big data and cognitive computing

MDPI

Article

Combination of Reduction Detection Using TOPSIS for Gene Expression Data Analysis

Jogeswar Tripathy [1], Rasmita Dash [1,*], Binod Kumar Pattanayak [1], Sambit Kumar Mishra [2,*], Tapas Kumar Mishra [2] and Deepak Puthal [3,*]

[1] ITER, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar 751030, India; jogeswar.tripathy@gmail.com (J.T.); binodpattanayak@soa.ac.in (B.K.P.)
[2] Department of Computer Science and Engineering, SRM University-AP, Amaravati 522502, India; kmtapas@gmail.com
[3] Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi 127788, United Arab Emirates
* Correspondence: rasmitadash@soa.ac.in (R.D.); skmishra.nitrkl@gmail.com (S.K.M.); dputhal88@gmail.com (D.P.)

Abstract: In high-dimensional data analysis, Feature Selection (FS) is one of the most fundamental issues in machine learning and requires the attention of researchers. These datasets are characterized by huge space due to a high number of features, out of which only a few are significant for analysis. Thus, significant feature extraction is crucial. There are various techniques available for feature selection; among them, the filter techniques are significant in this community, as they can be used with any type of learning algorithm and drastically lower the running time of optimization algorithms and improve the performance of the model. Furthermore, the application of a filter approach depends on the characteristics of the dataset as well as on the machine learning model. Thus, to avoid these issues in this research, a combination of feature reduction (CFR) is considered designing a pipeline of filter approaches for high-dimensional microarray data classification. Considering four filter approaches, sixteen combinations of pipelines are generated. The feature subset is reduced in different levels, and ultimately, the significant feature set is evaluated. The pipelined filter techniques are Correlation-Based Feature Selection (CBFS), Chi-Square Test (CST), Information Gain (InG), and Relief Feature Selection (RFS), and the classification techniques are Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and k-Nearest Neighbor (k-NN). The performance of CFR depends highly on the datasets as well as on the classifiers. Thereafter, the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method is used for ranking all reduction combinations and evaluating the superior filter combination among all.

Keywords: feature selection; machine learning; microarray gene extraction; pipelining; TOPSIS

Citation: Tripathy, J.; Dash, R.; Pattanayak, B.K.; Mishra, S.K.; Mishra, T.K.; Puthal, D. Combination of Reduction Detection Using TOPSIS for Gene Expression Data Analysis. *Big Data Cogn. Comput.* **2022**, *6*, 24. https://doi.org/10.3390/bdcc6010024

Academic Editor: Rao Mikkilineni

Received: 27 December 2021
Accepted: 16 February 2022
Published: 23 February 2022

1. Introduction

Over the years, researchers have been trying with microarray technology to track gene expression on a genomic scale. Cancer diagnosis and classification are possible through examining the expression of genes. The use of microarray technology to analyze gene expression has opened up a world of possibilities for studying cell and organism biology [1]. Nowadays, every researcher primarily focuses especially on the behavior of genes across the conditions of the experiment studied; however, recently, biomedical applications have fueled both the use of available technologies and the efficient implementation of new analytical tools to deal with these complex data. Microarray data analysis yields useful results that aid in the resolution of gene expression problems. Cancer categorization is one of the most significant uses of microarray data analysis. This reflects variations in the levels of expression of various genes. However, categorizing gene expression profiles is a difficult process that has been classified as an NP-Hard issue. As a result, not all genes

have a role in the development of cancer. In clinical diagnosis, a large percentage of genes are insignificant [2].

Researchers were able to examine hundreds of gene expression patterns concurrently using microarray technology, which is useful in a variety of disciplines, particularly in medicine, mainly for the detection of cancer; in biomedical research, categorizing patient's gene expression profiles has become a regular topic. The main issue is dealing with the dimensionality of microarray data [3]. As microarrays have such a vast dimension, efficient algorithm exploration becomes too difficult for analyzing gene expression features. There are more incorrect characteristics in the dataset, due to which the algorithm's accuracy suffers considerably [4]. Due to this reason in the pre-processing stage, feature selection approaches are used to extract meaningful information. The goal of the feature selection method is to identify the most significant characteristics from the microarray data to reduce the feature set and enhance classification accuracy. Using feature selection and classification approaches, gene expression analysis of cancer diagnosis has been made. However, combining an efficient feature selection method and classifier is a critical task to avoid incorrect drug selection [5].

In a machine learning pool, there exist three feature selection techniques: the filter approach, wrapper approach, and embedded approach. Filter techniques are an essential element of strategically selecting features due to the cheap cost of computation, making them suitable when data sizes are too big for a learning algorithm or when resources are limited. Filter methods may be split into two groups based on how they work. First, there are univariate methods: each characteristic is assessed independently in this category. The "relation" between a feature and the class label is taken into account here. Features are graded based on their "relationship" with other feature-class pairs. Mutual Information (MI) and Chi-square are a few examples of this category. Second, there are ultivariate methods: in this scenario, characteristics are accessed in sets to see how well the sets can distinguish across classes. The sets that can discriminate better are more likely to offer a more accurate classification. Ranking techniques focusing on score-based feature subset selection techniques are the most used filter approaches. In a microarray dataset, the ranking approach may be thought of as a crucial mechanism for picking the k most relevant genes. Since the number of features in microarray data might be quite enormous, the learning algorithm's accuracy is severely harmed. As a result, selecting the top k genes from microarray data is an important pre-processing step [6]. In typical microarray research, the high number of features and the relatively limited number of observations (samples) offer numerous statistical difficulties, which are referred to as the "curse of dimensionality" in machine learning. As a result, after normalizing and pre-filtering the original datasets, we use several feature selection techniques to extract compact sets of discriminative features before using classification algorithms [7]. However, choosing a filter approach for gene selection is critical, because one technique may produce the best results for one dataset while another produces the best results for another.

Inspired by the above analysis, which is discussed by several researchers, this paper proposes a pipeline of reduction combinations using filter approaches. Four feature ranking algorithms are taken into account in this model for obtaining a better feature subset from datasets, as shown in Table 1: Correlation-Based Feature Selection (CBFS), Chi-Square Test (CST), Information Gain (InG), and Relief Feature Selection (RFS). Then, the classification techniques such as Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), and k-Nearest Neighbor (k-NN) are used to classify the microarray databases. Forming the combination of four FS approaches, 16 pipelines are built up to reduce the feature subset in four phases to come up with more useful features. Considering 10 or fewer features, the final optimal feature subset is generated. With these reduced feature subsets, the performance of each pipeline is measured with the considered four classification techniques. In certain circumstances, it may become difficult to decide on a single parameter; thus, lastly, all reduction combinations are compared on different parameters such as accuracy, sensitivity, jaccard, specificity, and gmean. Thereafter, to finalize which reduction combination in

a pipeline is stable irrespective of the dataset, the TOPSIS approach is used for optimal decision making.

Table 1. Acronym and Description.

Acronym	Descriptions
ANN	Artificial Neural Network
BOFS	Bi-Objective Feature Selection
CBFS	Correlation-Based Feature Selection
CFR	Combination of Feature Reduction
CR	Reduction Combination
CST	Chi-Square Test
DBC	Distance-Based Clustering
DT	Decision Tree
EM	Expectation Maximization
FN	False Negative
FP	False Positive
FR	Feature Reduction
FS	Feature Selection
GEM	Gene Expression Microarray
GRM	Gray Relational Model
InG	Information Gain
k-NN	k-Nearest Neighbor
LR	Logistic Regression
MADM	Multi-Attribute Decision Making
MCDM	Multi-Criteria Decision Making
MI	Mutual Information
MIMAGA	Mutual Information Maximization and Genetic Algrithm
MLP Neural Nets	Machine Learning Perception Neural Networks
MOFSCE	Multi-Objective Feature Selection and Classifier Ensemble
PCA	Principal Compotent Analysis
RBF	Radial Basis Function
RF	Random Forest
RFS	Relief Feature Selection
SAW	Simple Additive Weighting
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine Recursive Feature Elimination
TN	True Negative
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
TP	True Positive

The organization of this analysis is as follows. Section 1 introduces a generalization of the research, Section 2 describes the recent related work about the sequence of feature ranking methods with different classifiers and TOPSIS, i.e., multi-criteria selection techniques used on microarray data. Section 3 describes the proposed model, and Section 4 includes methodologies used with a detailed explanation; evaluation metrics are also discussed here with the Multi-Criteria Decision Making (MCDM) technique. Section 5 includes the complete experimental work with datasets used and the result analysis. In Section 6, the conclusion and future work are discussed, and finally, a discussion about the study is presented in Section 7.

2. Related Work

Feature ranking approaches are now used everywhere, including analytical techniques, summarizing extraction, sequential data processing, multidimensional data processing, and many more. Several studies use various filter techniques for feature selection. Hence, it is very difficult to identify a filter approach that can extract superior features from the

datasets of the respective application. Again, the classification performance depends a lot on the extracted feature. Therefore, rather than sticking to a single filter approach, combinations of two or more filter approaches are applied in the pre-processing stage. Thus, the literature survey is as follows, which focuses on feature reduction in different stages, implementing filter approaches. A hybrid feature selection approach for illness identification has been presented by Namrata Singh et al. [8]. They use cross-validation for partitioning and multiple filter approaches for feature ranking with weighted scores. Furthermore, the sequential forward selection process is also employed as a wrapper technique to find out the subset of features. Compared to the benchmark classifiers, such as Naive Bayes, Support Vector Machine with Radial Basis Function, Random Forest, and k-Nearest Neighbor, it is experienced that the four-step hybrid ensemble filter selection strategy outperforms fourteen feature selection algorithms. The empirical results clearly show that the suggested hybrid approach surpasses the competing methods in terms of accuracy, sensitivity, specificity, F1-score, the area under curve evaluation measures, and the number of selected features.

Andrea Bommert et al. compared the most advanced feature selection strategies on high-dimensional datasets in their work [9]. They compared 22 filter techniques from several toolboxes on 16 high-dimensional classification datasets from distinct fields as well as the methods that select the features of a dataset in the same order. They concluded that some filter methods appear to perform better than others but failed to identify the highly reliable filter methods. Furthermore, they suggested that filter methods are dependent on the dataset.

Cosmin Lazar et al. [10] emphasized gene prioritization and filter-based feature selection techniques for informative feature extraction in Gene Expression Microarray (GEM) analysis. Rasmita Dash et al. [4] employed a pipelining of the ranking approaches presented in their work that addresses the difficulties associated with the filter approach. Few of the lower-ranked features are deleted at each level of the pipeline, resulting in a pretty decent subset of features being maintained at the end. The sequence for ranking approaches applied in the pipeline, on the other hand, is critical to ensuring that the significant genes are kept in the final subset. Out of four gene ranking methodologies, twenty-four separate pipeline models are developed during this experimental investigation. To discover the best pipeline for a given task, these pipelines are tested against seven distinct microarray databases. The Nemenyi post hoc hypothetical test confirms the grading system's result that a pipeline model is noteworthy.

