A Hybrid Feature Selection Based on Fisher Score and SVM-RFE for Microarray Data

Hind Hamla[1], Khadoudja Ghanem[2]
[1]Laboratory of Modelling and Implementation of Complex System, Department of Computer Science, University of Abdelhamid Mehri Constantine 2, Constantine, Algeria
[2]Laboratory of Modelling and Implementation of Complex System, Department of Computer Science, University of Abdelhamid Mehri Constantine 2, Constantine, Algeria
E-mail: hind.hamla@univ-constantine2.dz , khadoudja.ghanem@univ-constantine2.dz

Keywords: SVM-RFE, Fisher score, gene selection, microarray data

Received: March 22, 2023

Microarray data analysis has played a significant role in disease diagnosis and tumor type identification over the last two decades. However, due to the curse of dimensionality issues, microarray data classification remains a challenging task. This issue arises from a situation where the number of features is large, but the number of samples is small. As a result, dimension reduction techniques, specifically feature selection methods, are critical for removing non-informative features and improving cancer classification. This paper presents a Filter-embedded hybrid feature selection method to address the gene selection challenge in microarray data analysis. First, it selects the features with the highest Fisher score to create a candidate subset for the next embedded stage. Second, the proposed method employs support vector machine-recursive feature elimination (SVM-RFE) on the candidate subset to identify the optimal set of features to enhance cancer classification. Extensive experiments were conducted with ten high-dimensional microarray datasets to assess the efficacy of the proposed approach. The results show that the proposed method improves classifier performance significantly regarding classification accuracy, number of selected features, and computational efficiency.

Povzetek: Predstavljena je hibridna metoda izbire značilk z uporabo Fisherjeve ocene in SVM-RFE za izboljšanje natančnosti klasifikacije raka z analizo mikromrežnih podatkov.

1 Introduction

Over the last two decades, advances in microarray technology have enabled researchers to analyze thousands of genes simultaneously, which has been used in various applications such as disease classification [3]. Microarray data classification is an effective tool for early disease diagnosis and determining disease subtypes [9]. However, due to the curse of dimensionality, where the number of features is remarkably large (often thousands of features) while the number of samples is limited (often tens of samples), this task poses a significant challenge for machine learning algorithms [5]. In addition, a significant proportion of genes are irrelevant or redundant, affecting classifier performance [4]. Thus, gene selection methods have emerged as effective approaches for reducing dimensionality in microarray data. Gene selection methods seek to identify and eliminate redundant and irrelevant features to obtain a subset of the most informative features [32]. These methods have improved classification accuracy while reducing computational costs associated with classifiers [34].

Gene selection methods are broadly classified as filter, wrapper, and embedded methods. Filter methods select features independently from the learning classifier, based on statistical properties [3]. These methods are fast, but they produce a low classification accuracy [15]. The wrapper methods use the learning algorithm to evaluate a subset of selected features [3].Although they produce higher classification accuracy, they are computationally expensive. Therefore, when dealing with high-dimensional data, these methods are avoided [6]. Embedded methods select features during the learning process [31]. They are appropriate for analyzing microarray data due to their reduced computational demands compared to wrapper methods and enhanced efficiency compared to filter methods. [5]. Hybrid methods, which sequentially combine two or more feature selection methods from the same or different conceptual origins, have recently emerged [6] to leverage the strengths of diverse methodologies.

Many feature selection (FS) surveys for microarray data processing have been conducted. [2] compares feature selection methods including information gain, twoing rule, sum minority, max minority, Gini index, sum of variances, t-statistics, and one-dimension support vector machines. This study use two publicly available glioma gene expression datasets for evaluation. It was discovered that feature selection is important in the classification of gene expression data. In [7], the authors examined the importance and challenges of feature selection methods when dealing with high-dimensional data such as microarray and instruction detection. The paper emphasized the importance of efficient techniques for managing the computational complexity of high-dimensional data. Furthermore, open issues in feature selection are addressed, particularly in the context of big data and high-dimensional datasets.

The authors of [8] compared five filter methods: the F test, the T-test, the signal-to-noise ratio (S/R), ReliefF, and the Pearson product-moment correlation coefficient (CC). The study used five microarray datasets: leukemia, lung cancer, lymphoma, central nervous system cancer, and ovarian cancer. The results showed that combining the signal-to-noise ratio (S/R) with KNN classifiers produced the best classification accuracy. In [13], the researchers investigated the effect of popular filter methods (ReliefF, Mutual information, Chi-square, F-score, Fisher score, Laplacian, MRMR, and CMIM) on six well-known classifiers (random forest, logistic regression, K-Nearest Neighbour, decision tree, and Support Vector Machine). The experiment was carried out on ten high-dimensional microarray datasets, and the results revealed a distinct trend. Univariate filter feature selection techniques such as Mutual Information, F-score, and Fisher score outperformed multivariate techniques such as MRMR and CMIM. Only a few studies on embedded methods have been conducted. [12] assessed the efficacy of five embedded feature selection techniques: decision trees, random forests, lassos, ridges, and SVM-RFE. The experiment employed ten high-dimensional microarray datasets. The results highlight the SVM-RFE's superior accuracy performance.

This paper combines the embedded method's performance with the filter method's computational efficiency. the proposed method is divided into two stages: The Fisher score filter method is used in the first stage to select the most relevant features due to its effective performance with high-dimensional data [10]. Second, the selected subset is input for the embedded Support Vector Machine Recursive Feature Elimination SVM-RFE method. This combination improves classification accuracy while significantly reducing the number of selected features. Experiments were conducted on ten high-dimensional microarray datasets, including Colon, Central Nervous System CNS, Leukemia, Breast cancer, Lung cancer, Leukemia3-Classes, Leukemia4-Classes, Ovarian, Lymphoma, and MLL. The experimental setup consists of three major components:

– A comparative analysis of the proposed method with other filter methods combined with the same embedded method, SVM-RFE, specifically ReliefF_SVM-RFE and Mutual Information (MI)_SVM-RFE. In addition, we present SVM-RFE results without using a filter method. We avoid comparing the proposed method to the Minimum-Redundancy Maximum-Relevancy (MRMR) and Chi-square filter methods because they have already been studied [19] and [4].

