Revue d'Intelligence Artificielle
Vol. 37, No. 3, June, 2023, pp. 673-687
Journal homepage: http://iieta.org/journals/ria

A Comparative Analysis of Feature Selection Algorithms for Cancer Classification Using Gene Expression Microarray Data

Wafaa Mustafa Abduallah[1,2]

[1] Department of Information Technology Management, Duhok Polytechnic University, Duhok 42001, Iraq
[2] Department of Computer Science, Nawroz University, Duhok 42001, Iraq

Corresponding Author Email: wafaa.abduallah@dpu.edu.krd

https://doi.org/10.18280/ria.370316

Received: 20 April 2023
Accepted: 26 May 2023

Keywords:
feature selection, gene expression, cancer classification, ReliefF, Chi Square, ANOVA

ABSTRACT

DNA Microarray technology allows simultaneous analysis of gene expression levels, making it useful in cancer classification. However, analyzing Microarray data is challenging due to the large number of genes and their sparsity. Feature selection has emerged as an effective method to overcome these challenges. This research aims to study the impact of three well-known feature selection algorithms (ReliefF, Chi Square, and ANOVA) in enhancing the accuracy of gene expression profile classification. A three-stage approach was employed: data preprocessing, feature selection, and feature classification. The focus is on selecting relevant features to accurately represent the problem under study. Four classifiers, i.e., SVM, GNB, LDA, and KNN, were evaluated using the aforementioned feature selection algorithms.The proposed methodology was tested on 10 publicly available gene expression datasets. Using all genes, the SVM classifier showed the best accuracy and F1 scores, followed by the LDA classifier. When applying the ReliefF feature selection algorithm, the SVM classifier performed best with a 5% dataset ratio. Moreover, the ANOVA feature selection algorithm yielded optimal results with the SVM classifier at dataset ratios of 3%, 4%, and 5%. Lastly, the Chi-square feature selection consistently produced the best results with both SVM and GNB classifiers for all dataset ratios. The study underscores the significance of feature selection for improving gene expression profile classification accuracy. The findings of this research offer promising insights into the analysis of microarray data, which can be instrumental in enhancing the accuracy of cancer classification.

1. INTRODUCTION

DNA microarrays are powerful tools used in the field of genomics to study gene expression levels. They contain a wide range of genes, providing valuable information for disease detection and tumor classification. However, the curse of dimensionality makes machine learning methods unsuitable for analyzing microarray data [1].

The curse of dimensionality refers to the phenomenon where the performance of machine learning algorithms deteriorates as the number of features (dimensions) increases. In microarray data, each gene represents a feature, and with thousands of genes in a typical microarray, the dimensionality becomes extremely high. This high-dimensional space poses difficulties for accurate and efficient analysis, as traditional machine learning algorithms struggle to effectively handle such large feature spaces [1, 2].

Furthermore, microarray datasets are often characterized by sparsity, which means that only a small subset of genes exhibits significant changes in expression levels that are relevant to the task at hand, such as disease classification. Many studies have shown that the majority of features in microarray data do not contribute meaningfully to the prediction of labels, making their inclusion in the analysis inefficient and potentially detrimental to the performance of machine learning models [3].

Besides, the high dimensionality of microarray data requires substantial computational resources and time for processing. Performing computations in a high-dimensional space is computationally expensive and can lead to increased model training times and decreased efficiency. To address these challenges, it is crucial to reduce the dimensionality of microarray data and undertake data preprocessing techniques before applying classification algorithms. Dimensionality reduction techniques, such as feature\gene selection, can help identify the most informative genes and discard irrelevant or redundant ones. By reducing the feature space, the curse of dimensionality can be mitigated, enabling more effective analysis and improved performance of machine learning models on microarray data [2, 4].

Gene selection is the procedure of picking out a group of useful genes from a larger pool of candidates, which selects informative and relevant genes. This collection of genes has allowed scientists to learn a great deal about the genetics of the illness and the underlying processes. Additionally, this method can enhance the efficacy of cancer classification while reducing computing expenses [5, 6].

Recently, a lot of works have been done on gene selection methods [7]. These methods can be broadly classified as filter, wrapper, and embedded as shown in Figure 1 [8].

Without taking into account any learning algorithms, filter methods primarily rely on statistical measures to evaluate the importance of each gene in the training data. A certain number of genes with the highest statistical scores are chosen using this method of gene selection. Since these methods don't rely on a specific classifier, they may process data quickly. However, the rankings of important genes by each method are likely to be different, resulting that they may not be as trustworthy as other methods. So, this means that insignificant genes may be picked over essential ones [9]. There are many common filtering techniques such as Symmetric Uncertainty (SU), Chi-Squared Attribute Evaluator (Chi), Information Gain (IG), ReliefF, t-statistic, signal-to-noise ratio, and Gain Ratio Attribute Evaluator. It is possible to further categorize filter strategies as either univariate or multivariate approaches. Gene dependencies are taken into account by the latter but not the former [10].

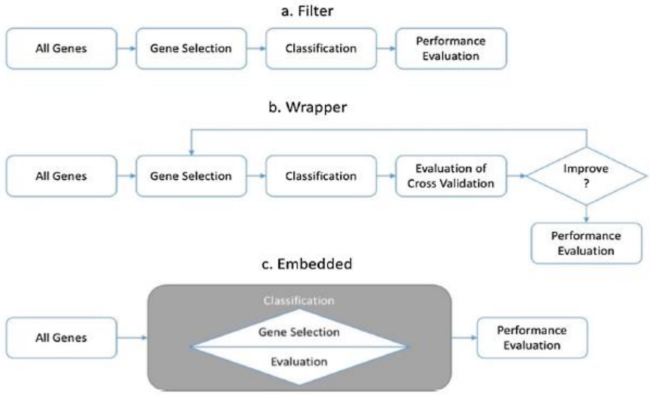


Figure 1. Featured selection methods [8]

Wrapper methods, in contrast to filters, evaluate the quality of the chosen genes by employing learning algorithms to provide the most accurate outcome for classification. These methods have a high rate of accuracy, but they are computationally costly and time-consuming, especially when the provided data comprises thousands of genes [11]. Some examples of such wrapper methods are Particle Swarm Optimization (PSO) [12], Simulated Annealing (SA) [13], and Genetic Algorithm (GA) [14]. On the other hand, embedded gene selection methods, integrate the gene-selection procedure within the learning Algorithm. This method calculates the importance of each gene while building the classifier. These methods have the potential to serve as both a classifier and a gene selector, but they will be quite computationally intensive when the number of genes is huge. Many embedded techniques are presented in the literature for solving multiclass problems such as, multi-task lasso and random forest feature selection [11, 15].

DNA microarrays can store the expression of up to 25,000 genes simultaneously [16], presenting a new challenge of research for gene classification. These features may have a lot of redundancy, and that not all of them are crucial to the classification process, especially knowing that the accuracy can affected negatively by the redundant features. Hence, the primary goal behind this work is to conduct an investigative study to compare the effectiveness of the three main feature selection methods. By doing so, we aim to demonstrate the importance of selecting relevant features that adequately represent the entire dataset. Ultimately, this endeavor is expected to improve the accuracy of cancer classification, a significant aspect in the field of cancer research and medical diagnostics. We will ensure that the research questions driving our study are clearly articulated. By comparing these three feature selection methods, we seek to answer questions such as: Which method yields the most accurate and representative feature subset for cancer classification using gene expression microarray data? How do these methods perform in terms of selecting features that contribute significantly to accurate predictions?

The rest of this paper is organized as follows. In section 2, the fundamental concepts of Microarray technology as well as listing the existing research methods for using genes in cancer classification. Section 3 describes the methodology of the three approaches that have been adopted in this study. The performance evaluation parameters and the experimental results are shown in section 4. In addition, the discussion is presented in section 5. Finally, section 6 gives the conclusion of this study.

2. FUNDAMENTAL CONCEPTS

The fundamentals of Microarray technology are discussed here. To begin, the basics of Microarray technology and gene expression on Microarrays are covered. Class prediction (classification) and its methodologies will be the focus of this investigation as well as the use of genes in cancer classification. The three main types of feature selection techniques Filter, wrapper, and the embedding approach will be reviewed.

2.1 Microarray gene expression

DNA Biologists now have a strong tool in microarray technology (commonly known as "DNA chips") for keeping tabs on how genes are being expressed in various tissues and organs [17]. With this method, scientists may check how many genes are being actively expressed all at once. There are often hundreds of genes (high dimensionality) and just a few of samples available in gene expression data. In addition, it has a ton of unnecessary and redundant components. Health professionals make extensive use of microarrays to study illness mechanisms and develop effective treatments [18].

Microarray technology has created a large microarray data collection reflecting gene expression developed from tissue and cell samples obtained. Gene expression data normally receives thousands of genes (sample). Therefore, these data are well known for their high, detailed, and broad range of detail [19]. Microarray evidence was instrumental in cancer detection and classification. Most microarray data sets contain thousands of genes, but there are a significant number of genes that do not make any effect on diseases. Intelligent algorithms to select genes are necessary because of microarray technology [20]. Microarray technology is an approach to explain the gene-by-gene interactions of genes. In addition to it, the microarray technique can quantify genes activity from the entire genome into one experiment [21].

The most common types of cancer in the human body are Leukemia [22]. Leukemia is a cancer of the blood-forming white cells in the marrow. WBC will be present in the blood of patients with cancer and is fatal. Leukemia can be categorized into two (2) types: chronic or acute leukemia. These types were classified based on when the disease starts, and the damage gets worst. Usually, chronic leukemia strikes individuals progressively and as it gets worse it compromises the adult or the elderly people. For acute leukemia rapidly becomes critical condition and usually occurs in children. For a person affected by chronic leukemia, it will not appear at early stages as the stage of malignant cancer and thus the person will not have early signs and symptoms for the disease [23]. Chronic Lymphocytic Leukemia (CLL) and Chronic Myelogenous Leukemia (CML) are two kinds of leukemia that are a cause of blood cancer [24].