Rasmita Dash et al. [11] offer an approach for microarray data Multi-Objective Feature Selection and Classifier Ensemble (MOFSCE), which works in two phases. The first phase is a pre-processing phase in which the Pareto front is utilized to identify relevant genes using a bi-objective optimization approach. In their study, 21 Bi-Objective Feature Selection (BOFS) models are created using seven feature ranking methodologies. The BOFS model's performance varies based on the dataset. As a consequence, the grading system is used to determine the stability of the BOFS models. The construction of a classifier ensemble, which obtains the selected characteristics from the identified BOFS model, is the second phase.

Mitsunori Ogihara et al. [12] have presented a comparative analysis of feature selection on gene expression data and multiclass categorization. The research offers eight feature selection approaches, which according to the findings are information gain, the towing rule, sum minority, max minority, Gini index, the sum of variances, one-dimensional SVM, and t-statistics. They have evaluated the feature's usefulness by evaluating the level of class predictability when the prediction is made by splitting a gene's whole range of expression into two sections. The results are typically satisfactory for datasets with a modest number of classes. Prediction accuracy is much worse for datasets with a high number of classes.

In another study, Mehdi Pirooznia et al. [13] examined the effectiveness of classification algorithms such SVM, RBF Neural Nets, MLP Neural Nets, Bayesian, Decision Tree, and Random Forest. k-fold cross-validation was used to calculate the accuracy. The efficacy of certain standard clustering approaches, such as K-means, DBC, and expectation

maximization (EM) clustering, has been tested on the datasets. Support Vector Machine Recursive Feature Elimination (SVM-RFE), Chi-squared, and CSF have been used to compare the efficiency of feature selection approaches. In each example, these approaches are used on eight separate binary class microarray datasets. By observing the increasing performance of the work completed by Rasmita Dash et al. [11], who used a three-stage dimensionality reduction strategy on microarray databases, as well as four distinct classifiers. In the first stage, statistical techniques are utilized to filter out irrelevant genes from the database. Thus, approximately 90% less significant characteristics are deleted. SNR is employed in the second step to drop a group of very noisy genes. Finally, the PCA approach is utilized to further reduce the dimension in the final stage. Then, these compressed data are evaluated using ANN, MLR, naïve Bayesian, and k-NN classifiers.

Alok Sharma et al. [14] also worked for feature selection by using transcriptome data for a classification problem. The findings of comparison investigations conducted by Changjing Shang and Qiang Shen et al. [15] highlight the relevance of appropriate feature selection in the creation of classifiers for usage in high-dimensional domains. The optimal classification performance for k-NN and Naive Bayes classifiers is attained using a subset of features determined by information gain ranking after a huge corpus of systematic tests. Naive Bayes may also perform well with a modest collection of linearly processed primary features in categorizing this challenging dataset. In addition, feature selection enhances classification accuracy while enhancing computing efficiency.

Mahnaz Behroozi and Ashkan Sami [16] looked at a dataset with a variety of sound recordings. The key contribution is to suggest a new separate classification framework that recommends using a unique classifier for each type of voice sample and presenting which vocal tests are more representative. They employed pre-processed data that was classified using four different algorithms: k-NN, SVM, discriminant analysis, and Naive Bayes. The k-NN classifier was built using the Euclidean distance metric, with k values of 1, 3, 5, and 7. With a scaling factor sigma($\sum$) of 3 and a penalty parameter (C) of 1, the SVM classifier was utilized with linear and radial basis kernels (RBF).

Huijuan Lu et al. [17] devised a hybrid feature selection approach that combines two gene algorithms. According to experimental results, the suggested MIMAGA-Selection approach greatly decreases the dimension of the gene expression dataset. The reduced gene expression dataset delivers superior classification accuracy when compared to traditional feature selection techniques. Thanyaluk Jirapech-Umpai and S. Aitken [18] proposed a technique to create classification models using microarray data using both supervised and unsupervised classifiers. The study focuses on the supervised classification problem in which data samples are assigned to a known class. The k-NN classifier is used in this investigation. k-NN classification is based on a distance function determined for pairs of samples in N-dimensional space, such as the Euclidean distance or Pearson's correlation. The class memberships of each sample's k closest neighbors, as calculated by the distance function, are used to classify it. Motivated by the above literature survey, different ideas are taken for feature ranking, classification, and multi-criteria decision-making methods in the different datasets. Hence, the purpose of this study is to assess the classification accuracy by finding out which feature ranking works better in sequencing the feature selection process for the classification of microarray samples. Here, sequencing of the feature selection is taken into consideration, which is a prerequisite for classification [19]. This reduction combination is evaluated concerning multiple performance metrics. In the final stage of implementation, TOPSIS is used to prove the outcome of this analysis.

3. Proposed Work

A rank-based approach is one of the dominant feature selection approaches in high-dimensional data analysis. This approach awards a rank based on one mathematical score to all the features in the original data. Top-ranked features are assumed to be highly informative, and a few of them are selected in the descending order of their rank. Each ranking algorithm is unique, which focuses on the score based on the ranking criteria.

However, feature selection based on a single ranking criterion for any dataset may not result in satisfactory prediction. Hence, different ranking techniques are assigned to various gene-based sequencing datasets on this ranking technique's-based feature evaluation scheme [20]. As a result, the informative gene sequence in one approach may not be the same in another. Thus, a filter technique that is useful in some problem spaces may not be successful for all datasets that refer to different applications. Hence, one ranking technique may outperform others for a specific type of problem. Thus, in this study, in place of extraction of a few top-ranked features using a single ranked-based filter technique, the merits of sequencing various filter techniques are considered. Furthermore, these genes are passed through a sequence of filter ranking approaches. In every stage of filtration, few highly ranked features are taken to the next level, and the rest are dropped.

The proposed model, which is shown in Figure 1, presents the high-level overall workflow that contains the highly spaced microarray database. It is scaled or normalized and validated with k-fold cross validation having $(k - 1)$ fold for training and 1 fold for testing. After validating the data properly, then it is passed through a block, where the proposed sequence of the feature ranking process with different classification is performed in a proper manner. That block of the proposed model is described briefly, as shown in the Figure 2 where sixteen feature reduction (FR) sequences are designed from the four feature ranking techniques known as reduction combinations (RC) i.e. RC1–RC16. For the sequencing of feature ranking, four ranking techniques are taken into consideration, which are in order and represented as follows: FR1: Pearson Correlation Coefficient-Based Feature Selection, FR2: Chi-Square Test, FR3: Information Gain, and FR4: Relief Method. In every stage of reduction, 80% lower-ranked genes are dropped, and thus, the reduced dataset at different levels are formulated. It is very difficult to evaluate which ranking technique works well on a dataset for a specific classifier. Hence, to design the most superior RC for highly spaced gene expression data, a multi-criteria-based decision-making technique was applied in which TOPSIS is implemented considering Accuracy, Specificity, Sensitivity, Jaccard, and Gmean as five evaluation techniques. The proposed model is presented in Figure 1 in two ways, where Figure 1 explains the high-level overall flow and shows the overall model design, and Figure 2 explains the details of proposed work with the Multi-Attribute Decision-Making (MADM) process for obtaining a better sequence and classifier combination.

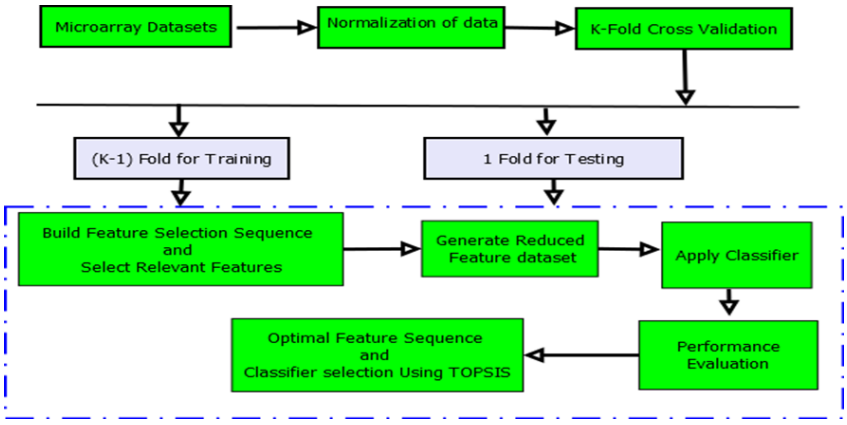

Figure 1. High-Level Flow Diagram of the Proposed Model.

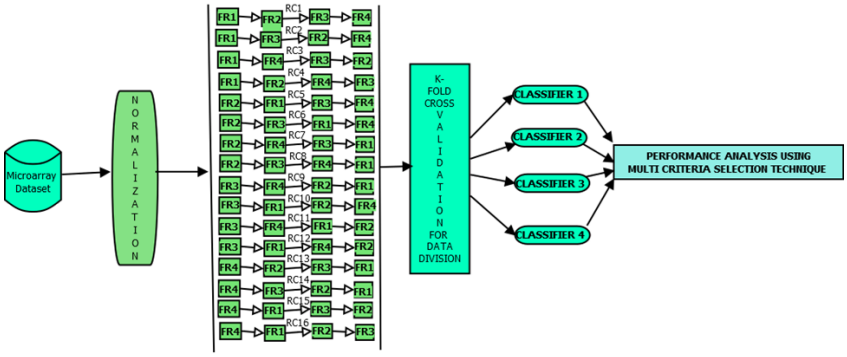

Figure 2. Detailed Flow Diagram of the Proposed Model.

4. Methodology

This section provides a brief overview of the machine learning techniques and tools used for data pre-processing, data classifications, performance metrics, and the TOPSIS approach.

4.1. Normalization

Normalization is often necessary when dealing with attributes on multiple scales [21]; otherwise, it may lead to a weakening in the impact of an equally essential attribute (on a smaller scale) due to other qualities having values on a greater scale. Generally, normalization is of three types: min–max normalization, Z-score normalization, and decimal scaling. In this research, min–max normalization is used and is discussed as follows.

The min–max normalization method is used for normalizing data. This technique can result in a poor data model when multiple attributes exist with different scale values when performing data mining operations, as shown in Equation (1). So, this technique is used for the implementation of this work for normalizing the data.

$$new\_V_{i,j} = (V_{i,j} - min\_A_j) / (max\_A_j - min\_A_j) \quad (1)$$

where $V_{i,j}$ is the original value for an instance of attribute j of record i. $new_V_{i,j}$is the new value. min_A_j is the minimum value of the attribute j in the original dataset (A). max_A_j is the maximum value of the attribute j in the original dataset (A).