– Investigation the impact of employing six well-established classifiers: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and K-Nearest Neighbour (KNN) on the feature subset selected by our proposed method.

– Finally, to highlight the effectiveness of the proposed method, we compared it with filter-wrapper methods ([30], [34], [23], [21], [24]) and with filter-embedded ones [19] and [4].

The paper is structured as follows: Section 2 examines related works on hybrid feature selection methods. Section 3 briefly describes the Fisher score algorithm and the SVM-RFE algorithm. Section 4 describes the proposed method in depth. Section 5 presents a comprehensive analysis of the experimental findings. Finally, Section 6 provides the conclusion and outlines potential future directions.

2 Related work

Numerous hybrid feature selection methods have been proposed to address the dimensionality reduction challenge and eliminate irrelevant and redundant features from microarray data. While most existing studies in the literature combine filter methods and wrapper methods [1], only a few works investigate the combination of embedded methods and filter methods. In this section, we will review some recent hybrid feature selection methods that have been published in the literature.

2.1 Hybrid wrapper-filter methods

Given their adaptability and efficiency in dealing with large-scale issues, meta-heuristics methods have attracted attention for solving gene selection problems [26]. However, these methods frequently necessitate a significant amount of computational time. Therefore, meta-heuristics have been combined with filter methods to narrow the search space and speed up the feature selection process [21]. Naik et al. [20] proposed a hybrid feature selection method combining the filter and wrapper methods. The Fisher score filter method was used to select a subset of features. The Binary Dragonfly Algorithm was used in the wrapper method to search for an informative subset of features, and the Radial Basis Function Neural Network was used as the learning model that evaluates the selected subset. Shukla [24] designed HMPAGA, a hybrid feature selection method that used an ensemble gene selection method to filter out noisy and redundant genes. It also used a multi-population adaptive genetic algorithm to identify high-risk difference genes. SVM and NB classifiers were used as objective functions.

Shulka et al. [25] proposed a two-stage feature selection method for microarray data recognition. In the first stage, noisy and redundant features were removed using a multi-layer approach and f-score filter methods. An adaptive genetic algorithm selected the most important features in the second stage. Zhang et al. [30] proposed IG-MBKH, a hybrid feature selection method that combines Information Gain and Modified Binary Krill Herd. The method was validated using nine high-dimensional microarray datasets, improving classification accuracy with fewer features. Zheng et al. [34] presented the K Value Maximum Reliability Minimum Redundancy Improved Grey Wolf Optimizer (KMR2IGWO), a hybrid feature selection method. MRMR was used in the filter stage to select K features, with K determined by the dataset's message. These features were then used as input for the IGHO algorithm, with the SVM classifier used to assess classification accuracy. KMR2IGWO's performance was validated using 14 microarray datasets, highlighting its superiority.

MIMAGA, a combination of mutual information maximization and adaptive genetic algorithm (AGA), was introduced by Lu et al. [17]. MIM was used to choose a subset of 300 features. Then, AGA was applied with the accuracy of ELM classifier serving as the fitness function. Sadeghian et al. [23]introduced a three-stage hybrid feature selection method named Ensemble Information Theory-based binary Butterfly Optimization Algorithm (EIT-bBOA). The method employed Minimal Redundancy-Maximal New Classification Information (MR-MNCI) in the initial phase to eliminate 80% irrelevant features. Subsequently, the Information Gain-binary butterfly optimization algorithm (IG-bBOA) optimized the first phase. In the final phase, an ensemble of ReliefF and the Fisher Score method was applied to the final feature subset. The method was evaluated using six well-known datasets. Ouadfel et al. [21] developed a two-stage feature selection method that used the ReliefF filter method to estimate feature relevance in the first stage. The top-ranked M features where then preselected. The second stage combined the binary Equilibrium Optimizer with a local search strategy based on Pearson coefficient correlation. The proposed method was evaluated on 16 UCI datasets and ten high-dimensional biological datasets.

2.2 Hybrid embedded-filter methods

In terms of computational time, embedded feature selection methods outperform wrapper methods. Though only a few embedded methods have been presented in the literature, [12] conducted a comparative study of the most common ones. SVM-RFE emerged as the most accurate method, with comparable execution time and selected features,. Furthermore, SVM-RFE has consistently demonstrated its efficacy [16]. Thus, many studies have proposed hybridization between filter and embedded methods that concentrate on combining SVM-RFE with filter methods. SVM-RFE has been shown to be effective in identifying informative genes in microarray data [33]. Mundra et al. [19] proposed a hybrid feature selection method combining MRMR and SVM-RFE. The approach's performance was assessed on four well-known microarray datasets. Almutiri and Saeed [4] introduced the ChiSVMRFE feature selection method based on the Chi Square Statistic and SVM-RFE. On ten microarray datasets, the proposed method was evaluated. Mishra et al. [18] combined SVM-RFE with the Bayesian T-test for gene selection, which resulted in improved classification accuracy, fewer selected genes, and a lower classification error rate.

