Hybrid Feature Selection and Ensemble Learning Methods for Gene Selection and Cancer Classification

Sultan Noman Qasem[1], Faisal Saeed[2]
Computer Science Department, College of Computer and Information Sciences[1]
Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia[1]
Computer Science Department, Faculty of Applied Science, Taiz University, Taiz, Yemen[1]
Information Systems Department, College of Computer Science and Engineering, Taibah University, Medina, Saudi Arabia[2]

***Abstract*—A promising research field in bioinformatics and data mining is the classification of cancer based on gene expression results. Efficient sample classification is not supported by all genes. Thus, to identify the appropriate genes that help efficiently distinguish samples, a robust feature selection method is needed. Redundancy in the data on gene expression contributes to low classification performance. This paper presents the combination for gene selection and classification methods using ranking and wrapper methods. In ranking methods, information gain was used to reduce the size of dimensionality to 1% and 5%. Then, in wrapper methods K-nearest neighbors and Naïve Bayes were used with Best First, Greedy Stepwise, and Rank Search. Several combinations were investigated because it is known that no single model can give the best results using different datasets for all circumstances. Therefore, combining multiple feature selection methods and applying different classification models could provide a better decision on the final predicted cancer types. Compared with the existing classifiers, the proposed assembly gene selection methods obtained comparable performance.**

Keywords*—Microarray; gene selection; ensemble classification; cancer classification; gene expression

I. Introduction

Gene expression is called the process of transcription of the Deoxyribo Nucleic Acid (DNA) sequence into Ribo Nucleic Acid (RNA). The expression frequency of a gene shows the average number of copies of the cell-produced RNA in that gene and is associated with the corresponding volume of protein [1].

Microarray is the technique for simultaneous measurements of the expression level in a single chip of tens of thousands of genes. Microarrays therefore provide an effective way to collect data that can be used to establish the pattern of expression of thousands of genes. In most classification issues, high gene expression data is a major challenge. Therefore, not all genes also lead to cancer. A broad variety of genes have no clinical importance or insignificance. However, incorrect diagnosis can also be accomplished by using both genes in the Microarray classification of gene expression. The two key explanations for low classification precision are two: large number of features (genes) against limited sample size and dimensional consistency in articulated data [2]. Subsequently, the decrease in dimensions is necessary. Standard machine learning methods have not been effective, since these methods are better suited when there are more samples than features.

In order to solve these problems, selection algorithms for dimension reduction or features (gene) were used. The gene selection methods are usually divided into three groups, namely filter, wrapper and embedded methods. The filter procedure requires the individual evaluation of each feature using its statistical characteristics in general. The wrapper approach uses training strategies to choose the best subset of features. By the precision of the particular classifier the efficiency of the wrapper technique is calculated. In the wrapper method evolutionary or bio-inspired algorithms are also used to direct the search process. The embedded approach aims for the best feature subset and is implemented in the classification scheme. The general structure for feature selection was recently complemented with hybrid and ensemble approaches. The filter and the wrapper approaches are designed to take advantage of hybrid. Extensive works have investigated this issue and proposed several methods such as [3-16].

Several feature selection methods have been applied. For instance, the authors in [17-19] proposed hybrid methods to combine filter and wrapper algorithms to overcome the disadvantage of each individual one. Conventional optimization algorithms are not efficiently working in the feature selection of large scale problems [20]. Alternatively, different meta-heuristic algorithms have been adapted for feature selection issues. Examples of these algorithms are Genetic Algorithm (GA) [21], Ant Colony Optimization [22], Simulated Annealing [23], and Particle Swarm Optimization (PSO) [24, 25]. In addition, a modified support vector machine (SVM) was also suggested to select the minimum possible genes [26]. Multi-objective version of bat algorithm for binary feature selection [27] and Genetic Bee Colony (GBC) algorithm [28] were successfully utilized in high dimensional datasets. Moreover, a hybrid feature selection algorithm was proposed that combines the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA) [19]. The reduced gene expression dataset presented higher classification accuracy compared with conventional feature selection algorithms.