4.2. Feature Ranking Techniques Used

Methods for extracting a subset of features from a larger dataset are known as feature selection methods. Furthermore, feature selection methods provide a feature subset from the raw dataset. It is divided into three types, i.e., filter approach, wrapper approach, and embedded approach. Feature ranking includes rating each feature using a specific approach and then picking genes based on their weights [22]. Each method employed allowed for the selection of a limited number of features. Some of the papers used different successful feature ranking techniques in a better way [4,11]. Filtering techniques use an easy-to-calculate measure to quickly rank features, with the highest-ranking features picked [23]. Here, four feature selection techniques are taken into consideration as in Table 1, i.e., CBFS, CST, InG, and RFS. CBFS [2] is a multi-variant filter technique that ranks features based on the correlation between performance assessment functions. It begins with a full set of features (genes). CBFS focuses on decreasing feature-to-feature correlations while improving feature-to-class correlations. The Chi-squared feature evaluation [24] merely displays the relative relevance of the original characteristics. Then, the user may choose which features to keep and which to reject based on this information. In Chi-squared feature selection, this test statistic between the feature and the target class is used to establish the

relevance of a feature. Equation (2) is used to construct the Chi-squared statistic, where the feature and class had no connection.

$$X^2 = \sum \frac{(A_i - B_i)^2}{B_i} \tag{2}$$

where A_i is the observed value and E_i is the expected value. The information gain serves as a simple initial filter for screening features. The Information $GainInG(C, B)$ [25] of a feature B, relative to a set of data C, is defined as:

$$InG(C, B) = E(C) - \sum_{x \epsilon Y(B)} \frac{|C_x|}{|C|} \tag{3}$$

In Equation (3), $Y(B)$ is the set of all possible values of feature Band C_x, the subset of C, for which feature B has the value of x. The first term is the entropy of the entire set. Given a randomly picked instance, Relief [26] searches for k of its nearest neighbours from the same class, which is referred to as nearest hits H, as well as k nearby neighbors from each of the distinct classes, which are referred to as nearest misses M. Relief selection computes feature relevance by showing the relationships between features and class labels. This approach, similar to nearest neighbor algorithms, applies weights to features based on the same-class and different-class examples that are nearest to every sample in the dataset. The adaptive formula for finding feature relevance is shown in Equation (4).

$$X_j = X_{j-1} - (Y_j - NH_j)^2 + (Y_j - NM_j)^2 \tag{4}$$

where X is an n-dimensional weight vector with n features. The closest same class and different class samples are represented by NH as 'NearHit' and NM as 'NearMiss'. j shows the number of iterations in the algorithm.

4.3. Classification Model

In this subsection, the authors have presented k-NN, LR, DT, and RF as classifiers [27] to classify the four cancer datasets. The summary of the classification techniques and performance evaluation matrix are described in the following section.

4.3.1. k-Nearest Neighbor (k-NN)

k-NN [28,29] chooses the class value of a new instance by examining a set of the k closest instances, as shown in Equation (6) in the training set and selecting the most frequent class value among them, with k set to five and Euclidean distance matrices used to calculate the similarity between two points. It stores the query data based on a similarity measure and the training data. k-NN parameter tuning is performed to improve the performance by selecting an appropriate value of k.

$$D(x, y) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2} \tag{5}$$

4.3.2. Logistic Regression (LR)

The LR classification model is a prominent option for modeling binary classifications [29]. LR creates a predictor variable by linearly combining the feature variables. Then, a logistic function is used to convert the values of this predictor variable into probabilities. Generally, this method is used for binary class prediction. It can also be applied to multiclass problems. This classification model's logistic equation is:

$$Y_i = \ln\left(\ln\left(\frac{x_i}{1 - x_i}\right)\right) \tag{6}$$

where X_i is the probability of the occurrence of event i.

4.3.3. Decision Tree (DT)

The DT may handle classification and regression issues to solve the classification problems [30]. It has two advantages:

- Decision Trees are made to resemble human decision-making abilities, making them simple to understand.
- Due to the tree-like structure of the decision tree, the logic behind the concept can be easily understood.

The Decision Tree is made up of three types of nodes: the first one is for decision making (commonly represented by a square), the second is for shaping the options (commonly represented by a circular pattern), and the last one is for representing the action (commonly represented by a triangle).

4.3.4. Random Forest (RF)

RF is a classification process that commonly employs an ensemble method, which utilizes multiple decision trees to classify data. It generates bootstrap templates from the Random Forest's original data, and for each bootstrap template, it grows a regression tree or raw classification. Rather than selecting only the best predictors for disclosure, it takes into account every node. It employs a random predictor selection and chooses the best separation among them. The parameter n-split, which specifies the number of splitting points to be evaluated for each feature, is required by RF [31]; higher values of n-split may result in more accurate predictions at the expense of increased computational load. We choose three values while leaving the other parameters at their default settings.

4.4. k-Fold Cross-Validation Method

The error rate of the classification algorithm is used to evaluate a classifier's performance. The error rate on the data that is used to train the classifier (training set) is not a trustworthy criterion. Indeed, such a method could cause the classifier to overfit the training data. To anticipate a classifier's performance, we must examine its error rate on a separate dataset that was not included in the training process (i.e., the test set). The k-fold cross-validation method is based on dividing the dataset into k sections at random. When the $k-1$ remainders are used for training, one portion is used for testing. This technique is performed k times to ensure that each part is tested only once. Then, the k-error estimates are averaged to produce a reliable overall error estimate. Varied k-fold cross-validation trials can result in different classification error rates due to the random selection of folds [32]. As a result, we repeat the k-fold cross-validation process k times to increase the error rate's reliability.

4.5. Performance Evaluation Criteria

At the final stage, the performance of each method was evaluated to determine which method could produce the best results [11]. To evaluate each of the methods used in this study, we used the parameters such as accuracy (Equation (7)), sensitivity (Equation (8)), specificity (Equation (9)), jaccard (Equation (10)), and gmean (Equation (11)) from the confusion matrix. The confusion matrix includes the terms FN (False Negative), FP (False Positive), TN (True Negative), and TP (True Positive). The definition of all performance metrics are as follows:

$$Accuracy = \left(\frac{TP_i + TN_i}{total\ number\ of\ sample}\right) * 100 \tag{7}$$

$$Sensitivity = \left(\frac{TP_i}{TP_i + FP_i}\right) * 100 \tag{8}$$

$$Specificity = \left(\frac{TN_i}{TN_i + FN_i}\right) * 100 \tag{9}$$

$$Jaccard = (\frac{TP_i}{TP_i + FP_i + FN_i}) * 100 \tag{10}$$

$$Gmean = \sqrt{Specificity * Sensitivity} \tag{11}$$

4.6. Multi-Criteria Decision Making (MCDM)

There are six steps involves in multi-criteria decision-making: (i) problem formulation, (ii) identification of requirements, (iii) goal setting, (iv) identification of various alternatives, (v) development of criteria, and (vi) identification and application of decision-making techniques. Some frequently referred multi-criteria techniques (also known as Multi-Attribute Decision Making (MADM)) are the Simple Additive Weighting (SAW), Gray Relational Model (GRM), and TOPSIS. TOPSIS is used to recommend one or more options from a large set of alternatives. The ranking of TOPSIS techniques was calculated using Microsoft Excel as a tool. TOPSIS can be used in situations where there are multiple feature selection algorithms to choose from, each with its own set of criteria such as accuracy, sensitivity, specificity, number of features, and so on [33]. So, TOPSIS techniques used for measurement are as follows:

Step 1: Firstly, create a matrix $M_{i,j}$ with 'm' number of rows corresponding to each feature reduction sequence and 'n' number of columns corresponding to the performance evaluation criteria and the number of classifiers used.

Step 2: For calculating normalized matrix $NM_{i,j}$:

$$NM'_{i,j} = \frac{NM_{i,j}}{\sqrt{\sum_{i=1}^{n}(NM_{i,j})^2}}, j = 1, \ \ldots, \ J \tag{12}$$

Step 3: For calculating weighted normalized matrix $WNM_{i,j}$:

$$WNM_{i,j} = NM'_{i,j} * W_j, \quad j = 1, \ \ldots, \ J \tag{13}$$

where W_j is the weight of the criterion and $\sum_{j=1}^{J} W_j = 1$, weights can be assigned randomly or according to the criteria.

Step 4: Then, calculate the ideal best solution from the combination of the best performance values as $V_j^+$and ideal worst from the combination of the worst performance values as V_j^- using the following formula:

$$V_j^+ = \{WNM_1^+, L, WNM_j^+\} = \{(max_i WNM_{i,j} \mid j \epsilon h), (max_i WNM_{i,j} \mid j \epsilon l)\} \tag{14}$$

$$V_j^- = \{WNM_1^-, L, WNM_j^-\} = \{(max_i WNM_{i,j} \mid j \epsilon h), (min_i WNM_{i,j} \mid j \epsilon l)\} \tag{15}$$

where h is the set of best performance values and l is is the set of worst performance values.

Step 5: After calculating the Ideal Best, calculate the separation measure from the Ideal Best:

$$S_i^+ = \sqrt{\sum_{j=1}^{m}(WNM_{i,j} - WNM_j^+)^2}, \ j = 1, \ \ldots, \ J \tag{16}$$

After calculating the Ideal Worst, calculate the separation measure from the Ideal Worst:

$$S_i^- = \sqrt{\sum_{j=1}^{m}(WNM_{i,j} - WNM_j^-)^2}, \ j = 1, \ \ldots, \ J \tag{17}$$

Step 6: Finally, calculate the relative closeness to the ideal solution performance score as follows:

$$P_i = (\frac{S_i^-}{S_i^+ + S_i^-}), \ j = 1, \ \ldots, \ M \tag{18}$$

5. Experimental Results

Simulation is conducted to consider a maximum of 10 numbers of genes for classification purposes, and an optimal set is extracted. Considering the reduced data, the 10 cross-validation process is implemented to come up with the training and testing data input for the classifiers. For this analysis, a few successful classification techniques for highly spaced data are taken such as k-NN, LR, DT, and RF [34]. From the classification output, it is observed that the performance of RCs varies with the order of the sequence, as these ranking techniques perform differently for different datasets and classifiers.

5.1. Dataset Description

Microarray data are high-dimensional data with a small number of samples or observations compared to the number of genes or attributes. The number of samples is in the range of hundreds, and the number of attributes is in the range of thousands. Among four datasets, Colon Cancer (D2) and Adenoma Cancer (D4) have two class levels, but another two datasets such as Brain Tumor (D1) has five class levels, and Breast Cancer (D3) has three class levels. Table 2 contains descriptions of the databases as follows.

Table 2. Description of microarray databases.

