]

Received: 21 October 2023 / Accepted: 15 January 2024

**Abstract**
Computational identification and classification of clinical disorders gather major importance due to the effective improvement of machine learning methodologies. Cancer identification and classification are essential clinical areas to address, where accurate classification for multiple types of cancer is still in a progressive stage. In this article, we propose a multiclass cancer classification model that categorizes the five different types of cancers using gene expression data. To perform efficient analysis of the available clinical data, we propose feature selection and classification methods. We propose a genetic clustering algorithm (GCA) for optimal feature selection from the RNA-gene expression data, consisting of 801 samples belonging to the five major classes of cancer. The proposed feature selection method reduces the 1621 gene expressions into a cluster of 21 features. The optimum feature set acts as input data to the proposed divergent random forest. Based on the features computed, the proposed classifier categorizes the data samples into 5 different classes of cancers, including breast cancer, colon cancer, kidney cancer, lung cancer, and prostate cancer. The proposed divergent random forest provided performance improvisation in terms of accuracy with 95.21%, specificity with 93%, and sensitivity with 94.29% which outperformed all the other existing multiclass classification algorithms.

**Keywords** Feature selection · Genetic cluster algorithm · Mutation · Divergent random forest · Multiclass classification

## 1 Introduction

A systemized analysis of gene expression data has become a gateway for computationally diagnosing different types of cancer. The vast number of gene expression data requires computational methodologies to perform in-depth analysis of the available genetic information. This can be satisfied by the artificial intelligence paradigms and their subfield machine learning that identifies the underlying pattern from the available data [1]. The current research works are carried out in identifying the gene expression that belongs to a specific type of cancer and the other normal data which comes under the category of the binary classification problem [2]. This binary classification using machine learning models had become an effective diagnosis tool during the continuous monitoring phase carried out during clinical examinations [3]. The limitation that can be understood after going through the working procedure of the binary classification problem, is the gene expression data are collected together very vast in numbers where there is not any constraint that the samples belong to only two classes i.e., a single type of cancer and the normal data samples. In the collected gene data, there might be several cancerous gene data that belong to more than one type of cancer and the asymptotic gene data [4]. This constrained event makes a pressing need for the development of a multiclass classification model with a patterned feature selection and an effective classification system. This type of model requires a benchmark dataset for the effective analysis of multiple classification systems. The multi-class classification system regarding cancer classification itself can be interpreted in two different ways. The initial approach is to consider a specific type of cancer and classify the multiple levels based on the severity of the

✉ L. Senbagamalar
s.malarresearch@gmail.com

S. Logeswari
logeswari.s@kce.ac.in

[1] Anna University, Chennai, Tamilnadu, India

[2] Karpagam College of Engineering, Coimbatore, Tamilnadu, India

Published online: 05 February 2024

cancer. Another approach is to take different types of cancerous data samples that are labeled one and categorize each data sample with the specific type using a multiclass classification system. In this article, we considered the second approach where we utilized a benchmark RNA-gene expression data that consists of samples that belong to five different types of cancers including breast, kidney, colon, lung, and prostate cancers. 801 data samples consist of 16,383 genes as a feature that represents the data samples.

Figure 1 represents the sample box plot of five data samples and ten gene expression data. From the above plot, we can interpret that each of the data has an elevation over the sample pane where the data with high relevance has domination towards the plot when compared with other normal data samples. The initial stage of data analysis is to extract valuable information from the available data samples. This is possible through machine learning algorithms and optimization strategies. Earlier binary and multiclass classification methodologies had utilized machine learning approaches as well as some of the methods adopted machine learning as well as optimization techniques as an ensemble approach and the data analysis had been made. This research article focuses on three dimensions to perform multiclass classification of cancer types from the RNA data.

- Optimal selection of features for dimensionality and complexity reduction of the RNA data sample is done using the proposed genetic cluster algorithm (GCA).
- A new feature space is framed based on the cancer types using the feature selection process.
- A multiclass classification for the available data samples is done through the proposed probabilistic divergent random forest.

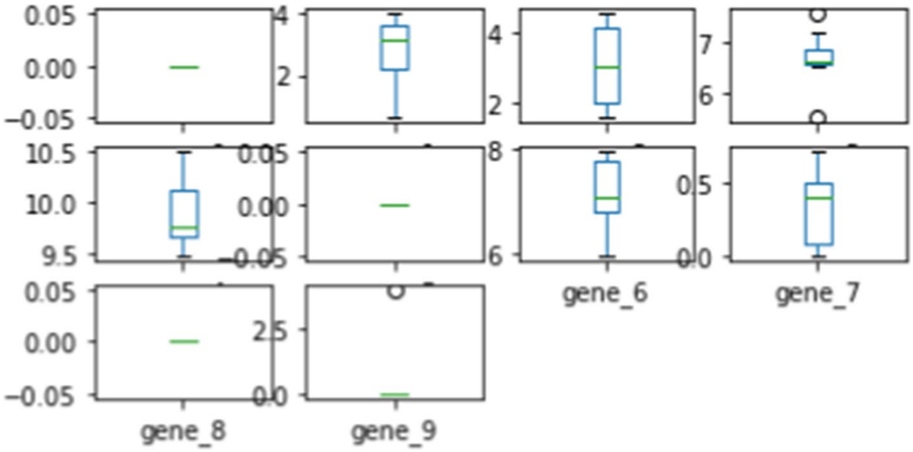
The major contribution of this research towards the clinical community is to identify the different types of cancer by extracting the gene expressions in clinical manner. Once gene expressions are obtained, then the computational methods can be completely utilized for breaking down each expression into data and to identify the underlying information. The flow of the article continues with Sect. 2 which deals with various literature surveys done for the feature selection and multiclass classification methods adopted in earlier research. Section 3 describes the implementation of the proposed GCA selection and the proposed divergent forest (DF) classifier. Experimental strategies and their comparison with the existing multiclass classification methods adopted for cancer classification are provided in the fourth section. Section 5 concludes and gives out the future direction to be adopted for further enhancement followed by the references.

## 2 Literature Survey

The data with higher dimensionality will always improve the difficulty in analyzing the underlying pattern which is considered a curse in the machine learning area. The primary concern while classifying the cancers is the high dimension of the available data sample that queries the reliability of the estimation and classification carried over the data [5]. Genetic data analysis based on the available DNA gene expression data is carried out widely for cancer analysis. DNA sequences consist of a wide number of genes which has enormous genes that are irrelevant to any of the cancer type considered for classification. This vast amount of genes that are available in the expression but do not support the classification has to be removed or to be hidden from the genetic data [6].