Huang et al. [14] enhanced the SVM-RFE's performance for gene selection by incorporating feature clustering, thereby reducing computational complexity and gene redundancy. Li et al. [16] proposed VSSRFE, an improved version of SVM-RFE that aimed to reduce time using a more efficient SVM classifier implementation. The results demonstrated the proposed method's efficiency in terms of time reduction. Combining wrapper or embedded methods with filter methods consistently improves classifier performance in terms of classification accuracy and computational efficiency, according to the aforementioned works. SVM-RFE, in particular, has demonstrated its ability to improve classification accuracy while optimizing feature dataset. This paper combines SVM-RFE, a leading embedded method, with the best filter method to further improve the results.

3 Background

This section describes the Fisher score and SVM-RFE methods.

3.1 Fisher score

The Fisher score algorithm is a well-known filter feature selection method that is applied to a subset of discriminative features. In summary, the algorithm works as follows: It begins by calculating the average and variance of each feature for each class. Then, it calculates scatter matrices between and within classes to assess the effectiveness of the features in differentiating various classes. The Fisher Scores are then calculated using these matrices, allowing for comparing different features. Features with higher Fisher Scores are considered more important for distinguishing between classes. We can rank the features and select the best based on their scores. The goal is to minimize the distances between samples in the same class while increasing the distances between samples in different classes [29]. Fisher scores f_i are calculated as follows:

$$SC_F\left(f_i\right) = \frac{\sum_{j=1}^{c} n_j \left(\mu_{i,j} - \mu_i\right)^2}{\sum_{j=1}^{c} n_j \sigma_{i,j}^2} \tag{1}$$

where, u_i is the mean of f_i feature, $n_j is$ the number of samples in the class j^{th}, $u_i j$ is the mean of f_i in the j^{th} class, and σ_{ij} is the variance of f_i in the j^{th} class. Usually, a higher Fisher score means the feature is vital for classification.

3.2 Support Vector Machine Recursive Feature Elimination (SVM-RFE)

SVM-RFE is an embedded feature selection method introduced by Guyon et al. [11]. This method employs a weight vector as a criterion for splitting, calculated as follows:

$$\mathrm{W} = \sum_{i=1}^{n} (y_i, x_i, \alpha_i) \tag{2}$$

where, i represents the number of features ranging from 1 to n, y_i is the labeled class of the sample x_i. α_i is the maximum class separation margin estimated from the training set. SVM-RFE works in a recursive manner, similar to iterative refinement. The entire feature set is initially used to train an SVM classifier. The algorithm then iteratively eliminates features with the lowest discriminative power, reducing the risk of the curse of dimensionality and overfitting. The features are then ranked according to their contribution to the classification task. The i^{th} ranking criterion is calculated as follows:

$$\mathrm{R} = W^2 \tag{3}$$

The higher the value of the ranking criterion, the more important the feature. Algorithm 1 depicts the detailed SVM-RFE algorithm.

Algorithm 1 Pseudocode of SVM-RFE

```
Input: F initial feature set
Output: R rank list
$R = \emptyset$
while $F \neq \varnothing$ do
    Train SVM with F
    Compute the weight vector using Equation 2
    Compute the ranking criterion using Equation 3
    Find feature with the lowest ranking criterion
    Update the Ranked list of features
    $\mathrm{R} = \mathrm{R} + \mathrm{F_i}$
    Update set of features
    $\mathrm{F} = \mathrm{F} - \mathrm{F_i}$
end while
```

4 Proposed method

Because of its low computational requirement, the Fisher score is a simple and efficient feature selection method that is particularly suitable for high-dimensional microarray data classification [28]. However, the Fisher score does not achieve satisfactory classification accuracy. SVM-RFE, on the other hand, has been successfully applied to gene selection problems. It has consistently outperformed several other embedded methods regarding classification accuracy while using a smaller feature set [12]. Nonetheless, one major disadvantage of SVM-RFE is the lengthy feature selection process, especially when dealing with high-dimensional data such as microarray [16]. This work proposes a hybrid feature selection method that combines the computational efficiency of the Fisher score filter method and the high performance of the SVM-RFE embedded method to capitalize on the strengths of both. Fig. 1 shows the flowchart of the hybrid filter-embedded method.

The following are the specifics of the proposed method:

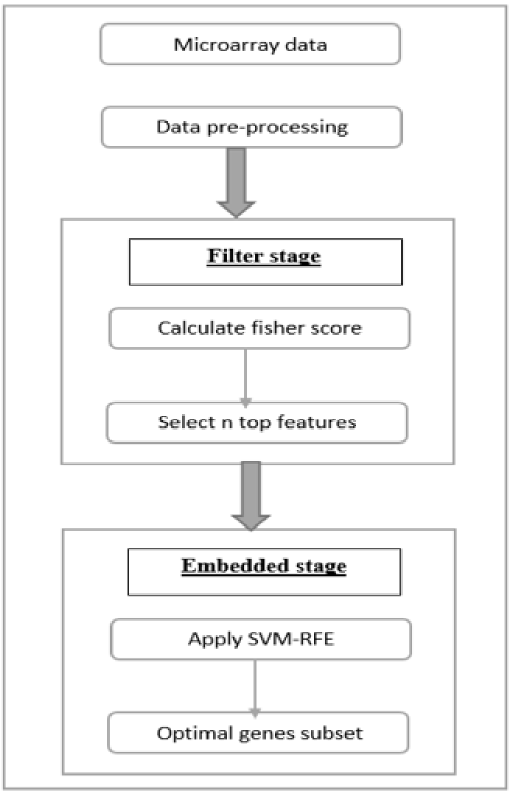

Figure 1: Flowchart of the proposed method.

1. Data pre-processing
   This first step involves replacing missing values with the mean value derived from all known gene values.

2. Filter stage

   Calculate Fisher score
   The Fisher score is used at this stage to eliminate redundant and irrelevant features. Eq. (1) calculates the Fisher score value for each feature, and the features are then sorted based on these values. The higher the Fisher score value, the more informative the feature is for classification.