Dataset	Total No. of Samples	No. of Samples in Each Class	No. of Genes	No. of Classes
Brain Tumor [35]	40	10	7129	5
		10		
		10		
		4		
		6		
Colon Cancer [34]	62	40	2000	2
		22		
Breast Cancer [34]	98	11	1213	3
		51		
		36		
Adenoma Cancer [36]	8	4	7086	2
		4		

5.2. Result Analysis

Before going to feature sequencing, first, datasets are normalized by min–max normalization, the values are converted to between 0 and 1 for each feature. After that, feature sets are reduced by 20% in each of the four steps adopted in the sequence. Further simulation is conducted to come up with an optimal set of features, fixing it within the number of 10. Then, the reduced feature sequences are validated by the 10-fold cross-validation approach by separating the training and testing data with respect to the considered classifiers; the classification outputs are presented in Tables 3 and 4. As a result of analyzing Tables 3 and 4, it is clear that none of the classifiers obtain the optimal results across all metrics, and the ranking of the top-performing model differs depending on the performance assessment measurement chosen. Hence, further analysis of classifiers and the FR sequencing approaches is performed by TOPSIS. Table 5 represents the result of the TOPSIS approach implemented on the Brain Tumor dataset, where rows are representing 16 number feature reduction sequences from FR1 to FR16 and columns are representing classifiers used such as k-NN, LR, DT, and RF with five performance criteria: accuracy (CR1), sensitivity (CR2), specificity (CR3), jaccard (CR4), and gmean (CR5). The classifica-

tion result on each dataset is given in Table 2 and converted to the corresponding weighted normalized matrix by using Equation (15) for each dataset, as shown in Tables 5–8. Then, the ideal best value v^+ is chosen from the set of combinations of best performance values for each dataset individually, and the ideal worst value v^- can also be chosen from the set of worst performance values, as given in Equations (14) and (15). The Ideal Best and Ideal Worst solution separation measure or Euclidean distance can be measured by using Equations (16) and (17), whose results are shown on the column as S_i^+ and S_i^- individually for each dataset in Tables 5–8. Finally, the ideal solution performance score can be found out by using Equation (18), which is represented as P_i for ranking of each feature reduction sequences individually with referencing to the datasets shown in Tables 5–8. As described above, Tables 6–8 represent the TOPSIS result analysis for the Colon Cancer, Breast Cancer, and Adenoma Cancer datasets. After obtaining the ranks of each FR, the performances of all FR are compared in Table 9, where the top three ranked feature reduction sequences with respect to all classifiers are taken into consideration. As it is difficult to choose which sequence performs better, the occurrence of the few top-ranking FR is also calculated in Table 10, from which it is observed that FR5 is superior. FR5 is coming in first rank expect for one classifier (i.e., LR), where the rank is second. Finally, in Figure 3, the number of occurrences of all FR coming in the top three ranking is shown. Here, it is also found that out of 48 occurrences (as in Table 9), FR5 is coming under the top three rank in 16 occurrences. However, for other FR cases, the occurrence is quite nominal. Hence, it can be said that the FR5 model can work better as the feature reduction sequence given for four datasets, as shown in Figure 4.

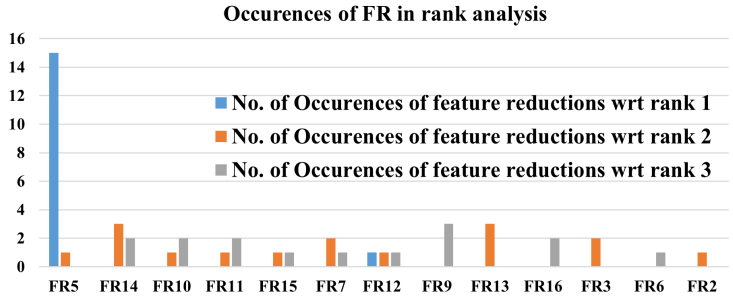

Figure 3. Occurrence Analysis for Feature Reduction Technique.

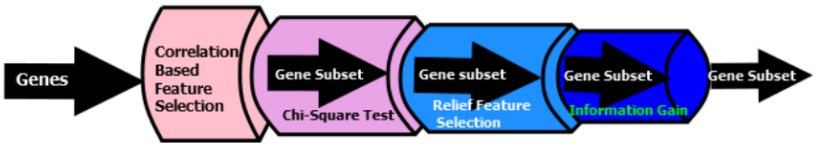

Figure 4. FR5 Model for Microarray Data.

Table 3. Classification Results of Four Datasets By Using k-NN, LR, DT, and RF Classifier with Different Feature Ranking Techniques.

		CR1				**CR2**				**CR3**				**CR4**				**CR5**			
		k-NN	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**
D1	FR1	0.89	0.92	0.91	0.89	0.78	0.91	0.92	0.91	0.97	0.91	0.85	0.93	0.89	0.94	0.91	0.94	0.87	0.91	0.88	0.92
	FR2	0.91	0.94	0.92	0.88	0.9	0.92	0.94	0.91	0.95	0.88	0.97	0.89	0.95	0.86	0.92	0.9	0.92	0.90	0.95	0.90
	FR3	0.94	0.93	0.92	0.92	0.88	0.92	0.9	0.91	0.9	0.92	0.93	0.94	0.94	0.93	0.91	0.9	0.89	0.92	0.91	0.92
	FR4	0.92	0.9	0.91	0.9	0.91	0.95	0.94	0.89	0.92	0.78	0.87	0.88	0.91	0.89	0.91	0.89	0.91	0.86	0.90	0.88
	FR5	0.97	0.95	0.91	1	0.99	0.96	1	0.91	0.98	0.95	0.99	1	1	0.99	0.97	0.98	0.98	0.95	0.99	0.95
	FR6	0.96	0.94	0.93	0.92	0.97	0.91	0.88	0.78	0.91	0.92	0.89	0.9	0.98	0.89	0.92	0.91	0.94	0.91	0.88	0.84
	FR7	0.9	0.91	0.91	0.89	0.92	0.93	0.89	0.91	0.93	0.97	0.98	0.82	0.97	1	0.9	0.89	0.92	0.95	0.93	0.86
	FR8	0.95	0.91	0.88	0.9	0.91	0.89	0.92	0.93	0.91	0.87	0.89	0.91	0.93	0.94	0.96	0.97	0.91	0.88	0.90	0.92
	FR9	0.94	0.91	0.92	0.93	0.92	0.91	0.93	0.88	0.9	1	0.93	0.91	0.9	1	0.95	0.99	0.91	0.95	0.93	0.89
	FR10	0.97	0.91	0.89	0.87	0.98	0.92	0.88	0.91	0.94	0.92	0.8	0.83	0.93	0.94	0.89	0.98	0.96	0.92	0.84	0.87
	FR11	0.95	0.94	0.93	0.9	0.93	0.91	0.85	0.93	0.92	0.95	0.92	0.87	0.94	0.96	0.91	0.92	0.92	0.93	0.88	0.90
	FR12	0.96	0.97	0.91	0.89	0.96	0.98	0.91	0.92	0.94	0.96	0.93	0.92	0.91	0.97	0.93	0.93	0.95	0.97	0.92	0.92
	FR13	0.99	0.93	0.92	0.93	1	0.97	0.93	0.95	0.89	0.88	0.99	0.89	0.94	0.95	0.97	0.9	0.94	0.92	0.96	0.92
	FR14	1	0.89	0.95	0.91	1	0.9	0.97	0.96	0.95	0.9	0.94	0.94	0.95	0.89	0.91	1	0.97	0.90	0.95	0.95
	FR15	0.97	0.94	0.92	0.94	0.9	0.91	0.89	0.92	0.88	0.9	0.96	0.97	0.89	0.93	0.95	0.98	0.89	0.90	0.92	0.94
	FR16	0.96	0.95	0.89	0.93	0.97	0.94	0.9	0.91	0.9	0.92	0.8	0.91	0.97	0.96	0.95	0.94	0.93	0.93	0.85	0.91
D2	FR1	0.98	0.94	0.89	0.91	0.91	0.94	0.89	0.79	0.89	0.92	0.91	0.89	0.91	0.88	0.92	0.93	0.90	0.93	0.90	0.84
	FR2	0.97	0.95	0.9	0.92	0.95	0.91	0.91	0.89	0.92	0.95	0.93	0.88	0.97	0.95	0.94	0.93	0.93	0.93	0.92	0.88
	FR3	0.92	0.92	0.91	0.89	0.9	0.89	0.88	0.95	0.93	0.91	0.9	0.91	0.89	0.92	0.91	0.93	0.91	0.90	0.89	0.93
	FR4	0.99	0.93	0.87	0.93	0.91	0.94	0.91	0.93	0.95	0.92	0.88	0.87	0.93	0.94	0.91	0.89	0.93	0.93	0.89	0.90
	FR5	1	0.99	0.92	0.93	1	1	0.99	0.98	0.98	0.99	0.99	1	0.99	1	1	0.99	0.99	0.99	0.99	0.99
	FR6	1	0.98	0.88	0.93	0.95	0.93	0.92	0.91	0.95	0.93	0.94	0.89	0.91	0.97	0.93	0.95	0.95	0.93	0.93	0.90
	FR7	0.96	0.95	0.93	0.91	0.86	0.88	0.87	0.93	0.91	0.94	0.93	0.92	0.92	0.93	0.95	0.98	0.88	0.91	0.90	0.92
	FR8	0.99	0.97	0.88	0.89	0.94	0.91	0.91	0.92	0.96	0.9	0.87	0.95	0.91	0.92	0.97	0.94	0.95	0.90	0.89	0.93
	FR9	0.99	0.96	0.92	0.92	0.95	0.93	0.92	0.89	0.95	0.94	0.97	0.9	0.89	0.96	0.97	0.91	0.95	0.93	0.94	0.89
	FR10	0.98	0.98	0.94	0.91	0.95	0.93	0.92	0.93	0.97	0.98	0.91	0.89	0.95	0.96	0.92	0.93	0.96	0.95	0.91	0.91
	FR11	0.97	0.94	0.91	0.89	0.98	0.97	1	0.94	0.9	0.95	0.94	0.96	0.92	0.96	0.97	0.89	0.94	0.96	0.97	0.95
	FR12	0.99	0.98	0.89	0.92	0.95	0.96	0.94	0.91	0.96	0.98	0.89	0.92	0.96	0.95	0.92	0.89	0.95	0.97	0.91	0.91
	FR13	0.96	0.95	0.93	0.94	0.92	0.78	0.92	0.91	0.99	0.98	0.93	0.89	0.93	0.96	0.89	0.91	0.95	0.87	0.92	0.90
	FR14	0.97	0.96	0.93	0.78	1	0.97	0.94	0.92	0.99	0.91	0.97	0.98	0.99	0.95	0.93	1	0.99	0.94	0.95	0.95
	FR15	0.98	1	0.92	0.87	0.96	0.97	0.96	0.95	0.98	0.96	0.95	0.93	0.89	0.91	0.94	0.96	0.97	0.96	0.95	0.94
	FR16	1	0.99	0.94	0.89	0.96	0.95	0.89	0.94	0.95	0.91	0.93	0.94	0.98	0.97	0.95	0.93	0.95	0.93	0.91	0.94

Table 4. Classification Results of Four Datasets By Using k-NN, LR, DT, and RF Classifier with Different Feature Ranking Techniques cont.