In early stage of acute leukemia, the irregular cells cannot damage the white blood cells. In the first stage, the leukemia cell increases rapidly and unregulated. Acute Lymphocytic Leukemia (ALL) and Acute Myelogenous Leukemia (AML) are two main types of leukemia AML, Figure 2 indicates the sample of blood microscopic images with different types of leukemia [23]. Figure 3 shows the visualization of the process in microarray analysis.

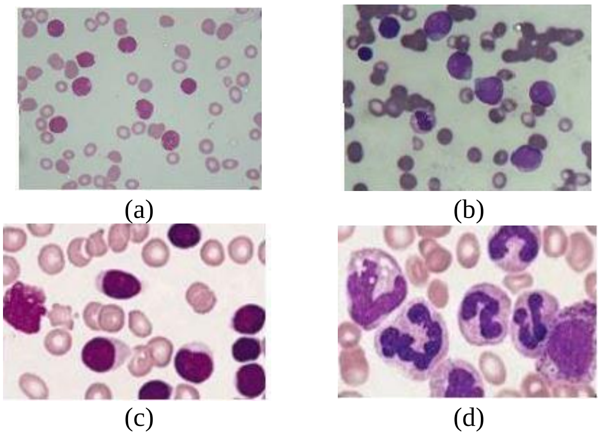


Figure 2. (a) Acute Lymphocytic Leukemia (ALL); (b) Acute Myelogenous Leukemia (AML); (c) Chronic Lymphocytic Leukemia (CLL); (d) Chronic Myelogenous Leukemia (CML)

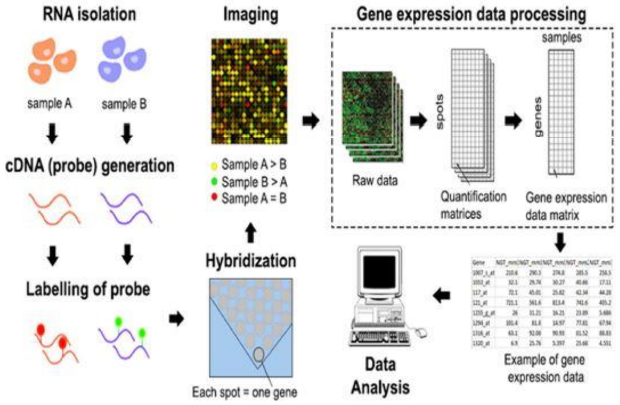


Figure 3. Visualization of the process in microarray analysis [24]

On the other hand, microarray data processing is a crucial step in biological function assessments, specifically for cancer classification. A large, high-dimensional dataset which includes useful genes, redundant genes, irrelevant genes, and noisy genes are generated from the microarray data. To solve the issues caused by a redundant or unnecessary gene, the scientists have developed a technique called gene selection [24].

Unfortunately, in microarray data, the size of features or genes is essentially larger than the number of samples. However, the microarrays gene expression data is so sparse that even a support vector machine classifier cannot get a reasonable result. Consequently, more reliable cancer classification requires a preprocessing phase of gene selection or feature selection prior to the classification itself [25].

2.2 Feature (Gene) selection

With regards to gene expression and in order to identify the genes that are of interest from microarray data, feature selection methods are frequently employed. Discovering genes with unique expression levels requires feature selection. Gene prioritization is another name for the feature (gene) selection process, as is biomarker discovery [1]. The analysis of microarray data is difficult because there are too few samples and too many features to work with. Microarray data is sparse because of the experimental procedure. The post-processing of many microarray data sets is affected by the presence of missing values. Singular Value Decomposition (SVD) based methods (SVDimpute), weighted K-nearest neighbors (KNNimpute), and row average are used to address this issue [26]. Filter, embedded, and wrapper methods are the three main categories of feature selection techniques. The effectiveness of these methods is conditional on how they influence the development of the underlying classification model [27]. The general framework for feature selections has recently been expanded to include new hybrid and ensemble methods. In the following paragraphs, we will discuss the fundamentals of these three categories and the algorithms that underlie them.

2.2.1 Filter approach

The filter method assesses each feature independently using its general statistical characteristics [1]. With the Filter method, no need to any kind of learning algorithm. As a result, it doesn't depend on the classifier. The usual features (genes) were ranked according to certain criteria, and those with the highest scores were chosen for further study. These features are then sent into a classifier or a wrapper technique. The most common types of filter approaches are: Mutual Information (MI) [28] where it calculate the dependency level between two independent features. So, in this procedure, how much one variable (X) have information about another (Y). Information Gain (IG) [29] is a uni-variant Filter method that calculate the amount of information a feature provides about a specific class. Thus, the feature that provides the most information is closely related, while the feature that are not related provides no information. A rise in entropy is the standard unit for measuring information gains (level of impurity). Therefore, a threshold is established, and features scoring above the threshold are picked for further analysis. High information gain leads to high class purity, which in turn increases the likelihood of obtaining the target class [9]. On the other hand, there is another type of Filter method called Minimum Redundancy Maximum Relevance (mRMR) [30], which is multi-variant Filter method with the goal of selecting features that maximize the relevance of the genes while reducing redundancies within each class. Therefore, mutually exclusive features that do not mimic each other are chosen. Another filter method called Laplacian Score (LS) [31] which is defined as is a feature filtering method that uses an unsupervised approach to identify relationships between features. This method is based on the Laplacian Eigen map and Preserving Projection, and it assumes that features belonging to the same class should be close to each other. The LS evaluates each feature based on its ability to preserve its own locality.

Another type of multi-variant Filter method known as Correlation-Based Feature Selection (CFS) [32] which ranks the features using the correlation between the heuristic evaluation functions, the multi-variant. The goal of CFS is to decrease the correlation between features and increase the correlation between features and the class. Finally, Fast correlation-based Filter (FCBF) [33], is a multivariate gene selection approach that starts with a full set of features (genes). It calculates gene dependencies using the symmetrical uncertainty (SU) measurement and identifies the optimal subset using a backward selection method informed by a sequential search. FCBF is an effective computational approach for finding redundant and irrelevant features. It assesses individual qualities and discovers major associations, then heuristically eliminates duplicate information. An internal ending condition causes it to terminate if no Between-Groups features can be found.

2.2.2 Wrapper approach

Wrapper approaches requires the use of learning methods to choose the best possible collection of features [34]. By combining the model hypothesis with the classifier in the search space, a more precise classification result can be achieved. The effectiveness of the wrapper technique is determined by the accuracy of the chosen classifier. This approach often employs evolutionary or biologically inspired algorithms to guide the search process [35]. The wrapper approach begins by generating a population of potential solutions, or feature subsets. These subsets are then evaluated using a learning strategy and a fitness function. Iteration is often used to improve the results. This approach can be computationally expensive and carries a higher risk of overfitting, but it generally produces better performance than the filter approach [34]. The most commonly used wrapper methods are:

Genetic Method (GA) [36] is a heuristic search algorithm that uses the principles of natural evolution and natural selection to find solutions to problems. The GA works by generating a population of potential solutions, called chromosomes, and applying three operations - selection, crossover, and mutation - to produce offspring with improved characteristics. The selection operation identifies the fittest chromosomes, which are then passed on to the next generation. In the crossover operation, two individuals are selected through the selection process. For each individual, a random crossover point is chosen and the two individuals exchange genetic material to create new offspring. Mutation is also included to maintain diversity in the population.

Another algorithm is evolutionary bioinspired from bees known as Artificial Bee Colony ABC) [37]. It takes its cues from the foraging habits of bees. It is based on three types of bees: employed bees, onlooker bees, and scout bees. Employed bees search for solutions, known as food sources, and share information about them with the onlooker bees, who remain in the hive and dance to communicate the location of the food sources. Onlooker bees select the best food sources discovered by the employed bees. If a food source does not improve, the employed bee becomes a scout bee and randomly searches for new food sources. The scout bees contribute to the diversity of the population by introducing new solutions into the search process.

Also, another algorithm is inspired by bird flocks, fish schooling patterns, and swarm theory recognized as Particle Swarm Optimization (PSO) [38]. It is a population-based optimization approach. Where, it involves a population of particles, each of which represents a candidate solution and has a position within the search space. The goal of PSO is for all particles to find the optimal position. To update their positions, particles change their velocity based on their own previous experiences and the best performance of their neighbors, until the optimal position is reached.

Moreover, another nature-inspired optimization method which is based on the behavior of grasshopper swarms called Grasshopper Optimization Algorithm (GOA) [39]. In the GOA, the positions of the grasshoppers represent candidate solutions. The position of each grasshopper is influenced by social interaction, the force of gravity, and wind advection. The GOA is used to calculate the proximity between two grasshoppers.

2.2.3 Hybrid (Ensemble) approach

The hybrid method is constructed so as to get benefits from both approaches: the filter and wrapper. As a result, it combines the excellent performance of the wrapper strategy with the computational economy of the filter approach [40]. The first step in its two-stage design reduces the dimension of the feature space. The next step is to use the wrapper technique to select best collection of features. The hybrid model may not be as accurate because the filter and wrapper are applied in separate steps [41].

The ensemble method operates under the premise that the results obtained by combining the findings of several experts is superior to those obtained by using the findings of a single expert. A single wrapper approach may produce excellent results on one dataset, but may not perform well on another. Accordingly, by combining multiple methods, the overall error rate is reduced [42].

Lu et al. [43] presented a new hybrid feature selection algorithm named MIMAGA-Selection based on the combination of both Mutual Information Maximization (MIM) Algorithm with the Adaptive Genetic Algorithm (AGA). Initially, MIM was used as a filter to identify genes with a high dependence on all other genes. The number of genes selected using MIM was set to 300. After then, AGA was initiated. So, the proposed algorithm, referred to as the AGA, was tested using six multi and binary cancer gene expression datasets. An extreme learning machine (ElM) was chosen as the classifier and the classification process was repeated 30 times. The authors used the same dataset with the same number of target genes to develop the MIMAGA-Selection technique and to evaluate the efficiency of three existing algorithms—sequential forward selection (SFS), ReliefF, and MIM with the ElM classifier. These finding proved that MIMAGA-Selection was more accurate than other feature selection algorithms currently in use. In addition, four distinct classifiers—a back propagation neural network (BP), a support vector machine (SVM), an extreme learning machine (ELM), and a regularized extreme learning machine—are used by the authors to classify the gene chosen via MIMAGA-Selection. The accuracy of all four classifiers is greater than 80%.