Feature selection methods had been a solution to identify the specific gene markers that are directly related to the specific cancer type. These methods are widely used in the appropriate selection of data dimensions while relating the genetic information to the cancer type that which the particular data belongs [7]. In feature selection, there are three different methods widely used that include filtering, wrapper,

**Fig. 1** Box plot of the sample data



and embedded feature selection methods. The effective set of features that provides the most valuable information about the sample considered for analysis is identified through the ranking process in the filtering feature selection technique. Filtering methods are done to rank the available features based on their impact on classifying the data samples. Statistical methods are also utilized as a filtering technique to rank the available set of features. Mutual information can be calculated for the available set of features and the most relevant set of features is identified through ranking the features. Now the redundancy between the selected features can be eliminated using any of the redundancy removal methods that include minimum redundancy and maximum relevance (mRmR) [8].

Wrapper methods are also adopted as feature selection methods which are used as a tool to fine-tune the results of the classifier after evaluating the performance. Irrespective of the computational expensiveness of the wrapper methods, they provide effective outcomes for the gene selection process [9]. Hence it becomes essential to adopt the selection method that is less computational and more effective in the selection of the genes. While adopting the wrapper approach it is necessary to consider the methods paradigms including identifying the methods to find the search space, identifying the features that support the classifiers, and the classifier performance evaluation [10, 11]. Another approach for appropriate feature selection is the embedded approach which is used to scrutinize the best set of features based on certain parameters. Regression methods are utilized as an embedded method in recent days to odd out the best set of features [12]. Several methods are adopted for solving the diverse grouping issues including genetic methods, support vector machines [13, 14], KNN [15], particle swam optimization (PSO) [16], and, so on. After performing several hybridizations and comparison strategies including GA-SVM, PSO-SVM, and artificial bee colony (ABC)-SVM, Zhu. et. al [17], concluded that genetic algorithms are highly effective in feature extraction from the original data and classification of the available data samples.

The classification accuracy of the GA method is also high in the case of diverse grouping problems when compared to other regular classifiers. While performing hybridization with support vector machines, genetic algorithms had even outperformed the other optimization methods including PSO and ABC methods. The other diversity search methods including Tabu search [18] and local search [19] had also compared for the performance evaluation of the hybridized GA-SVM method where the hybrid method had outperformed the searching methods in terms of accuracy, specificity, sensitivity, and also in time complexity. The optimal selection of genes is also an important area that is essential for cancer classification problems through microarray or through RNA-Seq data where GA plays a vital role in selecting the optimal set which is carried out in several types of research [20–23]. In healthcare sector, apart from the diagnosing mechanism, it also becomes essential for those mechanism to be prevented from external attacks and change overs in the original data [28]. The decision analytics is essential after the evaluation of performance criterion [29].

The kinds of literature gone through in the multiple-class cancer classification can be compared and analyzed in two different categories. The first one is extracting the features that support the classifier to perform multiple classifications based on the type of cancers considered for analysis. In recent literature [7–16], the researchers adopted the existing feature selection approaches i.e., either filter, wrapper, or embedded approaches using statistical methods or using optimization algorithms. Even though the statistical methods do the ranking of features based on their impact on the target outcome, sometimes it neglects the relevant features that could support the classifier to produce better performance. Dimensionality reduction methods adopted in literature [7–9] had reduced the characteristics of the gene expression data which had also limited the performance with a highest accuracy of only 82% which could not be a benchmark for clinical data analysis. Optimization and genetic methods [10–16] have given better performance with 86% accuracy but this leads to computational complexity. The need for a novel classification algorithm was due to the lack of performance by the existing classifiers in [17–22] since they fail to calculate the decision boundary from the features using probabilistic approaches.

Through several recent kinds of literature, our proposed work utilized a novel genetic clustering algorithm (GCA) for efficient feature selection and the classification is done through the proposed novel divergent random forest multiclass classification method.

## 3 Methods

We proposed two different strategies for selecting the features and for classifying the data samples. The selection of attributes is done by the proposed GCA which calculates the efficient feature set from the available data samples and the classification is done through the proposed probable divergent random forest classifier that performs multiclass classification to categorize each of the data samples in its appropriate class.

### 3.1 Genetic Clustering Algorithm for Feature Selection

Genetic algorithm is adopted widely to solve grouping problems in classifying the data samples belonging to genetic expression cancer RNA-Seq data. Our proposed GCA also utilizes the working footprints of the primary genetic algorithm which are shown in Fig. 2. The traversal from GA to GCA requires a conventional establishment of the encoding process. Solving the problems through optimization methods is essential in the scenario where attaining the optimal solution with the available data information is extremely difficult [25]. Charles Darwin proposed GA as a bio-inspired algorithm to solve optimization problems. This method will create a set of solutions for solving the optimization issues where every single solution performs a fight for survival in its ecosystem. Initially, population fitness is not considered as a parameter since the solutions will start at every instance in a random manner that might have a huge difference from the absolute solution to be obtained. Survival is accordingly based on best and worst fitness values obtained by individuals. The optimal solution will be identified through the Genetic Algorithm only after the evolution of several generations from which the fitness value and the lot in life of the individuals will act as a parameter to name the individual as an optimum one. The different parameters enriched with the data can be considered as a state and the different representations could be adopted to project the data in a better way using the genetic algorithm. The evolutionary process of the data can be balanced by incorporating the mutation operation which can overcome the sparsity and the data that could act as an outlier leading to the misclassification [24]. It is also evident that the localization of the irrelevant data within the feature vector can be performed by incorporating genetic algorithms. The data is moved out of the vector by cumulatively summing its occurrence [26]. The evolutionary process of the generations in the genetic algorithm is shown in Fig. 2. Several operations including fitness evaluation, individual selection for mate choice, cross-over, mutation of the generations, and survivor selection for the iterative generations are the steps involved in the evolution of generations in GAs.