   Select n top features
   The top n features the Fisher score method indicates are selected as candidate input for the embedded stage.

3. Embedded stage
   SVM-RFE is applied to the previously selected candidate inputs. SVM-RFE uses all the selected features

to train the SVM classifier. Each iteration removes the features with the lowest ranking criterion from the features set. This process is repeated until all features have been removed. The features are sorted in reverse order of removal, with the most recently removed features considered the most important.

4. Select optimal subset
Finally, SVM-RFE selects a subset of m most important features. The value of n and m is determined through experimentation, with m always being less than n (m<n). This selected subset constitutes the set of informative genes for classification.

The proposed hybrid feature selection approach effectively addresses the challenges of high-dimensional microarray data by combining the Fisher score and SVM-RFE methods. The classification accuracy and interpretability can be improved by selecting a small but informative subset of genes. This has the potential to greatly aid in disease diagnosis and tumor classification. Furthermore, the proposed method balances computational efficiency with classification performance, thereby contributing to bioinformatics and microarray data analysis.

5 Experimental results

In this section, we describe the experimental setup employed to evaluate our hybrid method'efficacy for genes selection from high-dimensional microarray datasets. The goal is to evaluate the efficacy of SVM-RFE when combined with MI, ReliefF, and Fisher scores to determine the best filter method for a microarray dataset using SVM-RFE. Furthermore, the selected gene subset will be tested using a variety of classifiers, including SVM, LR, DT, RF, NB, and KNN. The proposed method is then compared to other existing hybrid feature selection methods. We used a personal computer with an Intel Core i7 processor, 2.9 GHz, and 8 GB of RAM to conduct the experiments. The results presented in this paper are an average of five runs.

5.1 Datasets description

The proposed method is evaluated on ten high-dimensional microarray datasets [35]. The datasets include 2-classes, 3-classes, 4-classes, 5-classes. The number of samples in these datasets is ranged from 60 to 253, while the number of features in these datasets is ranged from 2,000 to 24,481. Table 1 presents detailed information about these datasets. For the evaluation step, we employ 10-fold cross-validation. In this procedure, the datasets are randomly divided into training and testing data subsets, with an 80% and 20% proportion, respectively. The final results are obtained through averaging fold outcomes, a practice employed to address potential issues related to class imbalance.

5.2 Performance measure

Cross-validation [27] is a well-known method for determining the misclassification rate. The data is randomly divided into k subsets of approximately equal size in k-fold cross-validation. The classifier is trained on k-1 folds and then tested on the last fold. This procedure is repeated until every k-fold is used as the test sub-set. The average of the recorded scores is used as the performance metric. In this work, we use several performance metrics, including accuracy, recall, precision, and F-measure, in addition to execution time, to assess the effectiveness of the proposed method.

Accuracy: the ratio of samples that are correctly predicted:

$$\text{Accuracy } = \frac{TP+TN}{TP+FN+FP+TN} \tag{4}$$

Recall: the ratio of the positive samples that are predicted as positive:

$$\text{Recall } = \frac{TP}{TP+FN} \tag{5}$$

Precision: the ratio of the positive prediction that is correct:

$$\text{Precision } = \frac{TP}{TP+FP} \tag{6}$$

F-measure: is a harmonic mean of the precision and recall:

$$\text{F-measure } = \frac{2 * \text{ Precision } * \text{ Recall}}{\text{Presision } + \text{ Recall}} \tag{7}$$

The indicators for evaluation are:
Acc: Accuracy
Rec: Recall
Pre: Precision
Fmes: F-measure
Nb-FS: number of selected features

5.3 Combination of filter methods with SVM-RFE

Table 2 displays the results of various feature selection methods combined with SVM-RFE. Using the gene subsets selected by these methods, we assess each classifier's accuracy, recall, precision, F-measure, and execution time. According to Table 2, the classification accuracy of the SVM classifier on the original dataset is not very interesting, especially for the breast and CNS datasets, where the classification accuracy did not reach 70%. However, feature selection methods enhance classification performance regarding accuracy, recall, precision, and F-measure. The proposed method consistently performs comparable or better than other feature selection methods. Notably, the execution time was reduced for all datasets after using feature selection methods. Moreover, the proposed method demonstrates remarkable efficacy by achieving 100% accuracy for nine out of ten datasets, using less than 1%

of the original genes. This finding demonstrates our proposed method's ability to identify informative genes for microarray data analysis. In some cases, SVM-RFE outperforms other feature selection methods, implying that the filter methods have eliminated some important features. Fig. 3 presents the number of selected features. The proposed method clearly achieves higher classification accuracy with fewer than 20 features.

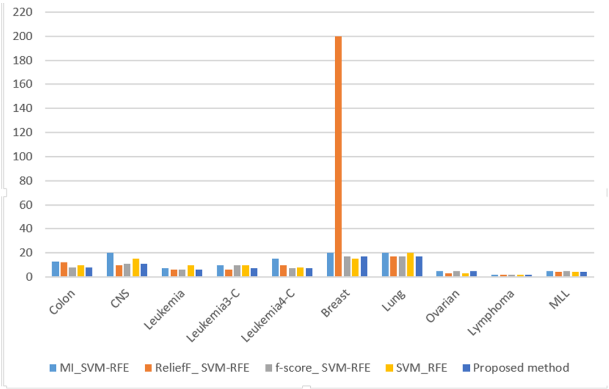

Figure 2: The number of selected features.