Pashaei et al. [44] provided a fresh approach to gene selection using Binary Black Hole Algorithm (BBHA) and Random Forest Ranking (RFR). By using RFR as a filter, the genes were sorted using RFRBBHA-Bagging. The top 500 genes were then combined to form a new gene subset that would be fed into BBAH. From the pool of candidate genes narrowed down in the previous stage, the Black Hole Algorithm was used to choose the set of genes that would ultimately serve as the basis for the organism. The suggested technique was assessed by applying the Bagging classifier to four sets of Microarray cancer data with 10-fold cross validation. This procedure is repeated 100 times. Finally, seven well-known classifiers were used to be compared with the RFR-BBHA-Bagging proposed method; the results showed that the proposed method had the highest accuracy across both datasets.

2.3 Classification

Data classification is a data mining approach used to forecast which of a fixed set of class labels will be applied to a given batch of data. In supervised learning methods like classification, the label for each category is explicitly stated. First, the model (classifier) is formed based on a set of training data, paired with a class label. This is known as the learning phase. Second, the model is applied to unknown data to forecast its class label, and third, the classifier's efficacy is evaluated [45]. The following sections present some of the most popular approaches that are usually utilized for classifying Microarray data.

2.3.1 Support vector machine (SVM)

Is a supervised machines learning algorithm. The primary focus of support vector machines (SVMs) is finding a hyperplane that splits the tuples into classes in the most efficient way possible. By utilizing the margin and the support vector, we can determine the hyperplane. The vectors (data points) used to form the hyperplane are used to derive the support vector. The margin is defined as the distance from the hyperplane to the nearest point. The hyperplane is the line that divides the data into two parts, where each portion ultimately belongs to a single class, where the data can be separated linearly. The optimal hyperplane is found by maximizing the margin, defined as the distance from the hyperplane's origin to the nearest data point (the support vector) in either class. Since this is the case, SVM looks for the hyperplane with the greatest Maximum Marginal Hyperplane (MMH). If the data cannot be separated linearly, then the approach is modified to become a nonlinear SVM, with the goal being the discovery of nonlinear hypersurfaces. Since most real-world data is unstructured and non-linearly separable so, the soft margin classifier is the method of choice. This allows some points to be on the wrong side of the hyperplane, which in turn allows the hyperplane to be violated at those points. SVM's main benefits are its efficiency and its suitability for use with high-dimensional data. More importantly, it performs well when there are more features than samples [45, 46].

2.3.2 K Nearest Neighbour (KNN)

K-nearest neighbor is a straightforward non-parametric instance-based supervised learning algorithm. K-NN works on the basis of the similarity measure, with new examples being classified by first looking for the K most similar ones among the training set (most similar instances, like neighbors). Using a distance metric, such as the Euclidean distance, allows us to evaluate how similar two objects are to one another. Once this is complete, the K members vote on how to divide up the new instances (similar neighbors). One of K-NN's benefits is how easy it is to apply. Furthermore, the input data does not require any particular distribution [47].

2.3.3 Gaussian naive bayes classifier

As a supervised learning algorithm, Gaussian naive Bayes (GNB) classification utilizes Bayes' theorem to place observations into one of many classes depending on the values of predictor variables. Under the naive assumption that the predictor variables are class-conditionally independent, GNB classifiers estimate the conditional probabilities that an observation belongs to a specific class given the values of the predictor variables [48].

2.3.4 Linear discriminant analysis (LDA)

The LDA algorithm is a popular choice among classification algorithms and it works by determining the variance values between and within different classes [49]. LDA employs a linear transformation to identify a projection matrix that maximizes the ratio of within-class to between-class variances in a lower dimensional space. The LDA transformation is calculated using the method of scattering matrices, which is based on Eigen decomposition. This method is often used for data with high dimensions in various applications. Discriminant analysis classifiers have numerous applications, including face recognition and image retrieval. Furthermore, LDA classifiers have been utilized in the medical field for various purposes, for example, analyzing electromyography signals, classifying lung cancer, and diagnosing breast cancer as stated in the study [50].

3. METHODOLOGY

This section presents a detailed structure of proposing an efficient approach for distinguishing among three main feature selection algorithms (RF, Chi Square, and ANOVA). These important algorithms implemented for microarray data classification to show the effectiveness of each one. The methodology stages consist of three main phases: data preprocessing, feature selection, and feature classification. Figure 4 shows the proposed block diagram for general feature selection algorithm structure. This process of the implementation provides the ability of selecting the most relevant features that represent the whole dataset, which consequently causes the accuracy increasing.

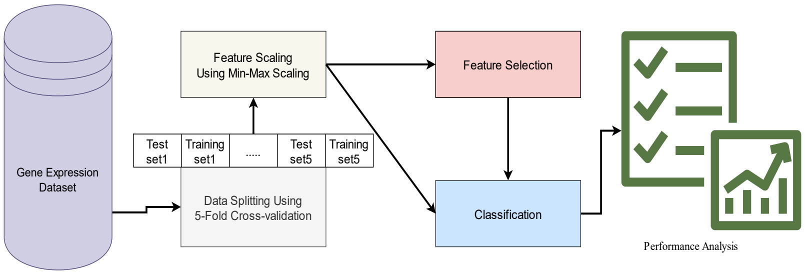


Figure 4. Proposed block diagram for general feature selection algorithm structure

3.1 First phase: Preprocessing

The first phase is feature scaling which is a method used to normalize the range of independent variables or features of data and can be considered as data preprocessing step. In this work, min-max scaling/normalization is applied, which is the simplest method and consists of rescaling the range of features to make it in the range [0, 1]. The general formula for normalization is given as [51]:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

where, max(x) and min(x) are the maximum and the minimum values of the features respectively.

3.2 Second phase: Feature selection

While, the second phase is feature selection. In this work, the three algorithms of feature selection are implemented.

3.2.1 Relief feature (RF) selection algorithm

Relief is a feature selection filtering process that estimates feature "quality" or "relevance" to the target notion by computing a proxy statistic (i.e., predicting endpoint value). Statistically speaking, these characteristics are given values between -1 (the worst) and +1 (the best) known as feature weights (W[A]= weight of feature'A'). (best). Notably, the first iteration of the Relief method could only be used to issues of binary classification and did not include any way to deal with missing data. The original Relief algorithm's Pseudo-code phases are depicted in Algorithm 1 [52].

Algorithm 1 Pseudo-code for the Original Relief Algorithm

```
Require: for each training instance a vector of feature values and the class value n← number of training instances
a ← number of features (i.e. attributes) Parameter: m←number of random training instances out of n used to update W initialize all feature weights W[A]:= 0.0
for i:=1 to m do randomly select 'target' instance Ri
    find a nearest hit 'H' and nearest miss 'M' (instances)
    for A:=1 to a do
        W[A]:=  W[A] – diff (A, Ri, H)/m+diff (A, Ri, M)/m
    end for
end for S
return the vector W of feature scores that estimate the quality of features
```

Relief algorithm is a widely used feature selection method that aims to identify relevant features by considering the differences between nearest neighboring samples. The Relief algorithm operates in two phases: the nearest hit and nearest miss phases. In each phase, it evaluates the relevance and importance of features by examining the differences between the feature values of the nearest instances with the same and different class labels, respectively. By computing the feature weights based on these differences, Relief assigns higher weights to features that contribute more to discriminating between classes. The Relief algorithm lies in its ability to capture feature interactions and dependencies. By comparing feature values within neighboring instances, Relief can identify features that exhibit consistent differences when class labels change. This approach enables the algorithm to prioritize features that are most informative for distinguishing between different classes in the dataset

3.2.2 ANOVA

The ANOVA test compares within-treatment variance with between-treatment variation in a given characteristic. Because of their ability to tell us whether or not a given feature adequately accounts for variation in the dependent variable, variances like these play a crucial role in this filtering strategy. The feature has not done a sufficient job of explaining the variance in the dependent variable if the variance within each treatment is bigger than the variance across treatments. If you want to do an ANOVA test, you'll need to calculate a F statistic for each feature, with the variance between treatments (SST, sometimes misunderstood as SSTotal) in the numerator and the variation within treatments (SSTotal) in the denominator. The test statistic is then put to the test by comparing it to the null hypothesis (H0: Mean value is equal across all treatments) and the alternative hypothesis (H: At least two treatments vary). The typical use of ANOVA is in situations where one of the variables is numerical and the other is categorical, such as when the input variables to a classification job are numerical and the goal variable is a categorical. The test findings can be utilized for feature selection, where characteristics unrelated to the target variable are eliminated. Pseudo-code phases of the ANOVA for feature selection is explained in Algorithm 2 [53].

Algorithm 2 Pseudo-code of ANOVA for Feature Selection

```
Input: M: Feature matrix of size S×G where S represents sample size and G represent feature size
Output: Select top N featu
for each feature fi do
    i=1,2, …. G
    Evaluate the value of MSB
    Compute the value of MSW
    Compute the F-Statistics value (Fi)
    Compute p-value (pi) for each  F-Statistics using the F-distribution table
    if pi < 0.001 then
        select the feature fi
        append  fi to feature matrix GM
    else
        feature  fi is discarded
    end if
    sort the features in ascending order of their p value
    if size of GM > 500 then
        select only top-500 features
    else
        keep the feature matrix GM as it is
    end if
end for
Return the features from the  feature matrix GM
```

When ANOVA is applied to microarray data analysis, it identifies genes with significant expression variations across different cancer subtypes. Also, it measures the variability of gene expression levels between groups and computes the F-statistic. Genes with large F-statistic values are deemed important for discriminating between cancer subtypes.