The working of the GA begins with the available population. Once the evaluation of fitness function is carried out, the obtained solution is marked as an offspring and it will be added to the current population. The new solution is created as a complement to the existing two solutions that have to be obtained earlier based on the offspring added to the population. In the solution code, a consistent modification will be made through the GA as a mutation. Once the recombination and mutation are performed then the offspring's fitness value will be calculated. Through the above steps if the population is increased then it is essential to perform the population control method which considers only the survivor individuals whereas others will be discarded from the population. The algorithm keeps on producing multiple generations until a stop condition is initiated to the algorithm. Those are conditions that might be a population count or the convergence criteria. The convergence is calculated on the basis of the best fitness value for the current entity and the mean best fitness value obtained for the gross population. To eradicate

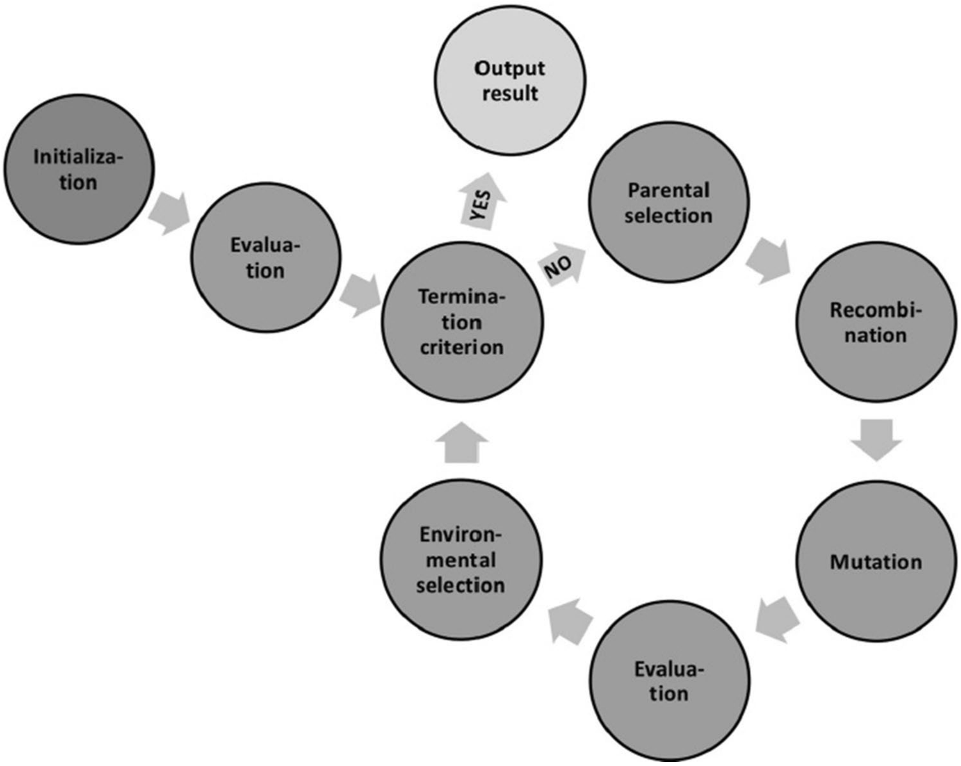


**Fig. 2** Workflow of the genetic algorithm

the error from the convergence we used an epsilon value of 0.0001.

$$f_{\text{mean}} - \epsilon \leq f_{\text{best}} \leq f_{\text{mean}} + \epsilon \tag{1}$$

At the stopping point, GA produces the individual set with the best fitness obtained from which one or more can be identified as an appropriate one. The proposed problem attains the best solution through encoding the obtained individuals. Our proposed GCA acts as an alternative to the GA to solve the clustering as well as grouping issues in the existing GA. The term 'Clustering' indicates the specialized encoding methods to produce hierarchy-oriented strategies in clustering-induced problems [3]. The novelty of GCA lies in its encoding strategy where the solutions are encoded as arrays with binary sub-arrays: part one is the assignment and part two is the grouping. The dual array parts belong to the natural numbers where the assignment coincides with the elements considered for categorization and the grouping matches with the number of groups considered for encoding. Let us consider Fig. 3 as an example of 14 elements encoding a grouping solution.

The individuals are named in the group from 1 to 4. In the figure, the initial element is not associated with any of the groups whereas the iterative second, third, and fourth elements are associated with the second group. Along with the encoding process, the proposed GCA is also improvised in terms of recombination and mutation operation. In the recombination process, the traditional GA has the cross-over with the mutation of another individual element whereas the proposed GCA introduces the grouping concepts in between the cluster of elements to perform the cross-over with the recombination and mutation of the elements. To perform the feature selection using the filtering approach the algorithm utilizes the mutual information-based statistical method to select the appropriate feature set. Mutual information between the features is calculated for the identification of relevance between the available features which are evolved through the GCA method. Mutual information has an effective fitness function that identifies the dependence between the feature vector considered for analysis. Once the relevance set is calculated using the mutual information equation as denoted in Eq. 2, the optimal set of features is calculated by the L1 regularized Logistic regression method that acts as an annexure to the filter method which is also denoted as an embedded approach for optimum feature calculation.

$$\mathrm{MD}(F, O) = \log_2\left(\frac{P(F, O)}{P(F)P(O)}\right) \tag{2}$$

The available dataset consists of 802 data samples and 16,382 gene expressions as a feature vector. The proposed feature selection method had identified the most dominating 21 genes and their cluster is calculated into four groups which all act as a new set of the feature vector for the dataset and it will act as input data to the DF classifier.

## 3.2 Divergent Forest for Multiple Classification

The dataset after performing the feature selection stage consists of 802 samples and 4 clusters of features that possess 21 genes with it. Since we have a labeled data set, the cancer types can be considered for multiple-class analysis. The key labels are included with the data that include BRCA denoting breast cancer, KIRC indicates the data label for kidney cancer, COAD indicates colon cancer occurring in the larger intestine, LUAD denoting Lung cancer, and PRAD indicates Prostate cancer. 16,382 gene expressions belong to any one type of cancer that is labeled in the dataset. Since we have more than two class labels, it is essential to adopt multiple classifications to categorize each of the data samples. Random forest is widely used as a multiclass classifier to categorize more than two classes labeled in the dataset. The limitation of using the random forest is it performs the node splitting by calculating the information gained between the nodes [3]. While considering the clinical data for analysis, the information might vary from time to time where adopting the IG strategy might lead to overfitting of the machine learning model. Thus to avoid such overfitting issues and to maintain the classification accuracy we proposed a divergent forest (DF) which utilizes the probable Kulback Leibler divergent (KLD) method to split the nodes for the forest. Unlike IG which calculates the difference in obtained information, the KLD compares how the distribution of the data varies from one to the other. The working of the DF is given in Fig. 4. Initially, the data is divided into several sets using the bootstrapping strategy, and for each of the sets, the difference in data distribution is calculated using KLD. Based on the node split the majority of voting is conducted between the classes and the class that obtains the major vote will be categorized for the data sample considered for analysis.