5.4 Evaluation of the application of different classifiers on the subset selected by the proposed method

Using the subset of features selected by our proposed method, we compare six popular classifiers: SVM, LR, DT, RF, NB, and KNN. The results in Table 3 indicate that:

- The six classifiers SVM, LR, DT, RF, NB, and KNN achieve comparable classification performance when using the subset of features selected by the proposed method. Based on this result, various classifiers can perform well when using the selected gene subset, indicating that the subset contains relevant and discriminative information.

- SVM consistently outperforms other classifiers regarding accuracy, recall, precision, and F-measure across all datasets. Due to its ability to find optimal hyperplanes for separating data points, SVM is effective for various datasets.

- DT has generally lower accuracy than other classifiers. This may be because it tends to overfit training data, especially when dealing with high-dimensional data sets.

- The fastest classifier is KNN, while RF is the slowest. However, RF still delivers competitive results despite its longer execution time, indicating its ability to handle high-dimensional data effectively.

- On the selected gene subset, the results obtained by the proposed method match those of the SVM classifier. Thus, the proposed method is valid and reliable due to this consistency.

5.5 Comparison of the proposed method with other hybrid methods

The performance of the proposed method was compared with several hybrid feature selection methods available in the literature, including filter-wrapper methods (IG-MBKH [30], KMR2IGWO [34], EIT-bBOA [23], RBEO-LS [21], and HMPAGA [24]). And filter-embedded methods (ChiSVM-RFE [4] and SVM-RFE with MRMR [19]). The comparison is based on classification accuracy and the number of selected features, as shown in Table 4. The symbol "-" means that information is unavailable.

The results Table 4 indicate that the proposed method achieves a comparable classification accuracy while selecting a reduced subset of features. It attains the highest classification accuracy for all datasets except the Colon dataset, with a small number of genes. Moreover, though a direct execution time comparison was not performed, embedded methods consume less time than wrapper methods, as demonstrated in [22]. This finding suggests that the proposed method is more efficient considering the execution time.

6 Conclusion and future work

Microarray data is well known for being high-dimensional and highly redundant. Thus, feature selection methods are critical in removing irrelevant and redundant features. This paper proposes a hybrid feature selection method that combines the Fisher score and SVM-RFE. The proposed method is divided into two stages. The Fisher score filter method selects a candidate subset of features in the first stage. The subset is then used as input for the SVM-RFE to further reduce the number of features to less than 20. The proposed method outperforms other methods such MI_-SVM-RFE, ReliefF_SVM-RFE, and SVM-RFE in terms of accuracy, recall, precision, F-measure, number of selected features, and runtime in experimental evaluations on ten high-dimensional datasets, some of which had over 20,000 features. In addition, we compared the proposed method to several methods proposed in the literature. According to the results, the proposed method consistently achieved higher classification accuracy and selected a smaller number of features for most datasets. These findings demonstrate the efficacy of the proposed method in addressing the challenges of high-dimensional microarray data analysis.

Table 1: Datasets description.

Datasets	Number of instances	Number of features	Number of classes
Colon tumor	62	2000	2
CNS	60	7129	2
Leukemia	72	7129	2
Breast cancer	97	24481	2
Lung_cancer	203	12600	5
Ovarian cancer	253	15154	2
Leukemia 3 classes	72	7129	3
Leukemia 4 classes	72	7129	4
Lymphoma	62	4026	3
MLL	72	12582	3

Table 2: The experimental result of filter-embedded methods.

Dataset	Methods	Acc	Rec	Pre	Fmes	Time	Nb-FS
Colon	No_FS	79.16%	83.33%	77.66%	79.30%	0.0075	-
	SVM-RFE	96.0%	96.66%	97.5%	96.57%	0.001	10
	ReliefF_SVM-RFE	95.5%	91.66%	100.0%	94.66%	0.002	12
	MI_SVM-RFE	98.00%	96.66%	100.0%	98.00%	0.002	10
	Proposed	98.00%	100.0%	97.5%	98.57%	0.001	8
CNS	No_FS	68.83%	60.0%	63.33%	56.66%	0.0322	-
	SVM-RFE	98.00%	95.0%	100.0%	96.66%	0.001	15
	ReliefF_SVM-RFE	98.00%	95.0%	100.0%	96.66%	0.003	9
	MI_SVM-RFE	98.00%	95.0%	100.0%	96.66%	0.003	15
	Proposed	100.0%	100.0%	100.0%	100.0%	0.003	11
Leukemia	No_FS	96.33%	95.0%	96.66%	94.66%	0.061	-
	SVM-RFE	94.57%	86.66%	100.0%	91.33%	0.001	8
	ReliefF_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.001	6
	MI_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.002	6
	Proposed	100.0%	100.0%	100.0%	100.0%	0.001	6
Leukemia 3-C	No_FS	94.57%	97.0%	94.0%	96.0%	0.056	-
	SVM-RFE	100.0%	98.0%	99.0%	99.0%	0.001	10
	ReliefF_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.004	7
	MI_SVM-RFE	98.57%	99.0%	95.0%	97.0%	0.005	7
	Proposed	100.0%	98.0%	99.0%	99.0%	0.0	7
Leukemia 4-C	No_FS	91.29%	98.0%	96.0%	96.0%	0.0663	-
	SVM-RFE	100.0%	99.0%	99.0%	99.0%	0.0009	8
	ReliefF_SVM-RFE	98.57%	99.0%	99.0%	99.0%	0.001	8
	MI_SVM-RFE	100.0%	99.0%	99.0%	99.0%	0.0	7
	Proposed	100.0%	99.0%	99.0%	99.0%	0.0	7
Breast	No_FS	65.47%	50.0%	57.49%	52.66%	0.5012	-
	SVM-RFE	98.57%	100.0%	97.5%	98.57%	0.001	15
	ReliefF_SVM-RFE	94.64%	93.33%	95.0%	93.14%	0.0	200
	MI_SVM-RFE	98.57%	96.66%	100.0%	98.0%	0.0	20¨
	Proposed	100.0%	100.0%	100.0%	100.0%	0.002	17
Lung cancer	No_FS	93.90%	91.0%	96.0%	93.0%	0.2905	-
	SVM-RFE	100.0%	99.0%	99.0%	99.0%	0.003	20
	ReliefF_SVM-RFE	98.75%	96.0%	97.0%	97.0%	0.002	17
	MI_SVM-RFE	98.75%	98.0%	96.0%	97.0%	0.003	20
	Proposed	100.0%	97.0%	97.0%	97.0%	0.003	17
Ovarian	No_FS	100.0%	100.0%	100.0%	100.0%	0.4832	-
	SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0	3
	ReliefF_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0	3
	MI_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0	5
	Proposed	100.0%	100.0%	100.0%	100.0%	0.001	5