3.2.3 Chi-square algorithm

A chi-square test is used in statistics to test the independence of two events. Given the data of two variables, we can get both (Observed count *O* and Expected count *E*). Chi-Square measures how *E* and *O* deviates each other. The formula of chi-square is [54]:

$$x_c^2 = \frac{(O_{i\ -E_i})^2}{E_i} \tag{2}$$

where:

c: degree of freedom
O: observed value(s)
E: expected value(s)

Consider a situation in which you need to establish a connection between an independent category feature (predictor) and a dependent category feature (response). Our goal in feature selection is to zero down on those characteristics that are reliant on the outcome. In cases when the two properties being tested are unrelated, the Chi-Square value will be lower since the observed count will be closer to the predicted count. Because the Chi-Square value is so high, we cannot accept the null hypothesis of independence. What this means is that features with larger Chi-Square values are more responsive to the response and hence more suitable for model training. Among the feature selection algorithms, Chi-Square is a popular choice. To streamline the classification procedure, this method is used to eliminate extraneous features [55].

In the context of microarray data analysis, Chi Square feature selection method focuses on identifying relevant genes by assessing the association between gene expression levels and cancer subtypes. It quantifies the dependence between a gene's expression level and class labels through contingency tables and calculates the chi-square statistic. Higher chi-square values indicate stronger associations, making those genes more informative for cancer classification.

3.3 Third phase: Classification

Finally, the third phase is Classification, where four classifiers are used to evaluate the quality of the features that have been selected in the previous phase. The utilized classifiers are: SVM, GNB, LDA, and KNN. These classifiers will be used to obtain the accuracy of the depended feature selection algorithms: RF, ANOVA, and Chi-square. The implementation will be done using all and (1%, 2%, 3%, 4%, and 5%) ratios of ten important datasets.

The model was trained and evaluated in this study using K-fold cross-validation. Multiple categories have been established for the dataset. There are K groups, or 'folds,' where K is the total number of groupings. Simultaneously, cross-validation is a method for testing ML models. One such method is called k-fold cross-validation, and it involves splitting a dataset into k groups, with k^{-1} of those groups training a model and the remaining k groups testing and evaluating it. In this method, the model is trained k times, with each iteration being evaluated by a different fold. In K-fold cross-validation, it means that everyone in every fold helps to train and evaluate the model. One example of cross-validation is shown in Figure 5 (5-fold). The picture shows that there are 5 different subsets, or "folds," of the dataset; four of these folds are responsible for training the model, while the fifth evaluates its progress after each iteration. At the end we evaluate the score of the model by finding the mean of all five scores that we have obtained through the iterations. In our study, we also have employed 5-fold cross-validation.

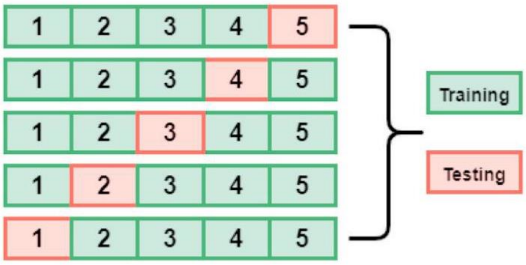


Figure 5. Graphical representation of K-fold cross-validation

4. EXPERIMENTAL RESULTS

This section deals with presenting a detailed description of the famous ten depended datasets in this work. Then, the performance analysis of used algorithms with these datasets will determined via standard equations. The results of the implementation for all the four algorithms using all datasets are determined, listed, and plotted. Then, those of each algorithm determined and plotted.

4.1 Datasets description

Ten microarray datasets have been selected for this work where, they are briefly described in Table 1 and were obtained from (Http://Csse.Szu.Edu.Cn/Staff/Zhuzx/Datasets.Html). However, the reason behind selecting these datasets was; they include both binary and multiclass classes where, the latter of which is more difficult to distinguish.

Table 1. Description of depended microarray datasets

Datasets	#Samples	#Genes	#Classes	Class Distributions
ALL-AML	72	7129	2	'ALL': 47, 'AML': 25
ALL-AML-3	72	7129	3	'AML': 25, 'B-cell': 38, 'T-cell': 9
ALL-AML-4	72	7129	4	'B-cell': 38, 'BM': 21, 'PB': 4, 'T-cell': 9
Breast Cancer	97	24481	2	'non-relapse': 51, 'relapse': 46
CNS	60	7129	2	'0': 39, '1': 21
Colon Tumor	62	2000	2	'Normal': 22, 'Tumor': 40
Lung Cancer	203	12600	5	'1': 139, '2': 17, '3': 6, '4': 21, '5': 20
MLL	72	12582	3	'ALL': 24, 'AML': 28, 'MLL': 20
Ovarian Cancer	253	15154	2	'Cancer': 162, 'Normal': 91
SRBCT	83	2308	4	1: 29, 2: 11, 3: 18, 4: 25

4.2 Performance analysis

All the experiments of this work were carried out using Python programming environment on a PC with Intel(R) Core (TM) i7-4702MQ CPU and 8.00 GB RAM. In addition, the libraries of (sklearn, numpy, and ReliefF) were used to conduct the experiments of the proposed methodology. Regarding the evaluation metrics used for obtaining the results, the following were taken into account [56]:

$$\text{Accuracy} = \frac{\mathit{Correctly\ Classified\ Samples}}{\mathit{Total\ Samples}} \times 100 \tag{3}$$

$$\text{Precision} = \frac{\text{True Positive}}{\text{Total Predected Positive}} \times 100 \tag{4}$$

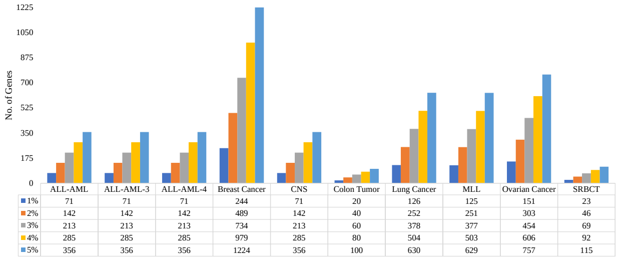
$$\text{Recall} = \frac{\text{True Positive}}{\text{Total Actual Positive}} \times 100 \tag{5}$$

$$\text{F1} - \text{Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \tag{6}$$

where, accuracy is the fraction of observations for which a prediction was correct, Precision is the fraction of positively predicted observations for which a prediction was correct, Recall is the fraction of positively predicted observations for which a prediction was correct, and F1-Score is the weighted average of Precision and Recall. In this study, we apply the SVM, GNB, LDA, and KNN classifiers to 10 datasets and run them through the aforementioned assessment metrics to get a sense of how they affect the performance of the actual system.

4.3 Experimental results

To demonstrate the efficiency of four different classifiers investigated in this study, four different experiments were conducted on above mentioned datasets as explained in the following sections. From the total number of each dataset illustrated in Table 1, there are five ratios of them been selected to be depended in the experiments (2, 3, and 4). The ratios are (1%, 2%, 3%, 4%, and 5%). Figure 6 shows these ratios for all ten datasets.



	ALL-AML	ALL-AML-3	ALL-AML-4	Breast Cancer	CNS	Colon Tumor	Lung Cancer	MLL	Ovarian Cancer	SRBCT
1%	71	71	71	244	71	20	126	125	151	23
2%	142	142	142	489	142	40	252	251	303	46
3%	213	213	213	734	213	60	378	377	454	69
4%	285	285	285	979	285	80	504	503	606	92
5%	356	356	356	1224	356	100	630	629	757	115

Figure 6. Different ratios of selected genes to be used by the three algorithms

Experiment 1: Comparative Analysis Using All Genes

Experiment-1 involved the utilization of raw gene expression data from ten microarray datasets. The data was initially divided into training and testing sets using a fivefold cross-validation approach. To ensure equal contribution of all genes, the gene data of each dataset underwent scaling with the Min-Max scaling method. Four different classifiers were then independently trained on each scaled training set to assess their performance on unseen scaled data (testing set). The evaluation metrics used were Accuracy and F1, and the results are presented in Table 2.

The results from Experiment-1, as shown in Table 2, indicate the following findings. For the SVM classifier, the highest accuracy and F1 values were achieved when using the Ovarian Cancer and SRBCT datasets. On the other hand, the GNB classifier demonstrated the best accuracy and F1 scores when utilizing the SRBCT dataset. In the case of both the LDA and KNN classifiers, the Ovarian Cancer dataset yielded the best accuracy and F1 values.

These observations provide insights into the performance of different classifiers when applied to the scaled testing data of the microarray datasets in Experiment-1.

Table 2. Performance analysis of SVM, GNB, LDA, and KNN on ten microarray datasets without gene section algorithm

#	Dataset	SVM		GNB		LDA		KNN	
		Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
1	ALL-AML	98.57	98.50	98.57	98.56	90.10	89.10	85.90	83.87
2	ALL-AML-3	97.24	95.77	94.38	92.31	83.24	78.74	80.38	78.47
3	ALL-AML-4	94.48	92.89	90.19	90.39	77.62	74.45	78.95	78.77
4	Breast Cancer	67.11	62.97	49.47	43.43	59.00	55.25	54.84	50.96
5	CNS	65.00	58.50	65.00	62.11	68.33	62.18	58.33	52.70
6	Colon Tumor	87.18	86.15	61.79	60.49	79.23	75.31	77.69	75.03
7	Lung Cancer	94.09	92.23	89.12	87.91	94.57	88.94	87.62	83.69
8	MLL	97.24	96.81	94.57	94.21	89.24	88.64	83.43	82.17
9	Ovarian Cancer	100.00	100.00	89.72	88.84	100.00	100.00	94.05	93.59
10	SRBCT	100.00	100.00	98.75	98.57	56.25	52.57	82.28	82.95

Experiment 2: Comparative Analysis Using a Subset of Genes Selected By RF

For Experiment-2, the RF feature selection algorithm was employed to select specific ratios from the datasets, using different classifiers. Figures 7 to 14 depict the results obtained from this experiment. Each classifier has two corresponding figures, one representing the accuracy and the other representing the F1 values. The findings reveal that the SVM classifier achieved the highest accuracy and F1 scores when utilizing the 5% ratio of the Ovarian Cancer dataset (Figures 7 and 8). On the other hand, the GNB classifier demonstrated its best performance when using the MLL dataset with a 5% ratio (Figures 9 and 10). Furthermore, the LDA classifier yielded its best results when depending on the Ovarian Cancer dataset with a 5% ratio (Figures 11 and 12). Lastly, the KNN classifier showed optimal performance with the Ovarian Cancer dataset at a 4% ratio (Figures 13 and 14).