In addition, a binary version of Black Hole Algorithm called BBHA was proposed for solving feature selection problem in biological data. However, the tested classifiers were under tree family, and other kinds of classifiers were not assessed [29]. Along this line, the assessment of different classifiers such as artificial neural network (ANN) [30] and

fuzzy decision tree algorithm [31] has been made upon microarray data. In addition, the two evolutionary algorithms of PSO and GA are usually used in wrapper form [17, 20]. PSO is known to be a memory enabled algorithm compared with other algorithms, it requires few parameters to be adjusted, so it is simple and efficient [18, 32]. Kar et al. [33] proposed a PSO–adaptive K-nearest neighbors (KNN) based gene selection method and they used a heuristic for selecting the optimal values of K, while the classification accuracies have been tested using SVM algorithm. Furthermore, Jain et al. reported a two phase hybrid model for cancer classification, integrating Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO) using Naive-Bayes as the only classifier [34].

Moreover, Almutiri and Saeed [35], proposed a new combination for gene selection that utilized Chi Square and SVM Recursive Feature Elimination. This proposed method was called ChiSVMRFE and considered as ranking method. The top 10% of the genes were selected based on the high obtained weights and then SVM-RFE was used to remove the genes with lower weights. Only 10 features were selected and fed to several machine learning methods such as random forest, decision tree, K-nearest neighbors Naïve Bayes, and neural networks to enhance the cancer classification process.

The objectives of this paper are to propose a hybrid feature selection methods using the combination of filter and wrapper methods and apply them with different machine learning and ensemble learning methods to improve the performance of cancer classification.

The rest of the paper is structured as follows: Materials and Methods are provided in Section II. The experimental design is presented in Section III. Section IV shows the results and discussion. The conclusion and future work are presented in Section V.

II. Materials and Methods

A. Datasets

The proposed methods have been applied on four high dimensional microarray datasets for gene expression of different types of cancers. In addition to Breast Cancer and Brain Cancer dataset, Lung Cancer, Leukemia Cancer, Central Nervous System Cancer (CNS) datasets as shown in Table I. In the previous studies, other datasets have been used such as SRBCT, Prostate, Ovarian, MLL, Lymphoma, Leukemia and Colon, but the dimensionality of the genes for these methods is not too high and the applied feature selection and machine learning methods on these datasets obtained satisfactory performance.

TABLE I. Description of Datasets

Dataset	# Features	# Instances	# Classes
Brain	5597	42	5(10,10,10,4,8)
Breast	24481	97	2(46,51)
Lung	12600	203	5(139,17,6,21,20)
CNS	7129	60	2(21,39)

The Brain cancer [36] dataset includes 42 samples, with 5597 genes and five classes. The Breast dataset [37] includes 97 samples; with 24,481 genes. From these samples, 46 were classified as cancer. The Lung dataset [38] includes 203 samples with five classes. The number of features are 12,600 genes. Finally, the CNS dataset includes 60 samples, among these samples, only 21 are classified as cancer. The number of features are 7129 genes.

B. Hybrid Feature Selection Methods

In this study, several combinations between Filter-based and Wrapper-based feature selection methods have been done to suggest the better hybrid method. In Filter-based method, the information gain was used to reduce the dimensionality 1% and 5%. After that several wrapper-based methods were applied to investigate on the performance of gene selections, which are Best First, Greedy Stepwise, and Rank Search. Two classification methods were used in each wrapper method, which are: K-nearest neighbors and Naïve Bays. Fig. 1 shows the overall methods used in this study.