		CR1				**CR2**				**CR3**				**CR4**				**CR5**			
		k-NN	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**	**k-NN**	**LR**	**DT**	**RF**
D3	FR1	0.94	0.93	0.88	0.91	0.91	0.91	0.9	0.89	0.92	0.92	0.91	0.95	0.93	0.91	0.9	0.89	0.91	0.91	0.90	0.92
	FR2	0.93	0.92	0.91	0.9	0.94	0.96	0.98	0.97	0.91	0.94	0.97	0.96	0.93	0.91	0.98	0.9	0.92	0.95	0.97	0.96
	FR3	0.89	0.91	0.92	0.87	0.92	0.91	0.9	0.92	0.93	0.89	0.96	0.91	0.94	0.93	0.94	0.93	0.92	0.90	0.93	0.91
	FR4	0.94	0.93	0.88	0.92	0.97	0.96	0.89	0.9	0.91	0.93	0.94	0.92	0.96	0.98	0.92	0.89	0.94	0.94	0.91	0.91
	FR5	0.95	0.92	0.91	0.9	1	1	0.98	0.99	0.99	1	1	0.98	1	1	0.99	0.98	0.99	1.00	0.99	0.98
	FR6	0.97	0.93	0.89	0.89	0.96	0.9	0.95	0.97	0.91	0.99	0.96	0.97	0.98	0.91	0.93	0.95	0.93	0.94	0.95	0.97
	FR7	0.94	0.89	0.9	0.91	0.93	0.94	0.9	0.91	0.98	0.95	0.96	0.97	0.98	0.97	0.95	0.94	0.95	0.94	0.93	0.94
	FR8	0.95	0.91	0.92	0.89	0.95	0.91	0.88	0.92	0.94	0.95	0.97	0.92	0.92	0.89	0.95	0.94	0.94	0.93	0.92	0.92
	FR9	0.96	0.97	0.89	0.91	0.95	0.98	0.97	0.94	0.94	0.91	0.94	0.95	0.97	0.93	0.95	0.93	0.94	0.94	0.95	0.94
	FR10	0.97	0.95	0.93	0.92	0.94	0.98	0.89	0.92	0.97	0.93	0.94	0.92	0.95	0.96	0.93	0.92	0.95	0.95	0.91	0.92
	FR11	0.96	0.88	0.86	0.93	0.98	0.97	0.94	0.96	0.95	0.94	0.93	0.9	0.95	0.94	0.92	0.91	0.96	0.95	0.93	0.93
	FR12	0.94	0.93	0.92	0.91	0.99	1	1	0.9	0.96	0.97	0.91	0.89	0.9	0.89	0.93	0.94	0.97	0.98	0.95	0.89
	FR13	0.92	0.97	0.93	0.93	0.99	0.93	0.91	0.99	0.99	0.93	0.92	0.93	0.93	0.94	0.96	0.96	0.99	0.93	0.91	0.96
	FR14	0.97	0.92	0.93	0.95	0.96	0.89	0.91	0.92	1	0.95	1	0.91	0.97	0.98	0.92	1	0.98	0.92	0.95	0.91
	FR15	0.97	0.92	0.91	0.93	0.93	0.92	0.95	0.94	0.97	0.94	0.92	0.94	0.95	0.9	1	0.93	0.95	0.93	0.93	0.94
	FR16	0.95	0.92	0.91	0.88	0.95	0.93	0.95	0.94	1	0.95	0.89	0.93	0.96	0.97	0.92	0.91	0.97	0.94	0.92	0.93
D4	FR1	0.96	0.94	0.91	0.89	0.92	0.94	0.9	0.94	0.91	0.93	0.9	0.92	0.93	0.95	0.89	0.91	0.91	0.93	0.90	0.93
	FR2	0.96	0.94	0.92	0.9	0.92	0.93	0.94	0.89	0.99	0.96	0.88	0.89	0.91	0.94	0.95	0.95	0.95	0.94	0.91	0.89
	FR3	0.98	0.97	0.91	0.89	0.95	0.99	0.92	0.97	0.93	0.96	0.95	0.97	0.97	0.96	0.93	0.94	0.94	0.97	0.93	0.97
	FR4	1	0.98	0.94	0.92	0.94	0.89	0.89	0.9	0.91	0.97	0.93	0.95	0.95	0.93	0.92	0.89	0.92	0.93	0.91	0.92
	FR5	1	0.98	0.92	0.92	0.97	0.99	1	1	1	0.99	0.99	1	0.99	1	0.98	0.98	0.98	0.99	0.99	1.00
	FR6	1	0.99	0.93	0.94	0.93	0.96	0.94	0.92	0.98	0.93	0.96	0.94	0.92	0.93	0.94	0.95	0.95	0.94	0.95	0.93
	FR7	0.98	0.97	0.93	0.93	0.94	0.95	0.97	0.92	0.99	0.95	1	0.96	0.96	0.99	0.98	0.95	0.96	0.95	0.98	0.94
	FR8	0.99	0.95	0.92	0.91	0.93	0.95	0.94	0.91	0.92	0.94	0.91	0.9	0.95	0.94	0.94	0.92	0.92	0.94	0.92	0.90
	FR9	0.98	0.99	0.94	0.93	0.95	0.89	0.94	0.95	0.95	0.97	0.95	0.93	0.96	1	0.96	0.98	0.95	0.93	0.94	0.94
	FR10	0.98	0.97	0.96	0.92	0.96	0.95	0.94	0.89	0.88	0.89	0.96	0.94	0.91	0.92	0.93	0.95	0.92	0.92	0.95	0.91
	FR11	0.95	0.96	0.93	0.89	0.97	1	0.92	0.91	0.93	0.95	0.92	0.91	0.92	0.95	0.96	0.96	0.95	0.97	0.92	0.91
	FR12	0.98	0.93	0.89	0.9	0.92	0.96	0.92	0.94	0.96	0.92	0.93	0.95	0.92	0.93	0.97	0.94	0.94	0.94	0.92	0.94
	FR13	0.96	0.97	0.92	0.91	0.99	0.93	0.92	0.93	1	0.95	0.97	0.94	0.89	0.92	0.98	0.91	0.99	0.94	0.94	0.93
	FR14	0.98	0.94	0.91	0.93	1	0.96	0.93	0.89	1	0.91	0.98	0.95	0.9	0.93	0.93	1	1.00	0.93	0.95	0.92
	FR15	0.99	0.91	0.93	0.93	0.98	0.95	0.94	0.91	0.97	0.95	0.97	0.95	0.94	0.95	0.92	0.96	0.97	0.95	0.95	0.93
	FR16	1	0.98	0.94	0.91	0.96	0.94	0.95	0.92	0.95	0.97	0.95	0.94	0.91	0.94	0.92	0.96	0.95	0.95	0.95	0.93

Table 5. Results of TOPSIS for Brain Tumor Dataset with Ranking of Feature Reduction Sequence using Dataset D1.

	k-NN									LR								
FR	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**
FR1	0.234	0.209	0.262	0.237	0.234	0.078	0.024	0.237	16	0.248	0.245	0.248	0.250	0.247	0.041	0.045	0.522	10
FR2	0.240	0.241	0.257	0.253	0.249	0.042	0.043	0.507	10	0.253	0.248	0.240	0.229	0.244	0.056	0.033	0.372	15
FR3	0.248	0.236	0.243	0.251	0.240	0.052	0.034	0.394	14	0.251	0.248	0.251	0.247	0.250	0.037	0.047	0.560	7
FR4	0.242	0.244	0.249	0.243	0.246	0.047	0.040	0.457	12	0.243	0.256	0.213	0.236	0.234	0.076	0.018	0.194	16
FR5	0.255	0.265	0.265	0.267	0.265	0.008	0.078	0.904	1	0.256	0.259	0.259	0.263	0.259	0.016	0.068	0.806	2
FR6	0.253	0.260	0.246	0.261	0.253	0.027	0.063	0.701	4	0.253	0.245	0.251	0.236	0.248	0.044	0.044	0.498	11
FR7	0.237	0.246	0.251	0.259	0.249	0.041	0.048	0.539	9	0.245	0.251	0.265	0.266	0.258	0.023	0.069	0.749	4
FR8	0.250	0.244	0.246	0.248	0.245	0.043	0.042	0.492	11	0.245	0.240	0.238	0.250	0.239	0.054	0.033	0.380	14
FR9	0.248	0.246	0.243	0.240	0.245	0.048	0.027	0.357	15	0.245	0.245	0.273	0.266	0.259	0.025	0.075	0.749	3
FR10	0.255	0.262	0.254	0.248	0.258	0.024	0.065	0.735	3	0.245	0.248	0.251	0.250	0.250	0.038	0.048	0.557	8
FR11	0.250	0.249	0.249	0.251	0.249	0.025	0.049	0.665	7	0.253	0.245	0.259	0.255	0.252	0.029	0.058	0.669	5
FR12	0.253	0.257	0.254	0.243	0.256	0.032	0.058	0.648	8	0.261	0.264	0.262	0.258	0.263	0.014	0.072	0.842	1
FR13	0.261	0.268	0.241	0.251	0.254	0.031	0.069	0.687	5	0.251	0.262	0.240	0.252	0.251	0.039	0.047	0.545	9
FR14	0.263	0.268	0.257	0.253	0.262	0.016	0.076	0.827	2	0.240	0.243	0.246	0.236	0.244	0.054	0.035	0.397	13
FR15	0.255	0.241	0.238	0.237	0.240	0.055	0.039	0.414	13	0.253	0.245	0.246	0.247	0.246	0.043	0.042	0.497	12
FR16	0.253	0.260	0.243	0.259	0.252	0.030	0.061	0.671	6	0.256	0.253	0.251	0.255	0.253	0.029	0.054	0.651	6
V+	0.263	0.268	0.265	0.267	0.265					0.261	0.264	0.273	0.266	0.263				
V−	0.234	0.209	0.238	0.237	0.234					0.240	0.240	0.213	0.229	0.234				
	DT									RF								
FR1	0.249	0.251	0.232	0.245	0.241	0.057	0.028	0.332	14	0.244	0.250	0.256	0.250	0.253	0.042	0.054	0.565	8
FR2	0.252	0.256	0.264	0.248	0.261	0.026	0.063	0.709	4	0.241	0.250	0.245	0.239	0.248	0.054	0.044	0.451	12
FR3	0.252	0.246	0.254	0.245	0.250	0.043	0.045	0.513	9	0.252	0.250	0.259	0.239	0.255	0.041	0.056	0.580	6
FR4	0.249	0.256	0.237	0.245	0.247	0.048	0.037	0.435	11	0.246	0.245	0.242	0.237	0.244	0.055	0.038	0.405	14
FR5	0.249	0.273	0.270	0.261	0.272	0.011	0.082	0.882	1	0.274	0.250	0.275	0.261	0.263	0.015	0.081	0.846	1
FR6	0.254	0.240	0.243	0.248	0.242	0.054	0.033	0.378	13	0.252	0.215	0.248	0.242	0.231	0.065	0.026	0.289	16
FR7	0.249	0.243	0.267	0.242	0.255	0.041	0.057	0.584	7	0.244	0.250	0.226	0.237	0.238	0.066	0.037	0.357	15
FR8	0.241	0.251	0.243	0.258	0.247	0.047	0.041	0.464	10	0.246	0.256	0.251	0.258	0.253	0.039	0.058	0.599	4
FR9	0.252	0.254	0.254	0.256	0.254	0.032	0.052	0.618	5	0.255	0.242	0.251	0.263	0.246	0.038	0.051	0.570	7
FR10	0.243	0.240	0.218	0.239	0.229	0.079	0.009	0.098	16	0.238	0.250	0.229	0.261	0.239	0.061	0.044	0.420	13
FR11	0.254	0.232	0.251	0.245	0.241	0.057	0.038	0.400	12	0.246	0.256	0.240	0.245	0.248	0.051	0.048	0.488	11
FR12	0.249	0.248	0.254	0.250	0.251	0.039	0.047	0.545	8	0.244	0.253	0.253	0.247	0.253	0.043	0.054	0.555	10
FR13	0.252	0.254	0.270	0.261	0.262	0.023	0.069	0.752	2	0.255	0.261	0.245	0.239	0.253	0.045	0.058	0.564	9
FR14	0.260	0.265	0.256	0.245	0.261	0.025	0.063	0.713	3	0.249	0.264	0.259	0.266	0.262	0.030	0.074	0.714	3
FR15	0.252	0.243	0.262	0.256	0.252	0.038	0.054	0.589	6	0.257	0.253	0.267	0.261	0.260	0.022	0.071	0.762	2
FR16	0.243	0.246	0.218	0.256	0.232	0.073	0.021	0.227	15	0.255	0.250	0.251	0.250	0.251	0.038	0.052	0.581	5
V+	0.260	0.273	0.270	0.261	0.272					0.274	0.264	0.275	0.266	0.263				
V−	0.241	0.232	0.218	0.239	0.229					0.238	0.215	0.226	0.237	0.231				