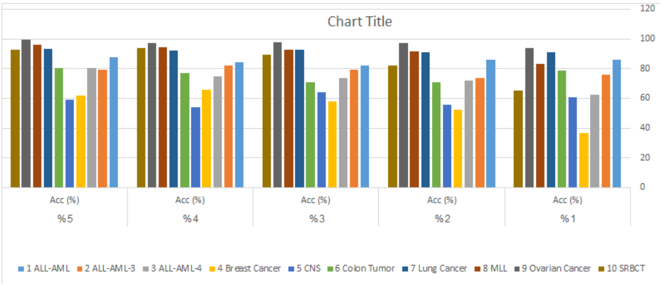


Figure 7. Accuracy evaluation metric of SVM for the ten microarray datasets using RF algorithm

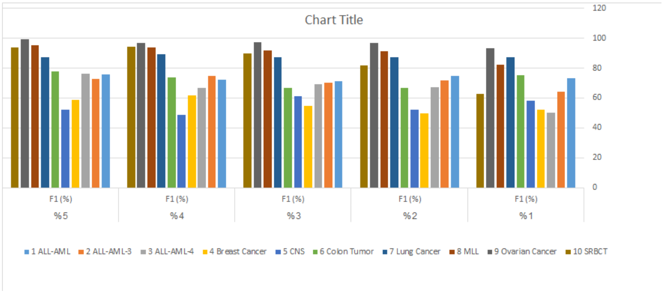


Figure 8. F1 evaluation metric of SVM for the ten microarray datasets using RF algorithm

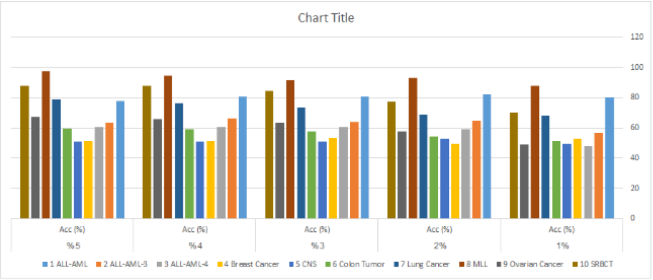


Figure 9. Accuracy evaluation metric of GNB for the ten microarray datasets using RF algorithm

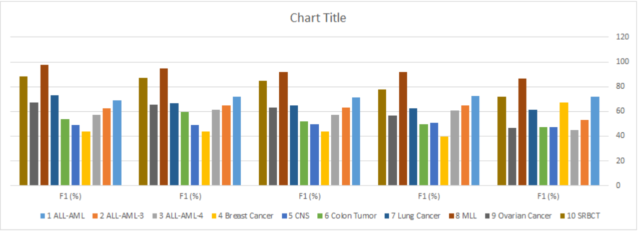


Figure 10. F1 evaluation metric of GNB for the ten microarray datasets using RF algorithm

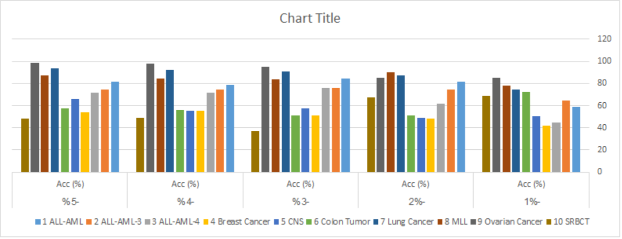


Figure 11. Accuracy evaluation metric of LDA for the ten microarray datasets using RF algorithm

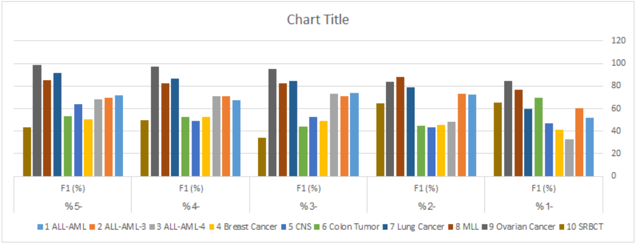


Figure 12. F1 evaluation metric of LDA for the ten microarray datasets using RF algorithm

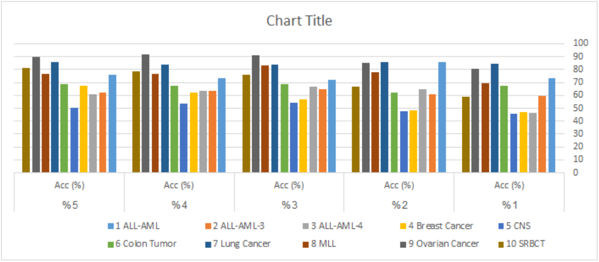


Figure 13. Accuracy evaluation metric of KNN for the ten microarray datasets using RF algorithm

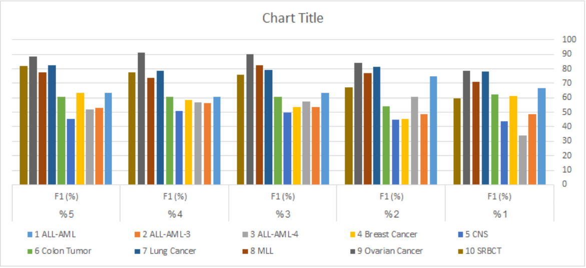


Figure 14. F1 evaluation metric of KNN for the ten microarray datasets using RF algorithm

These findings highlight the varying performance of classifiers based on different datasets and ratios. The results emphasize the importance of dataset selection and the impact it has on the performance of classifiers using the RF feature selection algorithm.

Experiment 3: Comparative Analysis Using a Subset of Genes Selected by ANOVA Algorithm

In Experiment-3, the ANOVA feature selection algorithm was employed to select specific ratios from the ten microarray datasets, using various classifiers. The evaluation metrics results (Accuracy and F1) for all classifiers (SVM, GNB, LDA, and KNN) are depicted in Figures 15 to 22. For the SVM classifier, the SRBCT dataset with ratios of 3%, 4%, and 5% produced the best accuracy and F1 scores, along with the Ovarian Cancer dataset that showed similar values across all ratios (Figures 15 and 16). The GNB classifier achieved its best results when utilizing ratios of 2%, 3%, and 4% from the SRBCT dataset (Figures 17 and 18). Similarly, the LDA classifier demonstrated its optimal performance when depending on the Ovarian Cancer dataset with ratios of 3% and 5% (Figures 19 and 20). Lastly, the KNN classifier exhibited the highest accuracy and F1 values when utilizing the SRBCT dataset with ratios of 2%, 3%, 4%, and 5% (Figures 21 and 22).

These findings emphasize the varying performance of classifiers based on different datasets and ratios when using the ANOVA feature selection algorithm.

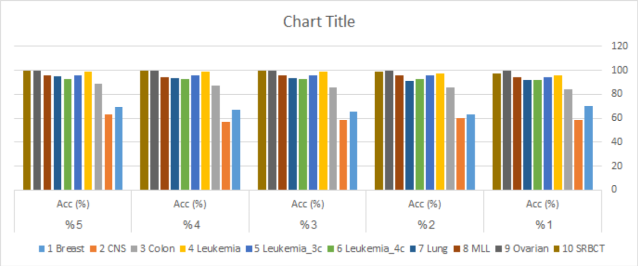


Figure 15. Accuracy evaluation metric of SVM for the ten microarray datasets using ANOVA algorithm

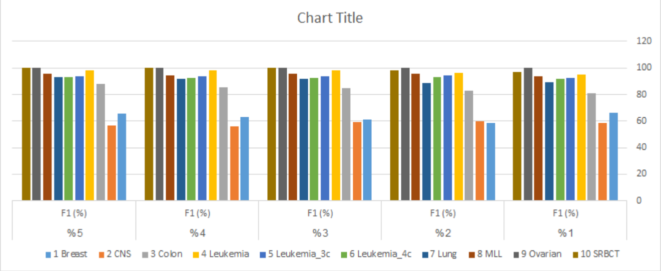


Figure 16. F1 evaluation metric of SVM for the ten microarray datasets using ANOVA algorithm

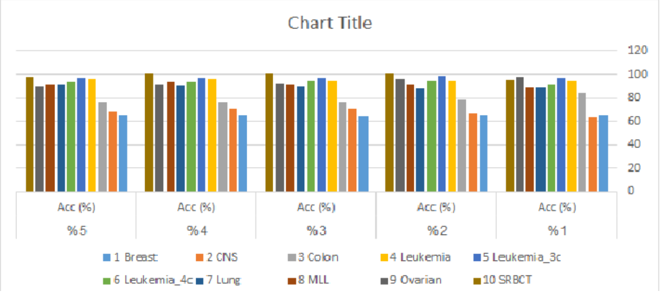


Figure 17. Accuracy evaluation metric of GNB for the ten microarray datasets using ANOVA algorithm

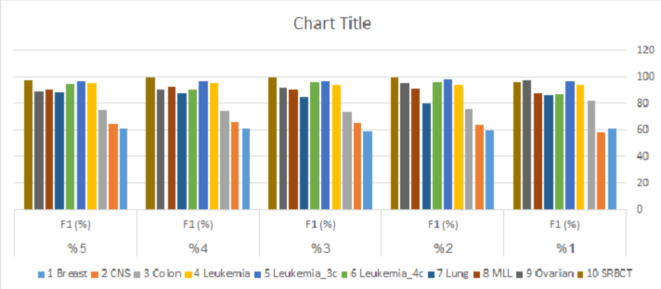


Figure 18. F1 evaluation metric of GNB for the ten microarray datasets using ANOVA algorithm