Lymphoma	No_FS	100.0%	98.0%	90.0%	94.0%	0.0507	-
	SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0005	2
	ReliefF_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0	2
	MI_SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0	2
	Proposed	100.0%	100.0%	100.0%	100.0%	0.001	2
MLL	No_FS	98.57%	100.0%	100.0%	100.0%	0.1222	-
	SVM-RFE	100.0%	100.0%	100.0%	100.0%	0.0005	4
	ReliefF_SVM-RFE	100.0%	98.0%	99.0%	98.0%	0.0	4
	MI_SVM-RFE	100.0%	98.0%	98.0%	98.0%	0.0	5
	Proposed	100.0%	100.0%	100.0%	100.0%	0.001	4

Table 3: The experimental results of applying different classifiers on the selected subset.

		SVM	LR	DT	RF	NB	KNN
Colon	Acc	98.00%	96.0%	79.23%	84.6%	79.0%	85.33%
	Rec	100.0%	96.66%	80.66%	86.99%	73.33%	91.66%
	Pre	97.5%	97.5%	71.93%	81.66%	80.0%	85.83%
	Fmes	98.5	96.57	76.21%	81.76%	75.14%	87.04%
	Time	0.001	0.001	0.001	0.061	0.0006	0.0001
CNS	Acc	100.0%	96.33%	77.83%	86.33%	85.50%	90.33%
	Rec	100.0%	90.0%	59.0%	69.0%	85.0%	85.0%
	Pre	100.0%	100.0%	96.66%	79.0%	81.66%	93.33%
	Fmes	100.0%	93.33%	60.73%	72.99%	81.33%	86.0%
	Time	0.0016	0.001	0.001	0.0970	0.001	0.0
Leukemia	Acc	100.0%	96.33%	92.80%	98.29%	100.0%	100.0%
	Rec	100.0%	95.0%	87.32%	96.0%	100.0%	100.0%
	Pre	100.0%	96.66%	95.32%	100.0%	100.0%	100.0%
	Fmes	100.0%	94.66%	90.37%	96.66%	100.0%	100.0%
	Time	0.0019	0.0013	0.0016	0.1091	0.0006	0.0005
Leukemia3-C	Acc	100.0%	98.00%	95.29%	96.73%	96.33%	96.90%
	Rec	98.0%	98.0%	94.4%	93.20%	93.0%	86.0%
	Pre	99.0%	99.0%	93.60%	97.20	97.0%	96.0%
	Fmes	99.0%	99.0%	93.40%	95.20	95.0%	90.0%
	Time	0.0004	0.001	0.001	0.0660	0.0011	0.0004
Leukemia4-C	Acc	100.0%	97.14%	87.41%	89.68%	88.35%	93.21%
	Rec	99.0%	90.0%	83.2%	79.4%	66.0%	87.0%
	Pre	99.0%	98.0%	90.2%	90.0%	64.0%	97.0%
	Fmes	99.0%	93.0%	85.8%	82.80%	64.0%	91.0%
	Time	0.0013	0.0015	0.0009	0.085	0.0006	0.0005
Breast	Acc	100.0%	95.0%	72.61%	84.09%	69.36%	85.95%
	Rec	100.0%	90.0%	60.16%	71.15%	35.0%	75.83%
	Pre	100.0%	97.5%	53.76%	87.83%	61.66%	92.66%
	Fmes	100.0%	91.57%	54.26	72.98%	43.33%	80.40%
	Time	0.002	0.001	0.001	0.070	0.001	0.0002
Lung	Acc	100.0%	92.54%	90.98%	91.72%	92.12%	92.56%
	Rec	97.0%	80.0%	78.20%	89.6%	75.0%	87.0%
	Pre	97.0%	91.0%	80.8%	93.0%	69.0%	95.0%
	Fmes	97.0%	84.0%	77.4%	91.0%	72.0%	90.0%
	Time	0.0021	0.0063	0.0030	0.1100	0.0009	0.0007
Ovarian	Acc	100.0%	100.0%	98.10%	99.04%	100.0%	100.0%
	Rec	100.0%	100.0%	97.21%	99.25%	100.0%	100.0%
	Pre	100.0%	100.0%	98.43%	99.55%	100.0%	100.0%
	Fmes	100.0%	100.0%	98.17%	98.92%	100.0%	100.0%
	Time	0.0014	0.0029	0.0016	0.0640	0.0014	0.0019

Lymphoma	Acc	100.0%	96.66%	96.33%	100.0%	98.33%	100.0%
	Rec	100.0%	87.0%	100.0%	98.8%	93.0%	100.0%
	Pre	100.0%	94.0%	100.0%	99.4%	99.0%	100.0%
	Fmes	100.0%	88.0%	100.0%	99.6%	96.0%	100.0%
	Time	0.0012	0.0010	0.0013	0.0873	0.0009	0.0008
MLL	Acc	100.0%	90.97%	93.95%	97.57%	96.90%	93.47%
	Rec	100.0%	87.0%	97.0%	97.2%	97.0%	91.0%
	Pre	100.0%	86.0%	97.0%	98.0%	96.0%	91.0%
	Fmes	100.0%	87.0%	97.0%	97.8%	96.0%	91.0%
	Time	0.0009	0.0008	0.0008	0.0730	0.0011	0.0001

Table 4: Comparison of the proposed method with other hybrid methods.