Table 6. Results of TOPSIS for Colon Cancer Dataset with Ranking of Feature Reduction Sequence using Dataset D2.

	k-NN									LR								
FR	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**
FR1	0.250	0.241	0.234	0.243	0.238	0.049	0.021	0.304	14	0.244	0.253	0.244	0.233	0.249	0.046	0.046	0.498	12
FR2	0.248	0.252	0.242	0.260	0.247	0.029	0.038	0.564	8	0.247	0.245	0.252	0.251	0.249	0.037	0.045	0.551	11
FR3	0.235	0.238	0.245	0.238	0.242	0.050	0.017	0.252	15	0.239	0.239	0.241	0.243	0.241	0.053	0.032	0.377	15
FR4	0.253	0.241	0.250	0.249	0.246	0.035	0.032	0.473	12	0.242	0.253	0.244	0.248	0.249	0.039	0.049	0.557	10
FR5	0.256	0.265	0.258	0.265	0.262	0.003	0.062	0.955	1	0.257	0.269	0.263	0.264	0.266	0.003	0.080	0.969	1
FR6	0.256	0.252	0.250	0.243	0.251	0.030	0.040	0.570	7	0.255	0.250	0.247	0.256	0.249	0.032	0.052	0.623	8
FR7	0.245	0.228	0.240	0.246	0.234	0.056	0.014	0.201	16	0.247	0.237	0.249	0.246	0.243	0.047	0.034	0.418	14
FR8	0.253	0.249	0.253	0.243	0.251	0.030	0.038	0.555	10	0.252	0.245	0.239	0.243	0.242	0.047	0.040	0.456	13
FR9	0.253	0.252	0.250	0.238	0.251	0.034	0.027	0.445	13	0.249	0.250	0.249	0.254	0.250	0.032	0.051	0.614	9
FR10	0.250	0.252	0.255	0.254	0.254	0.021	0.044	0.678	5	0.255	0.250	0.260	0.254	0.255	0.025	0.057	0.696	4
FR11	0.248	0.260	0.237	0.246	0.248	0.035	0.038	0.522	11	0.244	0.261	0.252	0.254	0.257	0.025	0.062	0.712	3
FR12	0.253	0.252	0.253	0.257	0.252	0.021	0.044	0.681	4	0.255	0.258	0.260	0.251	0.259	0.019	0.063	0.768	2
FR13	0.245	0.244	0.261	0.249	0.252	0.030	0.039	0.560	9	0.247	0.210	0.260	0.254	0.234	0.069	0.031	0.308	16
FR14	0.248	0.265	0.261	0.265	0.263	0.008	0.062	0.889	2	0.249	0.261	0.241	0.251	0.251	0.032	0.058	0.645	6
FR15	0.250	0.254	0.258	0.238	0.256	0.030	0.045	0.598	6	0.260	0.261	0.255	0.240	0.258	0.028	0.063	0.695	5
FR16	0.256	0.254	0.250	0.262	0.252	0.018	0.048	0.721	3	0.257	0.255	0.241	0.256	0.249	0.032	0.057	0.641	7
V+	0.256	0.265	0.261	0.265	0.263					0.260	0.269	0.263	0.264	0.266				
V−	0.235	0.228	0.234	0.238	0.234					0.239	0.210	0.239	0.233	0.234				
	DT									RF								
FR1	0.244	0.241	0.245	0.245	0.243	0.051	0.016	0.236	14	0.252	0.215	0.242	0.249	0.228	0.075	0.038	0.335	16
FR2	0.247	0.246	0.251	0.250	0.248	0.040	0.026	0.397	8	0.255	0.242	0.239	0.249	0.241	0.053	0.050	0.485	15
FR3	0.250	0.238	0.242	0.242	0.240	0.055	0.015	0.213	15	0.247	0.258	0.247	0.249	0.253	0.038	0.061	0.612	7
FR4	0.239	0.246	0.237	0.242	0.242	0.055	0.012	0.183	16	0.258	0.253	0.236	0.238	0.245	0.054	0.059	0.521	13
FR5	0.253	0.268	0.267	0.266	0.267	0.006	0.062	0.910	1	0.258	0.267	0.272	0.265	0.269	0.004	0.090	0.959	1
FR6	0.242	0.249	0.253	0.248	0.251	0.039	0.028	0.416	7	0.258	0.248	0.242	0.254	0.245	0.045	0.058	0.562	9
FR7	0.255	0.235	0.251	0.253	0.243	0.048	0.028	0.371	11	0.252	0.253	0.250	0.262	0.252	0.033	0.064	0.661	2
FR8	0.242	0.246	0.234	0.258	0.240	0.052	0.024	0.316	13	0.247	0.250	0.258	0.251	0.254	0.034	0.059	0.639	5
FR9	0.253	0.249	0.261	0.258	0.255	0.027	0.042	0.607	5	0.255	0.242	0.244	0.243	0.243	0.051	0.051	0.498	14
FR10	0.258	0.249	0.245	0.245	0.247	0.042	0.028	0.397	9	0.252	0.253	0.242	0.249	0.247	0.044	0.057	0.563	8
FR11	0.250	0.271	0.253	0.258	0.262	0.019	0.051	0.735	2	0.247	0.256	0.261	0.238	0.258	0.038	0.064	0.631	6
FR12	0.244	0.254	0.240	0.245	0.247	0.045	0.023	0.337	12	0.255	0.248	0.250	0.238	0.249	0.046	0.056	0.549	11
FR13	0.255	0.249	0.251	0.237	0.250	0.044	0.028	0.394	10	0.260	0.248	0.242	0.243	0.245	0.049	0.058	0.541	12
FR14	0.255	0.254	0.261	0.248	0.258	0.027	0.042	0.608	4	0.216	0.250	0.266	0.267	0.258	0.049	0.063	0.562	10
FR15	0.253	0.260	0.256	0.250	0.258	0.025	0.042	0.628	3	0.241	0.258	0.253	0.257	0.256	0.033	0.062	0.653	3
FR16	0.258	0.241	0.251	0.253	0.246	0.042	0.031	0.420	6	0.247	0.256	0.255	0.249	0.256	0.033	0.062	0.650	4
V+	0.258	0.271	0.267	0.266	0.267					0.260	0.267	0.272	0.267	0.269				
V−	0.239	0.235	0.234	0.237	0.240					0.216	0.215	0.236	0.238	0.228				

Table 7. Results of TOPSIS for Breast Cancer Dataset with Ranking of Feature Reduction Sequence using Dataset D3.