The input data i.e., RNA-Seq data is processed using the proposed GCA and the multiple classifications for each of the samples are done through DF classifier. The original data

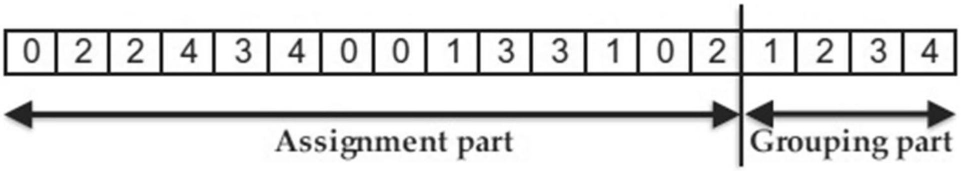


**Fig. 3** Encoding based on grouping for 14 elements

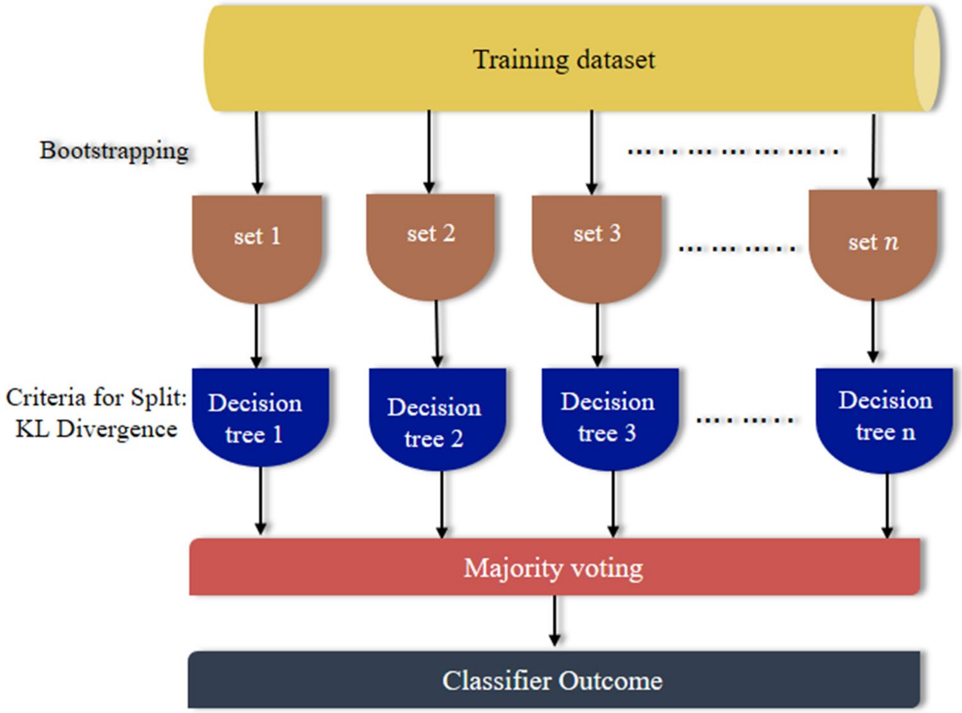


**Fig. 4** Workflow of the proposed divergent forest classifier

considered for analysis and the classification results are given in Sects. 4 and 5.

## 4 Dataset Description

The proposed GCA and DF are applied to the available data which has 16,382 gene expressions and 802 samples. From the overall dataset, the most influential genes that are responsible for the behavior of the data samples are considered for data analysis. The combinations of the most influential gene expressions are formed based on the five different groups and it acts as the input data for the multiclass classifier. The available samples and the entire dataset are further divided into three subsets namely the training set, the test set, and the validation set. The initial set is used to train the proposed machine learning model and the test is used for performance analysis. During this entire process, the validation set is kept as a hidden set, and once the halting condition for the generation evaluation is attained, the absolute outcomes concluded are calculated through the final set kept out for validation. The split up of the data is carried out in such a way that 80% of the data is utilized as the training set, 10% of the data is utilized as the test set, and the remaining 10% of the data available in the dataset is utilized as a validation set. To perform the multiclass classification from the available dataset we utilized 380 data samples from the entire gene expression data sample combination. The split up along with its class number is given in Table 1.

**Table 1** Classes and sample details from the dataset

| Category | Class description | Number of entities |
|---|---|---|
| BRCA | 1 | 78 |
| COAD | 2 | 78 |
| KIRC | 3 | 78 |
| LUAD | 4 | 78 |
| PRAD | 5 | 78 |
| Total | | 390 |

The above table of class descriptions is obtained after applying the resampling method to the originally available dataset. Synthetic minority oversampling technique (SMOTE) is utilized for avoiding the imbalance that occurs in the available dataset. Thus after the oversampling, there is a balance in the dataset with evenly matched numbers of samples in every category of cancer types. Now, the data has evenly distributed sample details and it can build an effective classifier.

## 5 Experimental Results

There are 5 gene set expressions and 390 samples are considered for analysis to perform the multiclass classification. Initially, each sample is categorized with its respective class then the overall classification for entire samples is performed. The performance is analyzed in terms of accuracy, specificity, and sensitivity. Accuracy is calculated as the summation of overall true predictions from the entire predictions made by the model. Specificity deals with the true predictions which are calculated by the summation of true positive and false negative predictions made by the model. Sensitivity deals with the negative predictions or false predictions made by the model. The performance obtained by PF is compared with existing and widely used multiple classifiers including logistic regression, multilayer perceptron, random forest, artificial neural networks, support vector machines, and KNN. Performance evaluation results obtained by the proposed and the existing methods are compared in the following sections for each cancer type considered for analysis.

### 5.1 Performance Evaluation of BRCA Classification

BRCA denotes the breast cancer class available in the gene expression RNA-Seq data considered for the analysis. There are 78 samples belonging to the category of BRCA and the remaining 312 data samples are other cancer types. The classification results obtained by the proposed and the existing classifiers are shown in Table 2.

From Table 2, it is evitable that the divergent forest (DF) had achieved the maximum accuracy of 94.09%, improvised specificity of 91.17%, and a maximum sensitivity of 93.33% which had outperformed all the other existing classifiers. The graph comparison of the results had been depicted in Fig. 5.

From Fig. 5, it is evitable that the DF had outperformed the accuracy achieved by the existing ANN by 6.26%, in specificity DF classifier had outperformed ANN by 8.91%, and in sensitivity, DF classifier outperformed ANN by 8.87%. Thus, DF classifier outperformed all the other classifiers in overall performance for the BRCA classification process.