	Colon		CNS		Leukemia		Leukemia3-C		Leukemia4-C	
	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS
[30]	96.47%	17.1	90.34%	14.7	100.0%	4.2	99.44	15.8	99.44	15.8
[34]	98.80%	8	-	-	-	-	-	-	-	-
[23]	92.0%	30	84.0%	30	-	-	-	-	-	-
[21]	100.%	6.13	-	-	100.0%	7.7	-	-	-	-
[24]	98.87%	16	-	-	98.84%	12.9	-	-	-	-
[4]	96.67%	10	-	-	-	-	-	-	-	-
[19]	91.68%	78	-	-	98.35%	37	-	-	-	-
Proposed	98.00%	8	100.0%	11	100.0%	6	100.0%	7	100.0%	7
	Breast		Lung cance		Ovarian		Lymphoma		MLL	
	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS	ACC	NB_FS
[30]	-	-	96.12%	23.8	100.0%	3.4	-	-	99.72%	11.1
[34]	-	-	99.3%	12	-	-	99.9%	10.6	/	-
[23]	-	-	-	-	-	-	94.0%	30	-	-
[21]	-	-	99.35%	9.1	-	-	-	-	-	-
[24]	94.15%	16.8	99.52%	12.9	-	-	-	-	-	-
[4]	-	-	-	-	100.0%	10	-	-	-	-
[19]	-	-	-	-	-	-	-	-	-	-
Proposed	100.0%	17	100.0%	17	100.0%	5	100.0%	2	100.0%	4

However, the number of selected features in the filter stage is determined empirically. In the future, we aim to create a mathematical function to determine the threshold based on the input dataset. Another objective is to improve the overall performance by incorporating more filter methods into the proposed method to eliminate irrelevant and redundant features.

References

[1] Muhammed Abd-Elnaby, Marco Alfonse, and Mohamed Roushdy. Classification of breast cancer using microarray gene expression data: A survey. *Journal of Biomedical Informatics*, 117:103764, 2021. URL `http://dx.doi.org/10.1016/j.jbi.2021.103764`.

[2] Heba Abusamra. A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Computer Science*, 23:5–14, 2013. URL `http://dx.doi.org/10.1016/j.procs.2013.10.003`.

[3] Russul Alanni, Jingyu Hou, Hasseeb Azzawi, and Yong Xiang. A novel gene selection algorithm for cancer classification using microarray datasets. *BMC medical genomics*, 12(1):1–12, 2019. URL `http://dx.doi.org/10.1186/s12920-018-0447-6`.

[4] Talal Almutiri and Faisal Saeed. Chi square and support vector machine with recursive feature elimination for gene expression data classification. In *2019 First International Conference of Intelligent Computing and Engineering (ICOICE)*, pages 1–6. IEEE, 2019. URL `http://dx.doi.org/10.1109/icoice48418.2019.9035165`.

[5] Verónica Bolón-Canedo and Amparo Alonso-Betanzos. *Microarray Bioinformatics*. Springer, 2019. URL `http://dx.doi.org/10.1007/978-1-4939-9442-7`.

[6] Verónica Bolón-Canedo, Noelia Sánchez-Marono, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282:111–135, 2014. URL `http://dx.doi.org/10.1016/j.ins.2014.05.042`.

[7] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. Feature selection for high-dimensional data. *Progress in Artificial Intelligence*, 5:65–75, 2016. URL `http://dx.doi.org/10.1007/s13748-015-0080-y`.

[8] Sara Haddou Bouazza, Khalid Auhmani, Abdelouhab Zeroual, and Nezha Hamdi. Selecting significant marker genes from microarray data by filter approach for cancer diagnosis. *Procedia Computer Science*, 127:300–309, 2018. URL `http://dx.doi.org/10.1016/j.procs.2018.01.126`.

[9] Zhipeng Cai, Randy Goebel, Mohammad R Salavatipour, and Guohui Lin. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. *BMC bioinformatics*, 8(1):1–15, 2007. URL `http://dx.doi.org/10.1186/1471-2105-8-206`.

[10] Hakan Gunduz. An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on parkinson's disease classification. *Biomedical Signal Processing and Control*, 66:102452, 2021. URL `http://dx.doi.org/10.1016/j.bspc.2021.102452`.

[11] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002. URL `https://link.springer.com/article/10.1023/A:1012487302797`.

[12] Hind Hamla and Khadoudja Ghanem. Comparative study of embedded feature selection methods on microarray data. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 69–77. Springer, 2021. URL `http://dx.doi.org/10.1007/978-3-030-79150-6_6`.

[13] Hind Hamla and Khadoudja Ghanem. A comparative study of filter feature selection methods on microarray data. In *International Conference on Computing and Information Technology*, pages 186–201. Springer, 2022. URL `http://dx.doi.org/10.1007/978-3-031-25344-7_18`.

[14] Xiaojuan Huang, Li Zhang, Bangjun Wang, Fanzhang Li, and Zhao Zhang. Feature clustering based support vector machine recursive feature elimination for gene selection. *Applied Intelligence*, 48(3):594–607, 2018. URL `http://dx.doi.org/10.1007/s10489-017-0992-2`.