	k-NN									LR								
FR	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**
FR1	0.248	0.238	0.241	0.244	0.240	0.043	0.016	0.267	15	0.251	0.241	0.244	0.242	0.243	0.047	0.018	0.275	15
FR2	0.245	0.246	0.238	0.244	0.242	0.040	0.016	0.281	14	0.249	0.254	0.249	0.242	0.252	0.036	0.029	0.446	11
FR3	0.235	0.241	0.243	0.247	0.242	0.042	0.012	0.225	16	0.246	0.241	0.236	0.248	0.239	0.052	0.014	0.216	16
FR4	0.248	0.254	0.238	0.252	0.246	0.032	0.027	0.457	11	0.251	0.254	0.246	0.261	0.250	0.029	0.037	0.563	4
FR5	0.251	0.262	0.259	0.263	0.261	0.006	0.049	0.892	1	0.249	0.265	0.265	0.266	0.265	0.014	0.058	0.811	1
FR6	0.256	0.251	0.238	0.257	0.245	0.031	0.033	0.518	9	0.251	0.238	0.262	0.242	0.250	0.040	0.032	0.447	10
FR7	0.248	0.244	0.257	0.257	0.250	0.024	0.033	0.581	6	0.240	0.249	0.252	0.258	0.250	0.034	0.032	0.484	7
FR8	0.251	0.249	0.246	0.242	0.248	0.033	0.023	0.410	12	0.246	0.241	0.252	0.237	0.246	0.047	0.020	0.301	14
FR9	0.253	0.249	0.246	0.255	0.248	0.026	0.017	0.405	13	0.262	0.260	0.241	0.248	0.250	0.034	0.038	0.527	5
FR10	0.256	0.246	0.254	0.250	0.250	0.024	0.032	0.570	7	0.257	0.260	0.246	0.256	0.253	0.026	0.040	0.609	2
FR11	0.253	0.257	0.249	0.250	0.253	0.021	0.034	0.616	5	0.238	0.257	0.249	0.250	0.253	0.036	0.032	0.468	9
FR12	0.248	0.259	0.251	0.236	0.255	0.030	0.032	0.517	10	0.251	0.265	0.257	0.237	0.261	0.032	0.045	0.579	3
FR13	0.243	0.259	0.259	0.244	0.259	0.023	0.037	0.619	4	0.262	0.246	0.246	0.250	0.246	0.036	0.033	0.475	8
FR14	0.256	0.251	0.262	0.255	0.257	0.014	0.042	0.756	2	0.249	0.236	0.252	0.261	0.244	0.041	0.031	0.431	12
FR15	0.256	0.244	0.254	0.250	0.249	0.027	0.031	0.540	8	0.249	0.244	0.249	0.240	0.246	0.044	0.021	0.319	13
FR16	0.251	0.249	0.262	0.252	0.255	0.018	0.038	0.671	3	0.249	0.246	0.252	0.258	0.249	0.032	0.032	0.504	6
V+	0.256	0.262	0.262	0.263	0.261					0.262	0.265	0.265	0.266	0.265				
V−	0.235	0.238	0.238	0.236	0.240					0.238	0.236	0.236	0.237	0.239				
	DT									RF								
FR1	0.243	0.241	0.241	0.238	0.241	0.052	0.009	0.153	16	0.250	0.238	0.254	0.238	0.246	0.046	0.021	0.311	13
FR2	0.251	0.263	0.256	0.260	0.260	0.013	0.046	0.783	2	0.247	0.259	0.257	0.241	0.258	0.032	0.035	0.527	5
FR3	0.254	0.241	0.254	0.249	0.248	0.037	0.028	0.435	8	0.239	0.246	0.243	0.249	0.245	0.043	0.015	0.261	16
FR4	0.243	0.239	0.249	0.244	0.244	0.047	0.016	0.253	15	0.253	0.240	0.246	0.238	0.243	0.047	0.017	0.263	15
FR5	0.251	0.263	0.264	0.262	0.264	0.008	0.053	0.867	1	0.247	0.264	0.262	0.263	0.263	0.015	0.050	0.773	1
FR6	0.246	0.255	0.254	0.246	0.255	0.029	0.032	0.522	7	0.245	0.259	0.259	0.255	0.259	0.022	0.040	0.641	3
FR7	0.248	0.241	0.254	0.252	0.248	0.037	0.027	0.422	10	0.250	0.243	0.259	0.252	0.251	0.031	0.030	0.492	8
FR8	0.254	0.236	0.256	0.252	0.246	0.040	0.030	0.432	9	0.245	0.246	0.246	0.252	0.246	0.038	0.020	0.340	12
FR9	0.246	0.260	0.249	0.252	0.254	0.026	0.034	0.565	3	0.250	0.251	0.254	0.249	0.253	0.029	0.029	0.503	7
FR10	0.257	0.239	0.249	0.246	0.244	0.043	0.025	0.366	12	0.253	0.246	0.246	0.247	0.246	0.038	0.021	0.353	11
FR11	0.237	0.252	0.246	0.244	0.249	0.040	0.022	0.347	14	0.256	0.256	0.241	0.244	0.248	0.037	0.027	0.425	9
FR12	0.254	0.268	0.241	0.246	0.254	0.032	0.040	0.555	5	0.250	0.240	0.238	0.252	0.239	0.046	0.018	0.276	14
FR13	0.257	0.244	0.243	0.254	0.244	0.039	0.028	0.413	11	0.256	0.264	0.249	0.257	0.256	0.019	0.042	0.685	2
FR14	0.257	0.244	0.264	0.244	0.254	0.034	0.038	0.534	6	0.261	0.246	0.243	0.268	0.245	0.032	0.038	0.542	4
FR15	0.251	0.255	0.243	0.265	0.249	0.030	0.037	0.556	4	0.256	0.251	0.251	0.249	0.251	0.029	0.030	0.510	6
FR16	0.251	0.255	0.235	0.244	0.245	0.043	0.024	0.360	13	0.242	0.251	0.249	0.244	0.250	0.039	0.021	0.353	10
V+	0.257	0.268	0.264	0.265	0.264					0.261	0.264	0.262	0.268	0.263				
V−	0.237	0.236	0.235	0.238	0.241					0.239	0.238	0.238	0.238	0.239				

Table 8. Results of TOPSIS for Adenoma Cancer Dataset with Ranking of Feature Reduction Sequence using Dataset D4.

			k-NN									LR						
FR	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**	**CR1**	**CR2**	**CR3**	**CR4**	**CR5**	S_i^+	S_i^-	**Pi**	**Rank**
FR1	0.245	0.242	0.238	0.249	0.240	0.043	0.014	0.239	15	0.245	0.248	0.246	0.250	0.247	0.032	0.021	0.388	13
FR2	0.245	0.242	0.259	0.244	0.250	0.034	0.031	0.478	9	0.245	0.245	0.254	0.248	0.249	0.031	0.024	0.438	10
FR3	0.250	0.249	0.243	0.260	0.247	0.028	0.028	0.497	7	0.252	0.261	0.254	0.253	0.257	0.015	0.040	0.728	2
FR4	0.255	0.247	0.238	0.254	0.243	0.036	0.023	0.386	13	0.255	0.234	0.256	0.245	0.245	0.038	0.028	0.423	12
FR5	0.255	0.255	0.262	0.265	0.258	0.009	0.049	0.847	1	0.255	0.261	0.261	0.263	0.261	0.004	0.050	0.931	1
FR6	0.255	0.244	0.257	0.246	0.250	0.029	0.032	0.523	6	0.258	0.253	0.246	0.245	0.249	0.029	0.031	0.513	7
FR7	0.250	0.247	0.259	0.257	0.253	0.021	0.038	0.646	3	0.252	0.250	0.251	0.261	0.251	0.021	0.034	0.620	4
FR8	0.252	0.244	0.241	0.254	0.243	0.036	0.022	0.381	14	0.247	0.250	0.248	0.248	0.249	0.029	0.025	0.458	8
FR9	0.250	0.249	0.249	0.257	0.249	0.025	0.016	0.399	12	0.258	0.234	0.256	0.263	0.245	0.034	0.036	0.521	6
FR10	0.250	0.252	0.230	0.244	0.241	0.045	0.014	0.238	16	0.252	0.250	0.235	0.242	0.243	0.041	0.022	0.351	16
FR11	0.242	0.255	0.243	0.246	0.249	0.033	0.022	0.402	11	0.250	0.263	0.251	0.250	0.257	0.019	0.039	0.673	3
FR12	0.250	0.242	0.251	0.246	0.246	0.034	0.025	0.417	10	0.242	0.253	0.243	0.245	0.248	0.035	0.022	0.383	14
FR13	0.245	0.260	0.262	0.238	0.261	0.029	0.042	0.594	5	0.252	0.245	0.251	0.242	0.248	0.033	0.025	0.432	11
FR14	0.250	0.263	0.262	0.241	0.262	0.025	0.045	0.644	4	0.245	0.253	0.240	0.245	0.247	0.036	0.021	0.373	15
FR15	0.252	0.257	0.254	0.252	0.256	0.018	0.037	0.672	2	0.237	0.250	0.251	0.250	0.251	0.032	0.025	0.442	9
FR16	0.255	0.252	0.249	0.244	0.250	0.030	0.027	0.480	8	0.255	0.248	0.256	0.248	0.252	0.025	0.033	0.567	5
V+	0.255	0.263	0.262	0.265	0.262					0.258	0.263	0.261	0.263	0.261				
V−	0.242	0.242	0.230	0.238	0.240					0.237	0.234	0.235	0.242	0.243				
			DT									RF						
FR1	0.246	0.241	0.237	0.236	0.239	0.053	0.008	0.132	16	0.243	0.254	0.245	0.240	0.249	0.043	0.020	0.316	12
FR2	0.249	0.251	0.232	0.252	0.242	0.044	0.022	0.337	14	0.246	0.241	0.237	0.251	0.239	0.054	0.016	0.230	15
FR3	0.246	0.246	0.251	0.246	0.248	0.035	0.025	0.416	10	0.243	0.262	0.258	0.248	0.260	0.025	0.039	0.611	2
FR4	0.254	0.238	0.245	0.244	0.242	0.045	0.021	0.316	15	0.252	0.243	0.253	0.235	0.248	0.047	0.020	0.303	14
FR5	0.249	0.267	0.261	0.260	0.264	0.011	0.055	0.831	1	0.252	0.270	0.266	0.259	0.268	0.008	0.057	0.882	1
FR6	0.251	0.251	0.253	0.249	0.252	0.026	0.033	0.557	6	0.257	0.249	0.250	0.251	0.249	0.035	0.028	0.444	8
FR7	0.251	0.259	0.264	0.260	0.262	0.012	0.052	0.815	2	0.254	0.249	0.255	0.251	0.252	0.032	0.031	0.492	4
FR8	0.249	0.251	0.240	0.249	0.246	0.037	0.023	0.380	13	0.249	0.246	0.239	0.243	0.243	0.050	0.012	0.195	16
FR9	0.254	0.251	0.251	0.254	0.251	0.026	0.034	0.572	4	0.254	0.257	0.247	0.259	0.252	0.029	0.035	0.551	3
FR10	0.259	0.251	0.253	0.246	0.252	0.026	0.036	0.576	3	0.252	0.241	0.250	0.251	0.245	0.043	0.023	0.349	11
FR11	0.251	0.246	0.243	0.254	0.244	0.037	0.026	0.408	12	0.243	0.246	0.242	0.253	0.244	0.045	0.021	0.314	13
FR12	0.241	0.246	0.245	0.257	0.246	0.039	0.027	0.410	11	0.246	0.254	0.253	0.248	0.253	0.032	0.029	0.474	6
FR13	0.249	0.246	0.256	0.260	0.251	0.029	0.037	0.567	5	0.249	0.251	0.250	0.240	0.251	0.039	0.022	0.362	10
FR14	0.246	0.249	0.259	0.246	0.254	0.029	0.034	0.539	8	0.254	0.241	0.253	0.264	0.247	0.039	0.036	0.477	5
FR15	0.251	0.251	0.256	0.244	0.254	0.027	0.034	0.551	7	0.254	0.246	0.253	0.253	0.249	0.035	0.029	0.454	7
FR16	0.254	0.254	0.251	0.244	0.252	0.028	0.032	0.534	9	0.249	0.249	0.250	0.253	0.249	0.035	0.027	0.433	9
V+	0.259	0.267	0.264	0.260	0.264					0.257	0.270	0.266	0.264	0.268				
V−	0.241	0.238	0.232	0.236	0.239					0.243	0.241	0.237	0.235	0.239				

Table 9. Ranking and Selection of Top Three FR.