### 5.2 Performance Evaluation for COAD Classification

COAD denotes the colon cancer samples available in the gene expression RNA-Seq data considered for the analysis. 78 samples are belonging to the category of COAD and the remaining 312 data samples are other cancer types. The

**Table 2** Performance comparison for BRCA evaluation

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 64.12 | 68.22 | 67.73 |
| Multilayer perceptron | 82.24 | 79.21 | 81.18 |
| Random forest | 76.19 | 81.21 | 79.93 |
| ANN | 87.83 | 82.26 | 84.46 |
| SVM | 69.23 | 71.11 | 70.06 |
| KNN | 67.11 | 74.15 | 69.19 |
| Divergent forest (DF)—proposed | 94.09 | 91.17 | 93.33 |

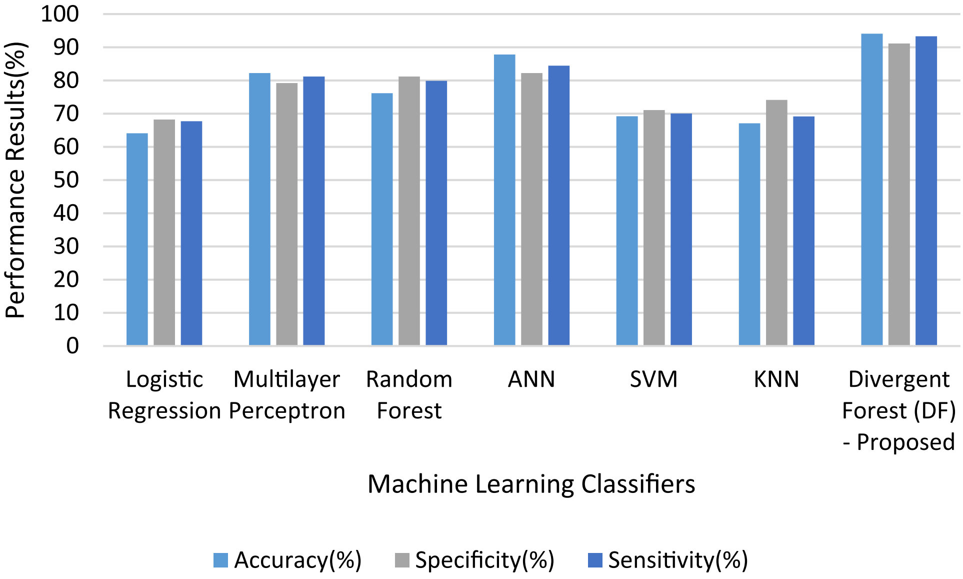


**Fig. 5** Performance comparison chart for BRCA evaluation

classification results obtained by the proposed and the existing classifiers are shown in Table 3.

From Table 3, it is evitable that the divergent forest (DF) had achieved a maximum accuracy of 95.12%, improved specificity of 93.56%, and a maximum sensitivity of 94.08% which had outperformed all the other existing classifiers. From Fig. 6, it is evitable that the proposed Divergent forest had outperformed the accuracy achieved by the existing ANN by 6.8%, in specificity DF classifier had outperformed ANN by 9.43%, and in sensitivity, DF classifier outperformed ANN by 7.36%. Thus, DF classifier outperformed all the other classifiers in overall performance for the COAD classification process.

## 5.3 Performance Evaluation for KIRC Classification

KIRC denotes the kidney cancer class available in the gene expression RNA-Seq data considered for the analysis. There are 78 samples belonging to the category of KIRC class and the remaining 312 data samples are other cancer types. The classification results obtained by the proposed and the existing classifiers are shown in Table 4.

From Table 4, it is evitable that the divergent forest (DF) had attained the maximum accuracy of 92.28%, improvised specificity of 90.73%, and a maximum sensitivity of 92.04% which had outperformed all the other existing classifiers. The graph comparison of the results had been depicted in Fig. 7 where DF had outperformed the accuracy achieved by the existing multilayer perceptron by 7.61%, in specificity DF classifier had outperformed multilayer perceptron by 3.41%, and in sensitivity, DF classifier had outperformed the multilayer perceptron by 5.49%. Thus, DF classifier outperformed all the other classifiers in overall performance for KIRC classification process.

## 5.4 Performance Evaluation for LUAD Classification

LUAD denotes the lung cancer data samples available in the gene expression RNA-Seq data considered for the analysis. There are 78 samples belonging to the category of LUAD class and the remaining 312 data samples are other cancer types. The classification results obtained by the proposed and the existing classifiers are shown in Table 5.

From Table 5, it is evitable that the divergent forest (DF) had attained a maximum accuracy of 96.84%, improvised

**Table 3** Performance comparison of COAD evaluation

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 63.19 | 67.64 | 67.13 |
| Multilayer perceptron | 83.30 | 80.06 | 82.24 |
| Random forest | 77.03 | 79.91 | 81.12 |
| ANN | 88.32 | 84.13 | 86.72 |
| SVM | 71.16 | 73.39 | 70.16 |
| KNN | 65.84 | 62.22 | 67.01 |
| Divergent forest (DF)—proposed | 95.12 | 93.56 | 94.08 |

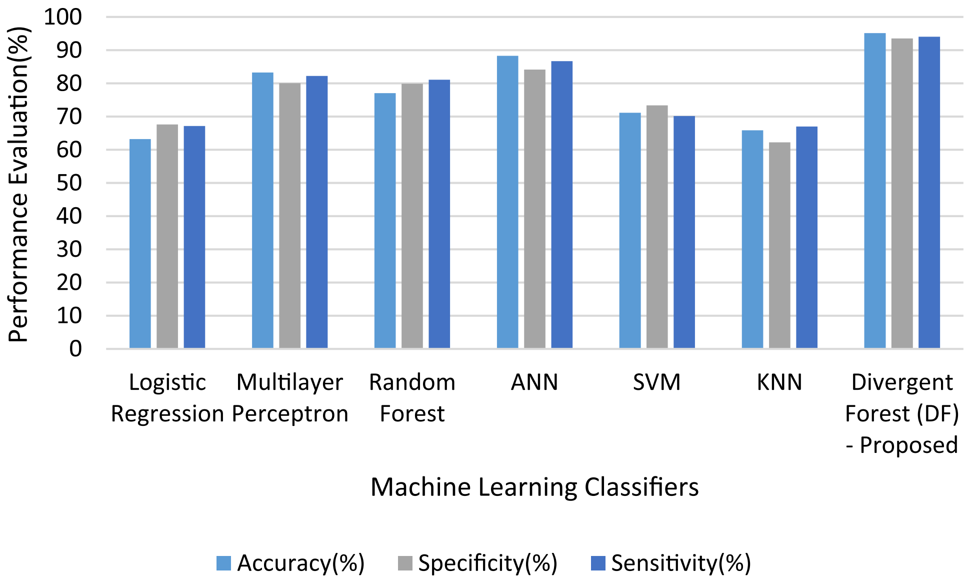


**Fig. 6** Performance comparison chart for COAD evaluation

specificity of 93.22%, and a maximum sensitivity of 94.65% which had outperformed all the other existing classifiers. The graph comparison of the results had been depicted in Fig. 8

From Fig. 8, it is evitable that the DF had outperformed the accuracy achieved by the existing multilayer perceptron by 9.16%, the specificity of the DF classifier had outperformed multilayer perceptron by 2.91%, the sensitivity of DF classifier had outperformed multilayer perceptron by 10.17%. Thus, DF classifier outperformed all the other classifiers in overall performance for the LUAD classification process.