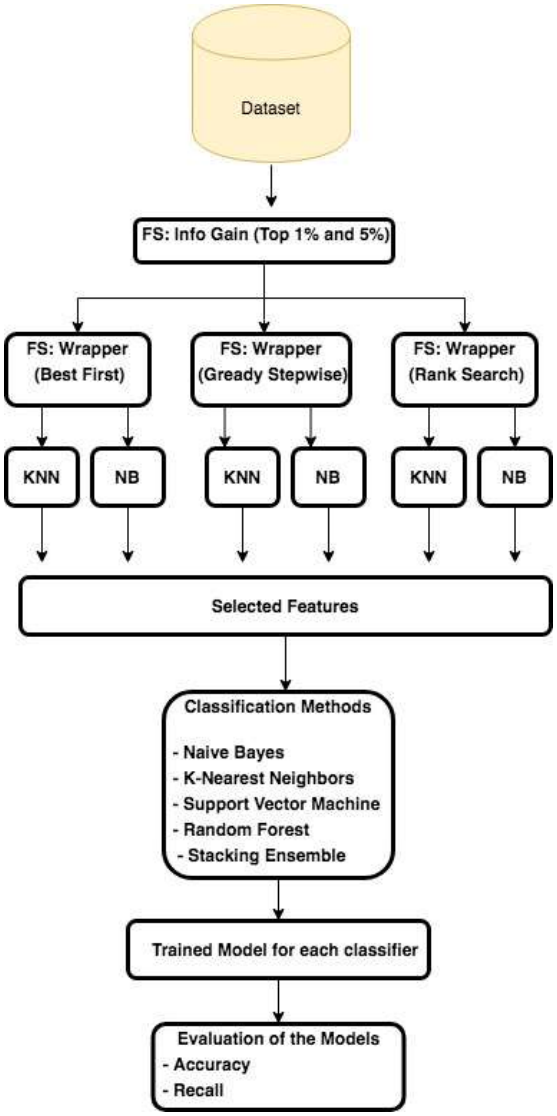

Fig. 1. The Developed Methods.

C. Machine Learning Methods

Several machine learning methods were applied for each combination in the feature selection step. These methods include individual and ensemble classification methods such as K-nearest neighbors, Naïve Bays, Support Vector Machine, Random Forests and Stacking Ensemble methods. The performance of these methods was evaluated before and after using the different combinations of feature selection and the best preforming methods were reported, as shown in Fig. 1.

III. Experimental Design

The experiments have been conducted on WEKA tool version 3.8. Each outcome of feature selection method has been fed to all machine learning methods (KNN, NB, SVM, RM and Stacking) in order to evaluate the performance of the gene selection and the cancer classification methods.

10-folds cross validation has been used for training and testing each dataset for all obtained combinations. The performance was evaluated using Accuracy and Recall measures, which are defined in the following equations (1) and (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where TP is true positive; TN is true negative; FP is false positive, and FN is false negative.

In addition, the performance of each method was compared before and after using features selection methods in order to discuss the enhancements obtained.

IV. Results and Discussion

The performance of the different combinations of feature selection and machine learning methods is shown in the tables below. The best performing method for each combination is bolded and the best performing method among all combinations for each dataset is shaded.

For Breast Cancer dataset, the performance of the used methods (using top 1% and 5% in the ranking method: information gain) are presented in Tables II and III.

As shown in Table II, the random forest method obtained the best accuracy and recall values with high dimensionality case (all features: 24481 and top 1% features: 244). However, after applying different combinations using ranking and wrapper methods, we found that Information Gain & Wrapper (NB & Best First) and Information Gain & Wrapper (NB & Gready Stepwise) obtained the best performance compared to all other methods/combinations before and after applying feature selection. Similarly, when the top 5% genes were selected in the ranking method, the performance of the used methods in Table III showed that random forest obtained the best results when high dimensional dataset was used, but when wrapper methods were applied, the combination of Information Gain and Wrapper (NB & Best First) obtained the best results. For Brain Cancer dataset, the results of used methods using the top 1% and 5% features are shown in Tables IV and V.