[15] Hengxun Li, Wei Guo, Guoying Wu, and Yanxia Li. A rf-pso based hybrid feature selection model in intrusion detection system. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 795–802. IEEE, 2018. URL `http://dx.doi.org/10.1109/dsc.2018.00128`.

[16] Zifa Li, Weibo Xie, and Tao Liu. Efficient feature selection and classification for microarray data. *PloS one*, 13(8):e0202167, 2018. URL `http://dx.doi.org/10.1371/journal.pone.0202167`.

[17] Huijuan Lu, Junying Chen, Ke Yan, Qun Jin, Yu Xue, and Zhigang Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, 256:56–62, 2017. URL `http://dx.doi.org/10.1016/j.neucom.2016.07.080`.

[18] Shruti Mishra and Debahuti Mishra. Svm-bt-rfe: An improved gene selection framework using bayesian t-test embedded in support vector machine (recursive feature elimination) algorithm. *Karbala International Journal of Modern Science*, 1(2):86–96, 2015. URL `http://dx.doi.org/10.1016/j.kijoms.2015.10.002`.

[19] Piyushkumar A Mundra and Jagath C Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE transactions on nanobioscience*, 9(1):31–37, 2009. URL `http://dx.doi.org/10.1109/tnb.2009.2035284`.

[20] Akshata Naik, Venkatanareshbabu Kuppili, and Damodar Reddy Edla. Binary dragonfly algorithm and fisher score based hybrid feature selection adopting a novel fitness function applied to microarray data. In *2019 International Conference on Applied Machine Learning (ICAML)*, pages 40–43. IEEE, 2019. URL `http://dx.doi.org/10.1109/icaml48257.2019.00015`.

[21] Salima Ouadfel and Mohamed Abd Elaziz. Efficient high-dimension feature selection based on enhanced equilibrium optimizer. *Expert Systems with Applications*, 187:115882, 2022. URL `http://dx.doi.org/10.1016/j.eswa.2021.115882`.

[22] Beatriz Remeseiro and Veronica Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in biology and medicine*, 112:103375, 2019. URL `http://dx.doi.org/10.1016/j.compbiomed.2019.103375`.

[23] Zohre Sadeghian, Ebrahim Akbari, and Hossein Nematzadeh. A hybrid feature selection method based on information theory and binary butterfly optimization algorithm. *Engineering Applications of Artificial Intelligence*, 97:104079, 2021. URL `http://dx.doi.org/10.1016/j.engappai.2020.104079`.

[24] Alok Kumar Shukla. Multi-population adaptive genetic algorithm for selection of microarray biomarkers. *Neural Computing and Applications*, 32(15):11897–11918, 2020. URL `http://dx.doi.org/10.1007/s00521-019-04671-2`.

[25] Alok Kumar Shukla, Pradeep Singh, and Manu Vardhan. A hybrid gene selection method for microarray recognition. *Biocybernetics and Biomedical Engineering*, 38(4):975–991, 2018. URL `http://dx.doi.org/10.1016/j.bbe.2018.08.004`.

[26] Alok Kumar Shukla, Diwakar Tripathi, B Ramachandra Reddy, and D Chandramohan. A study on metaheuristics approaches for gene selection in microarray data: algorithms, applications and open challenges. *Evolutionary Intelligence*, 13(3):309–329, 2020. URL `http://dx.doi.org/10.1007/s12065-019-00306-6`.

[27] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974. URL `http://dx.doi.org/10.1111/j.2517-6161.1974.tb00994.x`.

[28] Lin Sun, Xiao-Yu Zhang, Yu-Hua Qian, Jiu-Cheng Xu, Shi-Guang Zhang, and Yun Tian. Joint neighborhood entropy-based gene selection method with fisher score for tumor classification. *Applied Intelligence*, 49(4):1245–1259, 2019. URL `http://dx.doi.org/10.1007/s10489-018-1320-1`.

[29] J Yang, YL Liu, CS Feng, and GQ Zhu. Applying the fisher score to identify alzheimer's disease-related genes. *Genet Mol Res*, 15(2), 2016. URL `http://dx.doi.org/10.4238/gmr.15028798`.

[30] Ge Zhang, Jincui Hou, Jianlin Wang, Chaokun Yan, and Junwei Luo. Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdisciplinary Sciences: Computational Life Sciences*, 12:288–301, 2020. URL `http://dx.doi.org/10.1007/s12539-020-00372-w`.

[31] Huaqing Zhang, Jian Wang, Zhanquan Sun, Jacek M Zurada, and Nikhil R Pal. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):659–673, 2019. URL `http://dx.doi.org/10.1109/tkde.2019.2893266`.

[32] Xue Zhang, Zhiguo Shi, Xuan Liu, and Xueni Li. A hybrid feature selection algorithm for classification unbalanced data processsing. In *2018 IEEE International Conference on Smart Internet of Things (SmartIoT)*, pages 269–275. IEEE_, 2018. URL `http://dx.doi.org/10.1109/smartiot.2018.00055`.

[33] Ying Zhang, Qingchun Deng, Wenbin Liang, and Xianchun Zou. An efficient feature selection strategy based on multiple support vector machine technology with gene expression data. *BioMed research international*, 2018, 2018. URL `http://dx.doi.org/10.1155/2018/7538204`.

[34] Yuefeng Zheng, Ying Li, Gang Wang, Yupeng Chen, Qian Xu, Jiahao Fan, and Xueting Cui. Retracted: A hybrid feature selection algorithm for microarray data. *Concurrency and Computation: Practice and Experience*, 31(12):e4716, 2019. URL `http://dx.doi.org/10.1002/cpe.4716`.

[35] Zexuan Zhu, Yew-Soon Ong, and Manoranjan Dash. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, 40(11):3236–3248, 2007. URL `http://dx.doi.org/10.1016/j.patcog.2007.02.007`.