## 5.5 Performance Evaluation for PRAD Classification

PRAD denotes the prostate cancer data samples available in the gene expression RNA-Seq data considered for the analysis. There are 78 samples belonging to the category of PRAD class and the remaining 312 data samples are other cancer types. The classification results obtained by the proposed and the existing classifiers are shown in Table 6.

From Table 6, it is evitable that the divergent forest (DF) had attained a maximum accuracy as 97.74%, improvised specificity of 96.36%, and a maximum sensitivity

**Table 4** Performance comparison of KIRC evaluation

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 60.84 | 61.15 | 60.33 |
| Multilayer perceptron | 84.67 | 87.32 | 86.55 |
| Random forest | 83.31 | 80.84 | 81.18 |
| ANN | 82.34 | 79.76 | 80.22 |
| SVM | 84.54 | 79.92 | 83.62 |
| KNN | 68.86 | 63.11 | 66.63 |
| Divergent forest (DF)-proposed | 92.28 | 90.73 | 92.04 |

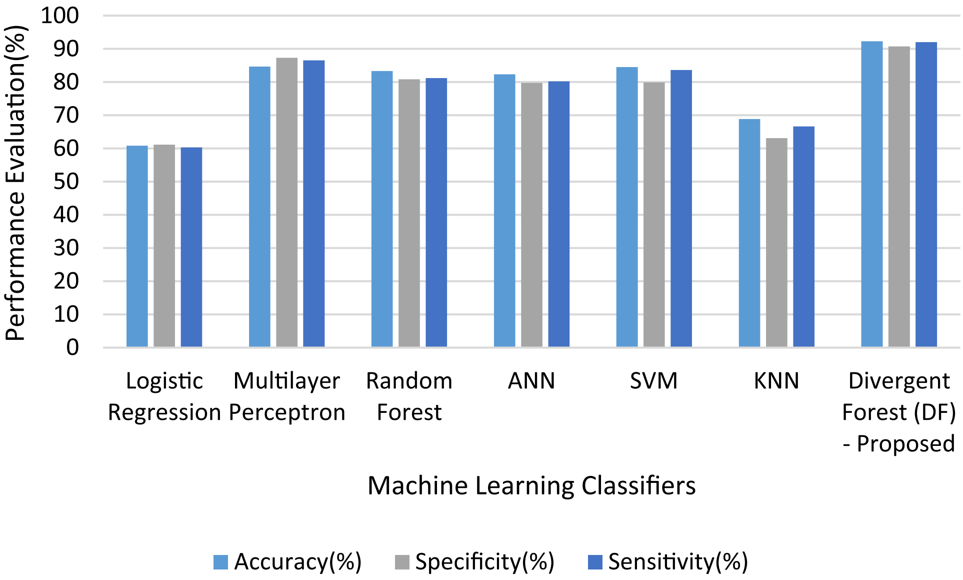


**Fig. 7** Performance comparison chart for KIRC evaluation

**Table 5** Performance comparison of LUAD evaluation

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 68.83 | 63.32 | 67.11 |
| Multilayer perceptron | 87.68 | 90.31 | 84.48 |
| Random forest | 86.11 | 82.12 | 83.39 |
| ANN | 86.69 | 83.01 | 81.19 |
| SVM | 85.52 | 82.28 | 83.09 |
| KNN | 70.17 | 68.88 | 71.21 |
| Divergent forest (DF)—proposed | 96.84 | 93.22 | 94.65 |

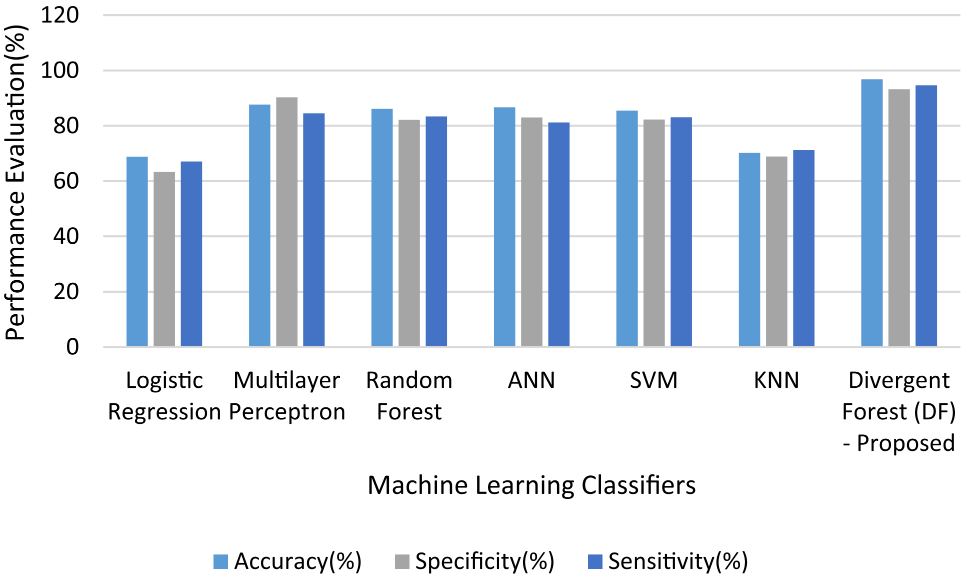


**Fig. 8** Performance comparison chart for LUAD evaluation

of 97.35% which had outperformed all the other existing classifiers. The graph comparison of the results had depicted in Fig. 9.

From Fig. 9, it is evitable that DF had outperformed the accuracy achieved by the existing random forest by 9.43%, the specificity of DF classifier had outperformed the existing random forest by 5.23%, and the sensitivity of DF classifier had outperformed the random forest by 7.13%. Thus DF classifier outperformed all the other classifiers in overall performance for the PRAD classification process.

### 5.6 Performance Evaluation for Overall Multiclass Classification

In earlier classification among the available data samples, each type of cancer data sample is categorized using the existing and the proposed classifiers. In this section using the multiclass classifiers all the available 390 data samples that belong to five different cancer classes including the BRCA, COAD, KIRC, LUAD, and PRAD classes are categorized using the existing and the proposed classifiers. The obtained results are shown in Table 7 where DF classifier outperformed all the other existing multiclass classifiers in terms of accuracy, specificity, and sensitivity.