TABLE II. The Feature Selection and Machine Learning Methods for Breast Cancer Dataset using the Top 1% Features

	No. of Features	**Measure**	**NB**		**SVM**	**RF**	**Stacking**
All Features:	24481	Accuracy	0.546	0.608	0.546	**0.659**	0.526
		Recall	0.546	0.610	0.546	**0.660**	0.526
Information Gain 1 %:	244	Accuracy	0.608	0.804	0.722	**0.845**	0.814
		Recall	0.608	0.804	0.722	**0.845**	0.814
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.577	**0.876**	0.629	0.814	0.763
		Recall	0.577	**0.876**	0.629	0.814	0.763
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.938**	0.804	0.784	0.856	0.835
		Recall	**0.938**	0.804	0.784	0.856	0.835
Info Gain & Wrapper (KNN & GreadyStepwise)	5	Accuracy	0.557	**0.845**	0.639	0.825	0.763
		Recall	0.557	**0.845**	0.639	0.825	0.763
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.938**	0.804	0.784	0.856	0.835
		Recall	**0.938**	0.804	0.784	0.856	0.835
Info Gain & Wrapper (KNN & RankSearch)	104	Accuracy	0.577	**0.866**	0.732	0.835	0.794
		Recall	0.577	**0.866**	0.732	0.835	0.794
Info Gain & Wrapper (NB & RankSearch)	2	Accuracy	**0.742**	0.660	0.711	0.732	0.670
		Recall	**0.742**	0.660	0.711	0.732	0.670

TABLE III. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BREAST CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	24481	Accuracy	0.546	0.608	0.546	**0.659**	0.526
		Recall	0.546	0.610	0.546	**0.660**	0.526
Information Gain 5 %:	1224	Accuracy	0.577	0.773	0.670	**0.814**	0.722
		Recall	0.577	0.773	0.670	**0.814**	0.722
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.557	**0.907**	0.557	0.845	0.845
		Recall	0.557	**0.907**	0.557	0.845	0.845
Info Gain & Wrapper (NB & Best First)	15	Accuracy	**0.969**	0.784	0.825	0.866	0.928
		Recall	**0.969**	0.784	0.825	0.866	0.928
Info Gain & Wrapper (KNN & GreadyStepwise)	6	Accuracy	0.577	**0.897**	0.588	0.825	0.814
		Recall	0.577	**0.897**	0.588	0.825	0.814
Info Gain & Wrapper (NB & GreadyStepwise)	12	Accuracy	**0.949**	0.825	0.753	0.876	0.876
		Recall	**0.948**	0.825	0.753	0.876	0.876
Info Gain & Wrapper (KNN & RankSearch)	669	Accuracy	0.577	0.845	0.691	**0.866**	0.856
		Recall	0.577	0.845	0.691	**0.866**	0.856
Info Gain & Wrapper (NB & RankSearch)	2	Accuracy	**0.742**	0.660	0.711	0.732	0.670
		Recall	**0.742**	0.660	0.711	0.732	0.670

As shown in Tables IV and V, it is clearly shown that there are high improvements when using the combined feature selection methods. The best reported method is KNN as classification method and Information Gain & Wrapper (KNN & Best First) as feature selection methods using the top 1% and 5% features. In addition, for the top 5% features, other combinations obtained the same best results which are KNN classifier with Information Gain & Wrapper (KNN & GreadyStepwise), NB classifier with Information Gain & Wrapper (NB & Best First) and NB classifier with Information Gain & Wrapper (NB & GreadyStepwise) feature selection methods.

For Lung Cancer Dataset, the best performing method is NB classifier with Information Gain & Wrapper (NB & Best First) feature selection method for the top 1 % features (as shown in Table VI), and KNN with Info Gain & Wrapper (KNN & Best First) and Info Gain & Wrapper (KNN & GreadyStepwise) for the top 5% features (see Table VII).