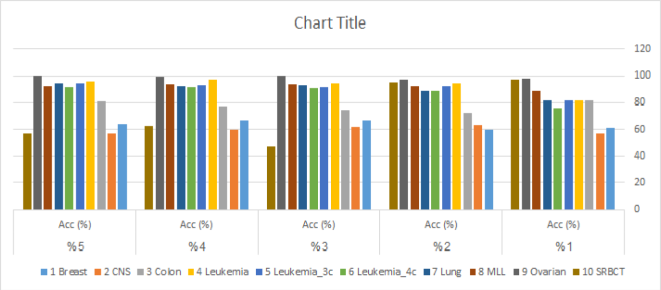


Figure 19. Accuracy evaluation metric of LDA for the ten microarray datasets using ANOVA algorithm

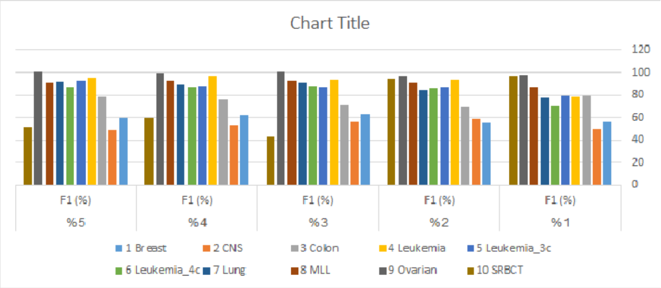


Figure 20. F1 evaluation metric of LDA for the ten microarray datasets using ANOVA algorithm

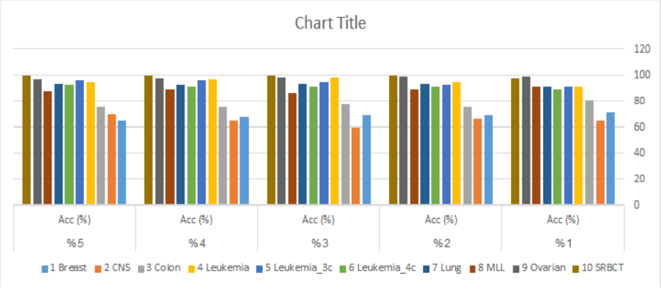


Figure 21. Accuracy evaluation metric of KNN for the ten microarray datasets using ANOVA algorithm

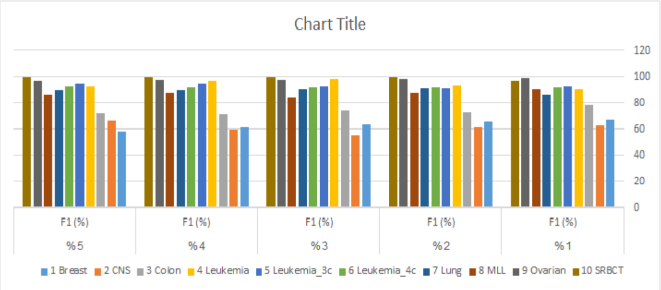


Figure 22. F1 evaluation metric of KNN for the ten microarray datasets using ANOVA algorithm

Experiment 4: Comparative Analysis Using a Subset of Genes Selected by Chi Square Algorithm

In Experiment-4, the Chi-square feature selection algorithm was utilized to select specific ratios from the ten microarray datasets, employing multiple classifiers. The evaluation metrics results (Accuracy and F1) for all classifiers (SVM, GNB, LDA, and KNN) are illustrated in Figures 23 to 30. For the SVM classifier, the SRBCT dataset with a 1% ratio exhibited the best accuracy and F1 scores, along with the Ovarian Cancer dataset that demonstrated the same values but for ratios of 2%, 3%, 4%, and 5% (Figures 23 and 24). The GNB classifier achieved its optimal performance when utilizing the SRBCT dataset across all ratios (Figures 25 and 26). Similarly, the LDA classifier demonstrated its highest accuracy and F1 values when depending on the Ovarian Cancer dataset with ratios of 4% and 5% (Figures 27 and 28). Lastly, the KNN classifier exhibited superior performance when using the SRBCT dataset across all ratios (Figures 29 and 30).

These results highlight the performance variation of classifiers based on different datasets and ratios when employing the Chi-square feature selection algorithm.

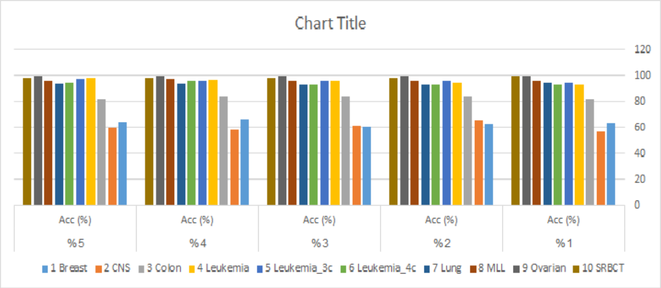


Figure 23. Accuracy evaluation metric of SVM for the ten microarray datasets using Chi Square algorithm

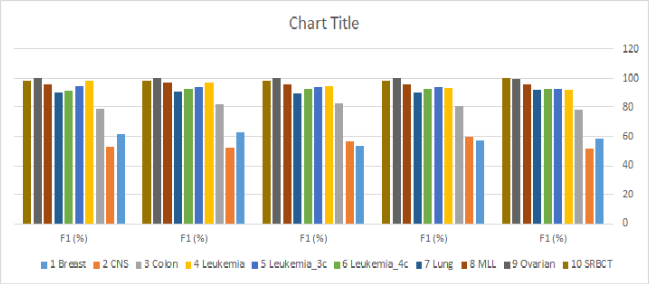


Figure 24. F1 evaluation metric of SVM for the ten microarray datasets using Chi Square algorithm

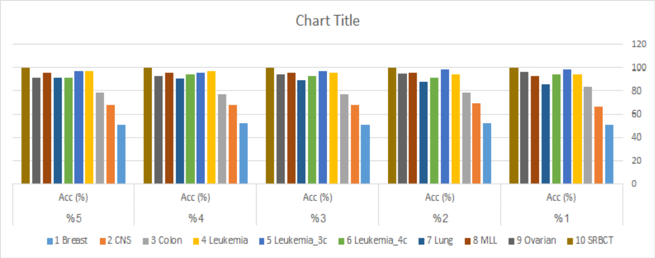


Figure 25. Accuracy evaluation metric of GNB for the ten microarray datasets using Chi Square algorithm

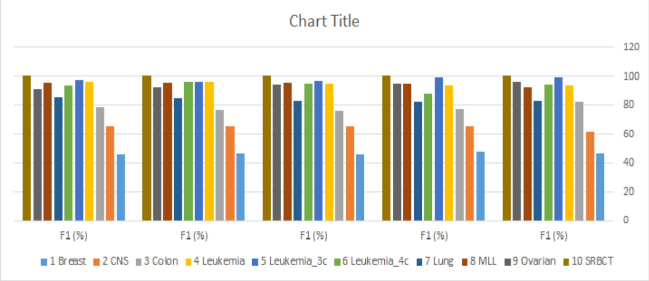


Figure 26. F1 evaluation metric of GNB for the ten microarray datasets using Chi Square algorithm

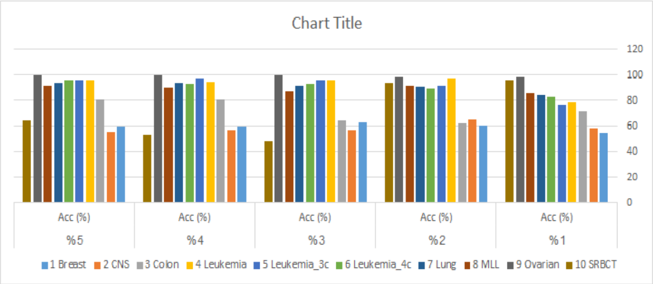


Figure 27. Accuracy evaluation metric of LDA for the ten microarray datasets using Chi Square algorithm

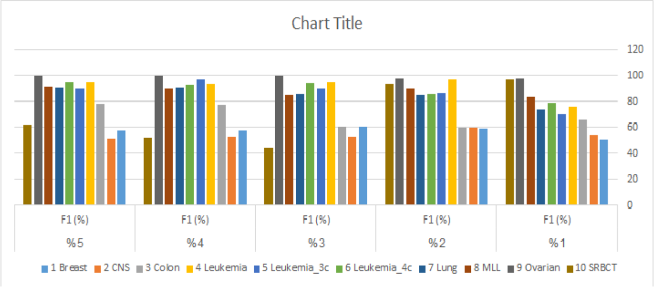


Figure 28. F1 evaluation metric of LDA for the ten microarray datasets using Chi Square algorithm

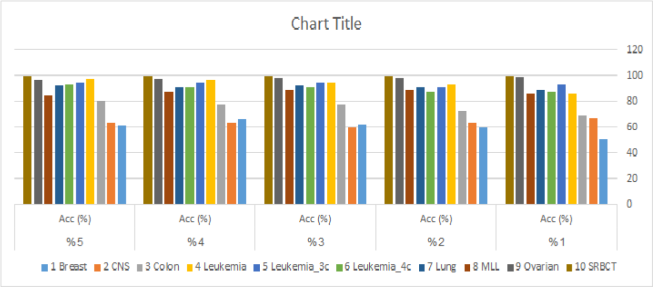


Figure 29. Accuracy evaluation metric of KNN for the ten microarray datasets using Chi Square algorithm

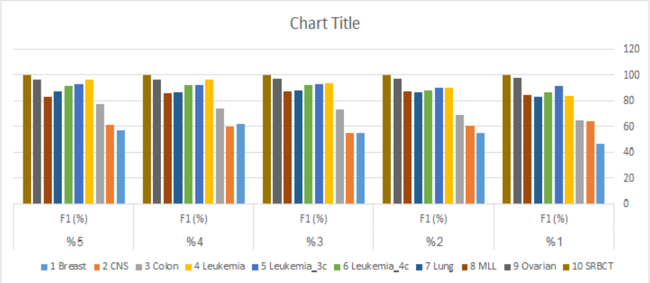


Figure 30. F1 evaluation metric of KNN for the ten microarray datasets using Chi Square algorithm

5. DISCUSSION

From the four experiments conducted, we have observed that the results obtained by the four tested classifiers tend to vary among the ten tested datasets. It is evident that the highest accuracy and F1 values for the ten tested datasets are not consistently the same. Several factors contribute to these variations.