From the above table, it is evitable the proposed divergent forest (DF) classification had attained a maximum accuracy of 95.21%, improvised specificity of 93%, and a maximum sensitivity of 94.29% which outperformed all the other existing classifiers. The graph comparison of the results had been shown in Fig. 10 where DF had outperformed the accuracy achieved by the existing ANN by 9.3%, and the specificity of the DF classification had outperformed the multilayer perceptron by 7.79%, and the sensitivity of DF classification had outperformed the multilayer perceptron by 9.94%. Thus, DF outperformed all the other existing classifiers in overall performance for the entire multiclass classification process.

Thus the experimental setup and the results obtained for the multiclass classification were discussed in this section. From the results obtained for the performance evaluation process that includes each of the five types of cancers and the overall data samples considered for multiclass classification, it is evitable that DF classifier which utilized the features selected through the proposed GCA had produced highest performance when compared to all the other existing machine learning methods.

**Table 6** Performance comparison of PRAD evaluation

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 70.11 | 71.36 | 69.24 |
| Multilayer perceptron | 85.55 | 89.19 | 87.34 |
| Random forest | 88.31 | 91.13 | 90.22 |
| ANN | 84.41 | 83.31 | 82.29 |
| SVM | 86.16 | 84.55 | 73.31 |
| KNN | 72.91 | 70.18 | 71.34 |
| Divergent forest (DF)-proposed | 97.74 | 96.36 | 97.35 |

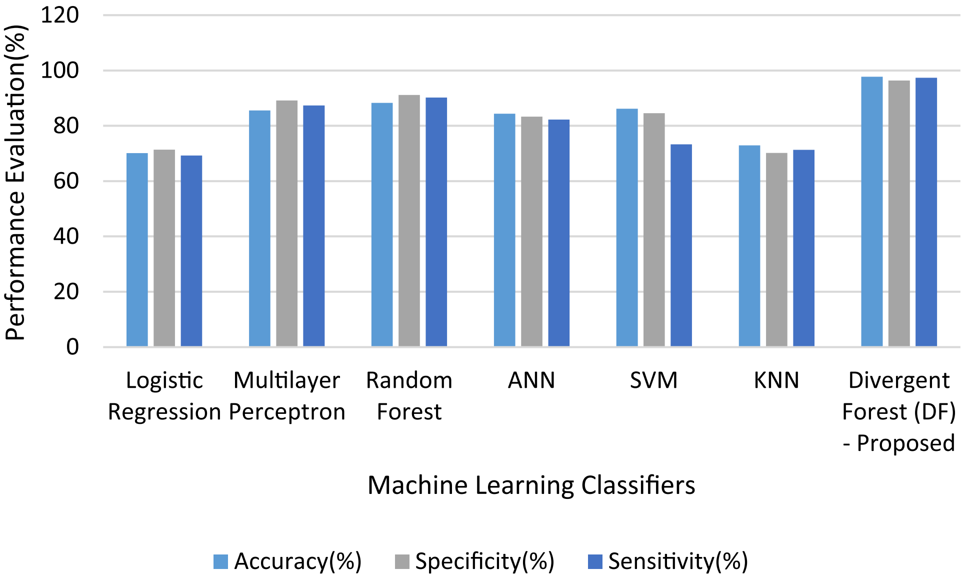


**Fig. 9** Performance comparison chart for PRAD evaluation

**Table 7** Performance comparison of overall multiclass classification

| Classifier | Accuracy (%) | Specificity (%) | Sensitivity (%) |
|---|---|---|---|
| Logistic regression | 65.41 | 66.33 | 66.30 |
| Multilayer perceptron | 84.68 | 85.21 | 84.35 |
| Random forest | 82.19 | 83.04 | 83.16 |
| ANN | 85.91 | 82.49 | 82.97 |
| SVM | 79.32 | 78.25 | 76.04 |
| KNN | 68.97 | 67.70 | 69.07 |
| Divergent forest (DF)-proposed | 95.21 | 93 | 94.29 |

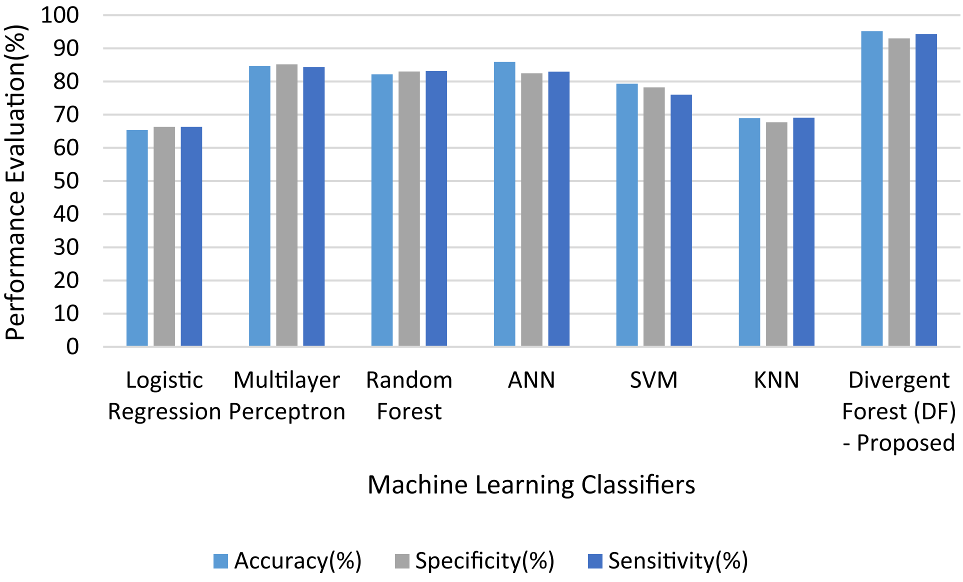


**Fig. 10** Performance comparison chart for overall multiclass classification

## 6 Conclusion and Future Scope

Diagnosis of cancer type based on computational methods had gathered its importance over the past decade which motivates individuals to undergo regular clinical examinations to treat the disease. Based on the type of data considered for analysis which includes text, voice, signals, images, etc., a large number of researches had been carried out for binary classification of samples that is to diagnose whether the data sample belongs to normal or abnormal category. Multiclass classification is the method of considering the data samples with more than two labels (multiclass) and categorizing them accordingly. To perform an efficient multiclass classification to categorize five different cancer types that including breast cancer, colon cancer, kidney cancer, lung cancer, and prostate cancer, we proposed a genetic cluster algorithm (GCA) for feature selection and a divergent forest (DF) classifier for multiple classifications of data samples. Through the various experimental results obtained it is evitable that the proposed feature selection and classification methods had achieved the highest performance with an accuracy of 95.21%, specificity of 93%, and sensitivity of 94.29%. In the future, a metaheuristic machine learning methodology can be performed for the optimistic selection of genetic expressions from the majority voting process which can analyze more gene expressions simultaneously for cancer-type classification with minimal time and computational complexity.