D	Classifier	Rank1	Rank2	Rank3
D1	k-NN	FR5	FR14	FR10
	LR	FR12	FR5	FR9
	DT	FR5	FR13	FR14
	RF	FR5	FR13	FR14
D2	k-NN	FR5	FR14	FR16
	LR	FR5	FR12	FR11
	DT	FR5	FR11	FR15
	RF	FR5	FR7	FR15
D3	k-NN	FR5	FR14	FR16
	LR	FR5	FR10	FR12
	DT	FR5	FR2	FR9
	RF	FR5	FR13	FR6
D4	k-NN	FR5	FR15	FR7
	LR	FR5	FR3	FR11
	DT	FR5	FR7	FR10
	RF	FR5	FR3	FR9

Table 10. Occurrence of Few Top-Ranking FR.

	No. of Occurrences of Feature Reductions Wrt Rank		
Feature Readuction	**Rank 1**	**Rank 2**	**Rank 3**
FR5	15	1	
FR14		3	2
FR10		1	2
FR111		1	2
FR15		1	1
FR7		2	1
FR12	1	1	1
FR9			3
FR13		3	
FR16			2
FR3		2	
FR6			1
FR2		1	

6. Discussion

The following are the contribution of this proposed work:

- This research work focuses on the feature selection and classification approaches for the gene expression cancer data analysis.

- After going through several research contributions of the last five years, it is observed that filter approaches are successful for hugely spaced data, as these are simple to implement and with less computational cost.
- When a single filter approach is applied, the selection approach may not be able to drop all redundant and insignificant features from the data. This may be due to the data characteristics and score function used in the selection process.
- Thus, a series of filter approaches is applied to rank the features, and thus, few top-ranked features are extracted.
- When a series of filter approaches is applied, which sequence will generate a significant feature is required to be evaluated. Thus, rather than based on a single classifier, multiple classifiers are used to find out the optimal sequence selection.
- Furthermore, this analysis is statistically proven to come up with optimal results.
- The method is generalizable for other diseases (especially for high-dimensional data), if the research challenge is similar as in microarray datasets. Since features are different for the different datasets hence by a little bit of modification, we can solve these issues such as RNA sequence and methylation data accordingly. If the dimensionality is low, then four-stage feature selection may not be required.

7. Conclusions and Future Work

After analyzing all the experimental studies and results analysis individually, it is concluded that the FR5 model works better on the given datasets on all presented classifiers. Finally, it can be concluded that the feature reduction sequence FR5, i.e., (*Correlation-Based Feature Selection* $\rightarrow$ *Chi-Squared Test* $\rightarrow$ *Relief Feature Selection* $\rightarrow$ *Information Gain*) is found to be the superior approach to other feature reduction combination techniques. This analysis is implemented on highly spaced medical data. As filter approaches are quite successful for huge data, this strategy can boost classifier efficiency while requiring little computing work. Future work of this research analysis can be extended using a few more successful feature selection approaches such as other filter approaches, wrapper techniques, and embedded approaches. In this research study, the key challenges are significant feature selection, as the data are huge. Thus, alternatively, this approach can also be successfully applied in application areas where similar challenges are seen.

Author Contributions: Conceptualization, J.T., R.D. and B.K.P.; methodology, J.T., R.D. and S.K.M.; software, J.T., R.D. and T.K.M.; validation, J.T., R.D., T.K.M. and S.K.M.; formal analysis, R.D. and D.P.; investigation, J.T., R.D. and B.K.P.; resources, J.T.; data curation, J.T., R.D. and B.K.P.; writing—original draft preparation, J.T., R.D. and S.K.M.; writing—review and editing, J.T., R.D. and S.K.M.; visualization, R.D., T.K.M. and D.P.; supervision, R.D., B.K.P. and D.P.; project administration, R.D., B.K.P. and D.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data supporting for this work can be found in http://csse.szu.edu.cn/staff/zhuzx/Datasets.html, https://www.kaggle.com/ahmedhamada0/brain-tumor-detection/metadata, http://biogps.org/dataset/tag/adenoma.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Herrero, J.; Vaquerizas, J.M.; Al-Shahrour, F.; Conde, L.; Mateos, A.; Díaz-Uriarte, J.S.R.; Dopazo, J. New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.* **2004**, *32*, 485–491. [CrossRef] [PubMed]
2. Almugren, N.; Alshamlan, H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access* **2019**, *7*, 78533–78548. [CrossRef]
3. Singh, R.K.; Sivabalakrishnan, M.J.P.C.S. Feature selection of gene expression data for cancer classification: A review. *Procedia Comput. Sci.* **2015**, *50*, 52–57. [CrossRef]

4. Dash, R.; Misra, B.B. Pipelining the ranking techniques for microarray data classification: A case study. *Appl. Soft Comput.* **2016**, *48*, 298–316. [CrossRef]
5. Glaab, E.; Bacardit, J.; Garibaldi, J.M.; Krasnogor, N. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS ONE* **2012**, *7*, 39932. [CrossRef]
6. Ghosh, K.K.; Begum, S.; Sardar, A.; Adhikary, S.; Ghosh, M.; Kumar, M.; Sarkar, R. Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data. *Expert Syst. Appl.* **2021**, *169*, 114485. [CrossRef]
7. Sahu, B.; Dehuri, S.; Jagadev, A.K. Feature selection model based on clustering and ranking in pipeline for microarray data. *Inform. Med. Unlocked* **2017**, *9*, 107–122. [CrossRef]
8. Singh, N.; Singh, P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemom. Intell. Lab. Syst.* **2021**, *217*, 104396. [CrossRef]
9. Bommert, A.; Sun, X.; Bischl, B.; Rahnenführer, J.; Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* **2020**, *143*, 106839. [CrossRef]
10. Lazar, C.; Taminau, J.; Meganck, S.; Steenhoff, D.; Coletta, A.; Molter, C.; de Schaetzen, V.; Duque, R.; Bersini, H.; Nowe, A. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *9*, 1106–1119. [CrossRef]
11. Dash, R.; Misra, B.B. A multi-objective feature selection and classifier ensemble technique for microarray data analysis. *Int. J. Data Min. Bioinform.* **2018**, *20*, 123–160. [CrossRef]
12. Li, T.; Zhang, C.; Ogihara, M. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* **2004**, *20*, 2429–2437. [CrossRef] [PubMed]
13. Pirooznia, M.; Yang, J.Y.; Yang, M.Q.; Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genom.* **2008**, *9*, 1–13. [CrossRef] [PubMed]
14. Sharma, A.; Imoto, S.; Miyano, S. A top-r feature selection algorithm for microarray gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *9*, 754–764.
15. Shen, Q.; Shang, C. Aiding classification of gene expression data with feature selection: A comparative study. *J. Comput. Intell. Res. (IJCIR)* **2006**, *1*, 68–76.
16. Behroozi, M.; Sami, A. A multiple-classifier framework for Parkinson's disease detection based on various vocal tests. *Int. J. Telemed. Appl.* **2016**, *2016*, 6837498. [CrossRef] [PubMed]
17. Lu, H.; Chen, J.; Yan, K.; Jin, Q.; Xue, Y.; Gao, Z. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing* **2017**, *256*, 56–62. [CrossRef]
18. Jirapech-Umpai, T.; Aitken, S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinform.* **2005**, *6*, 148. [CrossRef]
19. Dash, R. A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study. *J. King Saud-Univ.-Comput. Inf. Sci.* **2020**, *32*, 232–247. [CrossRef]
20. Singh, R.; Kumar, H.; Singla, R.K. TOPSIS based multi-criteria decision making of feature selection techniques for network traffic dataset. *Int. J. Eng. Technol.* **2014**, *5*, 4598–4604.
21. GeeksforGeeks. Data Normalization in Data Mining. 2019. Available online: https://www.geeksforgeeks.org/data-normalization-in-data-mining/ (accessed on 6 March 2018)
22. Abusamra, H. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Comput. Sci.* **2013**, *23*, 5–14. [CrossRef]
23. Hemphill, E.; Lindsay, J.; Lee, C.M.; Oiu, I.I.; Nelson, C.E. Feature selection and classifier performance on diverse bio-logical datasets. *BMC Bioinform.* **2014**, *15*, S4. [CrossRef] [PubMed]
24. Spencer, R.; Thabtah, F.; Abdelhamid, N.; Thompson, M. Exploring feature selection and classification methods for predicting heart disease. *Digit. Health* **2020**, *6*. [CrossRef]
25. Liu, S.; Xu, C.; Zhang, Y.; Liu, J.; Yu, B.; Liu, X.; Dehmer, M. Feature selection of gene expression data for cancer classification using double RBF-kernels. *BMC Bioinform.* **2018**, *19*, 396. [CrossRef]
26. Gunduz, H. An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification. *Biomed. Signal Process. Control.* **2021**, *66*, 102452. [CrossRef]
27. Mohapatra, D.; Tripathy, J.; Patra, T.K. Rice Disease Detection and Monitoring Using CNN and Naive Bayes Classification. In *Soft Computing Techniques and Applications*; Springer: Singapore, 2021; pp. 11–29.
28. Nazir, S.; Yousaf, M.H.; Velastin, S.A. Evaluating a bag-of-visual features approach using spatio-temporal features for action recognition. *Comput. Electr. Eng.* **2018**, *72*, 660–669. [CrossRef]
29. Assiri, A.S.; Nazir, S.; Velastin, S.A. Breast tumor classification using an ensemble machine learning method. *J. Imaging* **2020**, *6*, 39. [CrossRef]
30. Criminisi, A. Machine learning for medical images analysis. *Med Image Anal.* **2016**, *33*, 91–93. [CrossRef]
31. Ko, B.C.; Kim, S.H.; Nam, J.Y. X-ray image classification using random forests with local wavelet-based CS-local binary patterns. *J. Digit. Imaging* **2011**, *24*, 1141–1151. [CrossRef]
32. Tripathy, J.; Dash, R.; Pattanayak, B.K.; Mohanty, B. Agutomated Phrase Mining Using POST: The Best Approach. In Proceedings of the IEEE International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON), Odisha, India; pp. 1–6.

33. Dash, R.; Samal, S.; Dash, R.; Rautray, R. An integrated TOPSIS crow search based classifier ensemble: In application to stock index price movement prediction. *Appl. Soft Comput.* **2019**, *85*, 105784. [CrossRef]
34. Microarray Datasets. Available online: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html (accessed on 6 March 2018).
35. Brain Tumor Dataset. 2020. Available online: https://www.kaggle.com/ahmedhamada0/brain-tumor-detection/metadata (accessed on 6 March 2018)
36. Adenoma Datasets. Available online: http://biogps.org/dataset/tag/adenoma/ (accessed on 6 March 2018)