Firstly, the number of samples in each dataset differs, which impacts the performance of feature selection methods. The effectiveness of feature selection can be influenced by the availability of an adequate number of samples for learning patterns and making accurate predictions. Datasets with larger sample sizes tend to provide more reliable and robust results.

Additionally, the distribution of samples across the classes in the datasets also varies. Class imbalance can pose challenges in feature selection, as it may lead to biased or inaccurate feature importance estimations. The suitability of feature selection methods can depend on their ability to handle imbalanced class distributions effectively.

Furthermore, the nature of gene expression varies across different datasets. Some datasets may contain genes with less noise or have clearer patterns, while others may have more noise or complex variations. The efficacy of feature selection methods can be influenced by the characteristics of gene expression in a particular dataset.

To address these factors, we conducted four experiments, each focusing on different aspects. The first experiment encompassed the entire dataset scope, while the subsequent experiments were conducted using different ratios (1%, 2%, 3%, 4%, and 5%) of the datasets. We employed the RF, ANOVA, and Chi-square feature selection algorithms for Experiments 2, 3, and 4, respectively.

Regarding to Expriment-1, from the results illustrated in Table 2, it can be observed that when for the SVM classifier, the highest accuracy and F1 values obtained when using the datasets (Ovarian Cancer, and SRBCT). While for the GNB classifier, the best values of both (accuracy and F1) obtained when using the SRBCT dataset. However, the best values of (accuracy and F1) for both (LDA and KNN) classifier have been determined when using Ovarian Cancer dataset.

For Expriment-2, Figures 6 to 13 show all obtained results when using the RF feature selection algorithm to select the depended ratios from the specified datasets when using the depended classifiers. Hence, there are two figures for each classifier (one represents the accuracy and the other represents the F1 values). For the SVM classifier, the best values of both (accuracy and F1) obtained when depending the 5% ratio of Ovarian Cancer, as shown in Figures 6 and 7. While, for the GNB classifier, the best values obtained when using MLL dataset with 5%, as shown in Figures 8 and 9. Moreover, for the LDA classifier, the best values obtained when depending the Ovarian Cancer with 5% ratio, as shown in Figures 10 and 11. In addition, the Ovarian Cancer dataset with 4% produced best values when using the KNN classifier, as shown in Figures 12 and 13.

For Expriment-3, Figures 14 to 21 show all obtained results when using the ANOVA feature selection algorithm for the same purpose. For the SVM classifier, SRBCT dataset with ratios (3%, 4%, and 5%) gave best values beside the Ovarian Cancer dataset which has the same values but for all depended ratios, as shown in Figures 14 and 15. Adding to that, for the GNB classifier, the best values obtained when using (2%, 3%, and 4%) ratios of the SRBCT dataset, as shown in Figures 16 and 17. For the LDA classifier, the best values obtained when depending the Ovarian Cancer with (3% and 5%) ratios, as shown in Figures 18 and 19. Finally, the SRBCT dataset with (2%, 3%, 4%, and 5%) produced best values when using the KNN classifier, as shown in Figures 20 and 21.

For Expriment-4, Figures 22 to 29 show all obtained results when using the Chi-square feature selection algorithm for the same purpose. For the SVM classifier, SRBCT dataset with 1% ratio gave best values beside the Ovarian Cancer dataset, which has the same values, but for the (2%, 3%, 4%, and 5%) ratios, as shown in Figures 22 and 23. The SRBCT dataset with all depended ratios produced best values for the GNB classifier, as shown in Figures 24 and 25. Then, for the LDA classifier, the best values obtained when depending the Ovarian Cancer with (4% and 5%) ratios, as shown in Figures 26 and 27. In addition, the SRBCT dataset for all ratios provided best values when using the KNN classifier, as shown in Figures 28 and 29.

Overall, the performance variations among classifiers and feature selection methods can be attributed to several factors. Firstly, the dataset size plays a crucial role. Larger datasets tend to provide more representative samples, which can enhance the performance of classifiers. Similarly, complex datasets with intricate relationships between features may pose challenges for certain classifiers and feature selection methods. Additionally, noise levels in the datasets can impact performance. Noisy datasets with high levels of irrelevant or misleading information may negatively affect the performance of classifiers. Feature selection methods that effectively filter out noise can lead to improved classification accuracy. Moreover, the inherent properties of classifiers and feature selection methods also contribute to performance differences. Each classifier has its own assumptions, strengths, and limitations that may align differently with the characteristics of the datasets. Similarly, feature selection methods have different strategies for identifying relevant features, which can impact their effectiveness depending on the dataset's characteristics. By considering these factors, researchers and practitioners can gain insights into the observed performance differences and make informed decisions when selecting classifiers and feature selection methods for specific datasets. This understanding allows for more accurate and reliable microarray data classification.

6. CONCLUSIONS

In this research, an efficient approach is produced which capable of distinguishing among three significant feature selection algorithms (RF, Chi Square, and ANOVA). These important algorithms implemented for microarray data classification to show the effectiveness of each one. The proposed approach passing through three stages: data preprocessing, followed by feature selection, and ended by feature classification. The focusing of our proposed approach concerned with accuracy enhancing. Consequently, the implementation provided capabilities of selecting the most relevant features that represent the datasets, which accordingly causes to increase the accuracy.

Based on the results obtained using all genes of all datasets, we recommend relying on the SVM classifier to achieve the highest accuracy and F1 scores. The LDA classifier also demonstrates competitive performance and can be considered as a secondary option. When using the RF feature selection algorithm with different dataset ratios, it is preferable to utilize the SVM classifier, which yields the best results when utilizing 5% of the datasets. For the ANOVA feature selection algorithm and dataset ratios of 3%, 4%, and 5%, the SVM classifier consistently delivers superior outcomes. Lastly, when employing the Chi-square feature selection algorithm with all dataset ratios, both the SVM and GNB classifiers exhibit the best performance.

Our research has important implications for real-world applications involving microarray data. By following our recommendations, practitioners can improve accuracy and performance in microarray data analysis, leading to more reliable predictions for tasks like disease diagnosis and treatment response prediction. These recommendations consider specific dataset characteristics, enabling practitioners to select the most suitable feature selection algorithms and classifiers. The insights gained from our research are transferable to similar microarray datasets, reducing the need for trial and error in selecting appropriate techniques. Additionally, the principles we present can be applied to other domains with high-dimensional data, benefiting researchers and practitioners in bioinformatics, genomics, and data mining.

The findings of our research advance the understanding of microarray data classification and feature selection in several ways. Firstly, by comparing and evaluating multiple feature selection methods, we contribute to the existing knowledge base regarding their performance and suitability in microarray data analysis. Secondly, our recommendations for method selection based on dataset characteristics provide practical guidelines for researchers and practitioners, enabling them to make more informed decisions in their work. Lastly, by discussing the implications of our research for real-world applications, we bridge the gap between theoretical advancements and practical implementation, enhancing the effectiveness and reliability of microarray data analysis in real-world scenarios.

REFERENCES

[1] Sarbazi-Azad, S., Abadeh, M.S., Mowlaei, M.E. (2020). Using data complexity measures and an evolutionary cultural algorithm for gene selection in microarray data. Soft Computing Letters, 3: 100007. https://doi.org/10.1016/j.socl.2020.100007

[2] Espezua, S., Villanueva, E., Maciel, C.D., Carvalho, A. (2015). A Projection Pursuit framework for supervised dimension reduction of high dimensional small sample datasets. Neurocomputing, 149: 767-776. https://doi.org/10.1016/j.neucom.2014.07.057

[3] Seo, M., Oh, S. (2013). A novel divide-and-merge classification for high dimensional datasets. Computational Biology and Chemistry, 42: 23-34. https://doi.org/10.1016/j.compbiolchem.2012.10.005

[4] Xie, H.Z., Li, J., Zhang, Q.S., Wang, Y.D. (2016). Comparison among dimensionality reduction techniques based on Random Projection for cancer classification. Computational Biology and Chemistry, 65: 165-172. https://doi.org/10.1016/j.compbiolchem.2016.09.010

[5] Tabakhi, S., Najafi, A., Ranjbar, R., Moradi, P. (2015). Gene selection for microarray data classification using a novel ant colony optimization. Neurocomputing, 168: 1024-1036. https://doi.org/10.1016/j.neucom.2015.05.022

[6] Du, D.J., Li, K., Li, X., Fei, M.R. (2014). A novel forward gene selection algorithm for microarray data. Neurocomputing, 133: 446-458. https://doi.org/10.1016/j.neucom.2013.12.012

[7] Dashtban, M., Balafar, M., Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics, 110(1): 10-17. https://doi.org/10.1016/j.ygeno.2017.07.010

[8] Elyasigomari, V., Lee, D.A., Screen, H.R., Shaheed, M.H. (2017). Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. Journal of Biomedical Informatics, 67: 11-20. https://doi.org/10.1016/j.jbi.2017.01.016

[9] Bolón-Canedo, V., Sánchez-Marono, N., Alonso-Betanzos, A., Benítez, J.M., Herrera, F. (2014). A review of microarray datasets and applied feature selection methods. Information Sciences, 282: 111-135. https://doi.org/10.1016/j.ins.2014.05.042