**Author Contributions** SL, the first and corresponding author of the article developed the proposed methodologies and also identified the gene expression data. The results are computed from the identified data using the proposed feature selection and classification methods by her. LS, the second author had made a major contribution in framing the article, and the obtained results are cross-validated by her with the existing methodologies. All the authors of the article have read and approved the final article.

**Funding** No funding was received to assist with the preparation of this manuscript.

**Data Availability** Not applicable.

## Declarations

**Conflict of Interest** The authors declare that they have no conflict of interest.

## References

1. Rongjun, X.I., Khalil, I., Badsha, S., Atiquzzaman, M.: Collaborative extreme learning machine with a confidence interval for P2P learning in healthcare. Comput. Netw. **149**, 127–143 (2019)
2. Santhakumar, D., Logeswari, S.: Efficient attribute selection technique for leukaemia prediction using microarray gene data. Soft. Comput. **24**(18), 14265–14274 (2020)
3. Balajee, A., Venkatesan, R.: Machine learning based identification and classification of disorders in human knee joint–computational approach. Soft. Comput. **25**(20), 13001–13013 (2021)
4. Santhakumar, D., Logeswari, S.: Hybrid ant lion mutated ant colony optimizer technique for Leukemia prediction using microarray gene data. J. Ambient. Intell. Humaniz. Comput. **12**(2), 2965–2973 (2021)
5. Libbrecht, M.W., Noble, W.S.: Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**(6), 321–332 (2015)
6. Zhao, G., Wu, Y.: Feature subset selection for cancer classification using weight local modularity. Sci. Rep. **6**(1), 1–6 (2016)
7. Salem, H., Attiya, G., El-Fishawy, N.: Classification of human cancer diseases by gene expression profiles. Appl. Soft Comput. **50**, 124–134 (2017)
8. Pavithra, D., Lakshmanan, B.: Feature selection and classification in gene expression cancer data. In: 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1–6. IEEE (2017)
9. Tang, C., Cao, L., Zheng, X., Wang, M.: Gene selection for microarray data classification via subspace learning and manifold regularization. Med. Biol. Eng. Comput. **56**(7), 1271–1284 (2018)
10. Piao, Y., Ryu, K.H.: Detection of differentially expressed genes using feature selection approach from RNA-seq. In: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 304–308. IEEE (2017)
11. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. **3**(02), 185–205 (2005)
12. Alphonse, B., Rajagopal, V., Sengan, S., Kittusamy, K., Kandasamy, A., Periyasamy, R.: Modeling and multi-class classification of vibroarthographic signals via time domain curvilinear divergence random forest. J. Ambient. Intell. Humaniz. Comput. (2021)
13. Zheng, C.H., Huang, D.S., Shang, L.: Feature selection in independent component subspace for microarray data classification. Neurocomputing **69**(16–18), 2407–2410 (2006)
14. Maji, P., Das, C.: Relevant and significant supervised gene clusters for microarray cancer classification. IEEE Trans. Nanobiosci. Nanobiosci. **11**(2), 161–168 (2012)
15. Brimberg, J., Mladenović, N., Todosijević, R., Urošević, D.: Solving the capacitated clustering problem with variable neighborhood search. Ann. Oper. Res. **272**(1), 289–321 (2019)
16. Alam, S., Dobbie, G., Koh, Y.S., Riddle, P., Rehman, S.U.: Research on particle swarm optimization based clustering: a systematic review of literature and techniques. Swarm Evol. Comput. **17**, 1–3 (2014)
17. Zhu, X., Li, N., Pan, Y.: Optimization performance comparison of three different group intelligence algorithms on a SVM for hyperspectral imagery classification. Remote Sens. **11**(6), 734 (2019)
18. Palubeckis, G., Ostreika, A., Rubliauskas, D.: Maximally diverse grouping: an iterated tabu search approach. J. Oper. Res. Soc. **66**(4), 579–592 (2015)

19. López-Ibáñez, M., Paquete, L., Stützle, T.: Exploratory analysis of stochastic local search algorithms in biobjective optimization. In: Experimental methods for the analysis of optimization algorithms, pp. 209–222. Springer, Berlin (2010)
20. Bonilla-Huerta, E., Hernandez-Montiel, A., Morales-Caporal, R., Arjona-Lopez, M.: Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. IEEE/ACM Trans. Comput. Biol. Bioinform. Bioinform. **13**(1), 12–26 (2015)
21. Zhang, Y,, Deng, Q., Liang, W., Zou, X.: An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. BioMed Research International. 2018 Aug 30 (2018)
22. Salman, I., Ucan, O.N., Bayat, O., Shaker, K.: Impact of metaheuristic iteration on artificial neural network structure in medical data. Processes **6**(5), 57 (2018)
23. Feitosa Neto, A.A., Canuto, A.M., Xavier-Junior, J.C.: Hybrid metaheuristics to the automatic selection of features and members of classifier ensembles. Information **9**(11), 268 (2018)
24. Nabeeh, N.: Assessment and contrast the sustainable growth of various road transport systems using intelligent neutrosophic multi-criteria decision-making model. Sustain. Mach. Intell. J. **2** (2023)
25. Alenizi, J.A., Alrashdi, I., SFMR-SH.: Secure framework for mitigating ransomware attacks in smart healthcare using blockchain technology 2. SMIJ. **2**(2), 19 (2023)
26. Mohamed, Z., Ismail, M.M., Abd El-Gawad, A.: Sustainable supplier selection using neutrosophic multi-criteria decision making methodology. Sustain. Mach. Intell. J. **3** (2023)
27. García Díaz, P., Martínez Rojas, J.A., Utrilla Manso, M., Monasterio, E.L.: Analysis of water, ethanol, and fructose mixtures using nondestructive resonant spectroscopy of mechanical vibrations and a grouping genetic algorithm. Sensors **18**(8), 2695 (2018)
28. Lu, K.D., Wu, Z.G.: Multi-objective false data injection attacks of cyber–physical power systems. IEEE Trans. Circuits Syst. II Express Briefs **69**(9), 3924–3928 (2022)
29. Lu, K.D., Wu, Z.G.: Genetic algorithm-based cumulative sum method for jamming attack detection of cyber-physical power systems. IEEE Trans. Instrum. Meas. **71**, 1 (2022)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.