[10] Gao, L.Y., Ye, M.Q., Lu, X.J., Huang, D.B. (2017). Hybrid method based on information gain and support vector machine for gene selection in cancer classification. Genomics, Proteomics & Bioinformatics, 15(6): 389-395. https://doi.org/10.1016/j.gpb.2017.08.002

[11] Wang, A., An, N., Yang, J., Chen, G., Li, L., Alterovitz, G. (2017). Wrapper-based gene selection with Markov blanket. Computers in Biology and Medicine, 81: 11-23. https://doi.org/10.1016/j.compbiomed.2016.12.002

[12] Sahu, B., Mishra, D. (2012). A novel feature selection algorithm using particle swarm optimization for cancer microarray data. Procedia Engineering, 38: 27-31. https://doi.org/10.1016/j.proeng.2012.06.005

[13] Brusco, M.J. (2014). A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. Computational Statistics & Data Analysis, 77: 38-53. https://doi.org/10.1016/j.csda.2014.03.001

[14] Goswami, S., Saha, S., Chakravorty, S., Chakrabarti, A., Chakraborty, B. (2015). A new evaluation measure for feature subset selection with genetic algorithm. International Journal of Intelligent Systems and Applications, 7(10): 28-36. https://doi.org/10.5815/ijisa.2015.10.04

[15] Guo, S., Guo, D.H., Chen, L.F., Jiang, Q.S. (2016). A centroid-based gene selection method for microarray data classification. Journal of Theoretical Biology, 400: 32-41. https://doi.org/10.1016/j.jtbi.2016.03.034

[16] Arshak, Y., Eesa, A. (2018). A new dimensional reduction based on cuttlefish algorithm for human cancer gene expression. In 2018 International Conference on Advanced Science and Engineering (ICOASE), pp. 48-53. http://doi.org/10.1109/ICOASE.2018.8548908

[17] Babu, M.M. (2004). Introduction to Microarray Data Analysis. In: Berrar, D.P., Dubitzky, W., Granzow, M. (eds) A Practical Approach to Microarray Data Analysis. Springer, Boston, MA. https://doi.org/10.1007/0-306-47815-3_1

[18] Read, J., Brenner, S. (2001). Microarray Technology, in Encyclopedia of Genetics. New York, NY, USA: Academic, 1191.

[19] Othman, M.S., Kumaran, S.R., Yusuf, L.M. (2020). Gene selection using hybrid multi-objective cuckoo search algorithm with evolutionary operators for cancer microarray data. IEEE Access, 8: 186348-186361. https://doi.org/10.1109/ACCESS.2020.3029890

[20] Lai, C.M., Huang, H.P. (2021). A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique. Applied Soft Computing, 100: 106994. https://doi.org/10.1016/j.asoc.2020.106994

[21] Pino Angulo, A. (2018). Gene selection for microarray cancer data classification by a novel rule-based algorithm. Information, 9(1): 6. https://doi.org/10.3390/info9010006

[22] Hegde, R.B., Prasad, K., Hebbar, H., Singh, B.M.K., Sandhya, I. (2020). Automated decision support system for detection of leukemia from peripheral blood smear images. Journal of Digital Imaging, 33: 361-374. https://doi.org/10.1007/s10278-019-00288-y

[23] Harun, N.H., Bakar, J.A., Abd Wahab, Z., Osman, M.K., Harun, H. (2020). Color image enhancement of acute leukemia cells in blood microscopic image for leukemia detection sample. In 2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp. 24-29. https://doi.org/10.1109/ISCAIE47305.2020.9108810

[24] Hasri, N.M., Wen, N.H., Howe, C.W., Mohamad, M.S., Deris, S., Kasim, S. (2017). Improved support vector machine using multiple SVM-RFE for cancer classification. International Journal on Advanced Science, Engineering and Information Technology, 7(4-2): 1589-1594. https://doi.org/10.18517/ijaseit.7.4-2.3394

[25] Tang, Y., Zhang, Y.Q., Huang, Z. (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4(3): 365-381. https://doi.org/10.1109/TCBB.2007.1028

[26] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6): 520-525. https://doi.org/10.1093/bioinformatics/17.6.520

[27] Yang, P., Zhou, B.B., Zhang, Z., Zomaya, A.Y. (2010). A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data. BMC Bioinformatics, 11(1): 1-12. https://doi.org/10.1186/1471-2105-11-S1-S5

[28] Vergara, J.R., Estévez, P.A. (2014). A review of feature selection methods based on mutual information. Neural Computing and Applications, 24(1): 175-186. https://doi.org/10.1007/s00521-013-1368-0

[29] Hira, Z.M., Gillies, D.F. (2015). A review of feature selection and feature extraction methods applied on microarray data. Advances in Bioinformatics, 2015: 198363. https://doi.org/10.1155/2015/198363

[30] Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8): 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[31] He, X.F., Cai, D., Niyogi, P. (2005). Laplacian score for feature selection. Advances in Neural Information Processing Systems, 18.

[32] Hall, M. A. (1999). Correlation-based feature selection for machine learning. Doctoral Dissertation, the University of Waikato.

[33] Chuang, L.Y., Yang, C.H., Wu, K.C., Yang, C.H. (2011). A hybrid feature selection method for DNA microarray data. Computers in Biology and Medicine, 41(4): 228-237. https://doi.org/10.1016/j.compbiomed.2011.02.004

[34] Saeys, Y., Inza, I., Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19): 2507-2517. https://doi.org/10.1093/bioinformatics/btm344

[35] Moradi, P., Gholampour, M. (2016). A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Applied Soft Computing, 43: 117-130. https://doi.org/10.1016/j.asoc.2016.01.044

[36] McCall, J. (2005). Genetic algorithms for modelling and optimization. Journal of Computational and Applied Mathematics, 184(1): 205-222. https://doi.org/10.1016/j.cam.2004.07.034

[37] Xu, Y.F., Fan, P., Yuan, L. (2013). A simple and efficient artificial bee colony algorithm. Mathematical Problems in Engineering, 2013: 526315. https://doi.org/10.1155/2013/526315

[38] Wang, D.S., Tan, D.P., Liu, L. (2018). Particle swarm optimization algorithm: An overview. Soft Computing, 22(2): 387-408. https://doi.org/10.1007/s00500-016-2474-6

[39] Saremi, S., Mirjalili, S., Lewis, A. (2017). Grasshopper optimisation algorithm: Theory and application. Advances in Engineering Software, 105: 30-47. https://doi.org/10.1016/j.advengsoft.2017.01.004

[40] Brownlee, J. (2011). Clever Algorithms: Nature-Inspired Programming Recipes. Jason Brownlee.

[41] Jović, A., Brkić, K., Bogunović, N. (2015). A review of feature selection methods with applications. In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 1200-1205. https://doi.org/10.1109/MIPRO.2015.7160458

[42] Venkatesh, B., Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1): 3-26. https://doi.org/10.2478/cait-2019-0001

[43] Lu, H.J., Chen, J.Y., Yan, K., Jin, Q., Xue, Y., Gao, Z.G. (2017). A hybrid feature selection algorithm for gene expression data classification. Neurocomputing, 256: 56-62. https://doi.org/10.1016/j.neucom.2016.07.080

[44] Pashaei, E., Ozen, M., Aydin, N. (2016). Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA. In 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 308-311. https://doi.org/10.1109/BHI.2016.7455896

[45] Yin, Y., Kaku, I., Tang, J.F., Zhu, J.M. (2011). Data mining: Concepts, methods and applications in management and engineering design. Springer Science & Business Media.

[46] Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3): 273-297. https://doi.org/10.1007/BF00994018

[47] Cover, T., Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1): 21-27. https://doi.org/10.1109/TIT.1967.1053964

[48] Griffis, J.C., Allendorfer, J.B., Szaflarski, J.P. (2016). Voxel-based Gaussian naïve Bayes classification of ischemic stroke lesions in individual T1-weighted MRI scans. Journal of Neuroscience Methods, 257: 97-108. https://doi.org/10.1016/j.jneumeth.2015.09.019

[49] Altay, O., Ulas, M. (2018). Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In 2018 6th international symposium on digital forensic and security (ISDFS), pp. 1-4. https://doi.org/10.1109/ISDFS.2018.8355354

[50] Shashoa, N.A.A., Salem, N.A., Jleta, I.N., Abusaeeda, O. (2016). Classification depend on linear discriminant analysis using desired outputs. In 2016 17th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA), pp. 328-332. https://doi.org/10.1109/STA.2016.7952041

[51] Jain, S., Shukla, S., Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications, 106: 252-262. https://doi.org/10.1016/j.eswa.2018.04.008

[52] Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R. S., Moore, J.H. (2018). Relief-based feature selection: Introduction and review. Journal of Biomedical

Informatics, 85: 189-203. https://doi.org/10.1016/j.jbi.2018.07.014

[53] Baliarsingh, S.K., Vipsita, S., Dash, B. (2020). A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm. Neural Computing and Applications, 32(12): 8599-8616. https://doi.org/10.1007/s00521-019-04355-x

[54] Ross, A.H. (1999). Algorithm for calculating the noncentral chi-square distribution. IEEE Transactions on Information Theory, 45(4): 1327-1333. https://doi.org/10.1109/18.761294

[55] Haryanto, A.W., Mawardi, E.K. (2018). Influence of word normalization and chi-squared feature selection on support vector machine (SVM) text classification. In 2018 International Seminar on Application for Technology of Information and Communication, pp. 229-233. https://doi.org/10.1109/ISEMANTIC.2018.8549748

[56] Shukla, A.K., Singh, P., Vardhan, M. (2018). A two-stage gene selection method for biomarker discovery from microarray data for cancer classification. Chemometrics and Intelligent Laboratory Systems, 183: 47-58. https://doi.org/10.1016/j.chemolab.2018.10.009

NOMENCLATURE

x'	min-max scaling/normalization for features
x	value of the feature
N	number of training instances
M	number of random training instances out of n
W	weights of features
x_c^2	value of chi-square test
O	observed value(s)
E	expected value(s)

Greek symbols

α	number of features

Subscripts

c	degree of freedom