TABLE IV. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BRAIN CANCER DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	5597	Accuracy	0.714	0.762	0.691	0.786	**0.881**
		Recall	0.714	0.762	0.690	0.786	**0.881**
Information Gain 1 %:	56	Accuracy	0.810	0.881	0.833	**0.905**	0.833
		Recall	0.810	0.881	0.833	**0.905**	0.833
Info Gain & Wrapper (KNN & Best First)	9	Accuracy	0.904	**0.100**	0.810	0.880	0.905
		Recall	0.905	**0.100**	0.810	0.881	0.905
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.976**	0.881	0.786	0.952	0.857
		Recall	**0.976**	0.881	0.786	0.952	0.857
Info Gain & Wrapper (KNN & GreadyStepwise)	6	Accuracy	0.762	**0.952**	0.833	0.881	0.643
		Recall	0.762	**0.952**	0.833	0.881	0.643
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.976**	0.881	0.786	0.952	0.857
		Recall	**0.976**	0.881	0.786	0.952	0.857
Info Gain & Wrapper (KNN & RankSearch)	26	Accuracy	0.857	0.905	0.881	0.905	**0.929**
		Recall	0.857	0.905	0.881	0.905	**0.929**
Info Gain & Wrapper (NB & RankSearch)	9	Accuracy	0.881	0.857	0.857	0.881	**0.929**
		Recall	0.881	0.857	0.857	0.881	**0.929**

TABLE V. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR BRAIN CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	5597	Accuracy	0.714	0.762	0.691	0.786	**0.881**
		Recall	0.714	0.762	0.690	0.786	**0.881**
Information Gain 5 %:	280	Accuracy	0.810	0.857	**0.905**	0.881	0.810
		Recall	0.810	0.857	**0.905**	0.881	0.810
Info Gain & Wrapper (KNN & Best First)	11	Accuracy	0.905	**1.000**	0.810	0.881	0.810
		Recall	0.905	**1.000**	0.810	0.881	0.810
Info Gain & Wrapper (NB & Best First)	8	Accuracy	**1.000**	0.952	0.905	1.000	0.952
		Recall	**1.000**	0.952	0.905	1.000	0.952
Info Gain & Wrapper (KNN & GreadyStepwise)	9	Accuracy	0.786	**1.000**	0.810	0.833	0.810
		Recall	0.786	**1.000**	0.810	0.833	0.810
Info Gain & Wrapper (NB & GreadyStepwise)	8	Accuracy	**1.000**	0.952	0.905	**1.000**	0.952
		Recall	**1.000**	0.952	0.905	**1.000**	0.952
Info Gain & Wrapper (KNN & RankSearch)	191	Accuracy	0.833	0.905	0.929	**0.976**	0.905
		Recall	0.833	0.905	0.929	**0.976**	0.905
Info Gain & Wrapper (NB & RankSearch)	8	Accuracy	**0.929**	0.762	0.738	0.905	0.786
		Recall	**0.929**	0.762	0.738	0.905	0.786

Finally, the performance of the combined methods for CNS Dataset is presented in Tables VIII and IX. The results show that the best performing method is KNN classifier with Information Gain & Wrapper (KNN & Best First) and NB with Information Gain & Wrapper (NB & Best First) feature selection methods for the top 1 % features (as shown in Table VIII). In addition, the RF with the combination of Info Gain & Wrapper (KNN & RankSearch) obtained the same best results here. For the top 5% features, and KNN with Info Gain & Wrapper (KNN & Best First) consistently obtained the best results in this case as well.

By comparing the performances of all combined feature selection methods with different individual and ensemble machine learning methods, it is clearly shown that using these combinations with high dimensional datasets improved the cancer classification using all datasets used. The results in Tables II to IX showed that the best performing methods were KNN classifier with Information Gain & Wrapper (KNN & Best First) feature selection method and NB classifier with Info Gain & Wrapper (NB & Best First) feature selection method. Each one obtained the best five from eight cases using different datasets and different thresholds in the ranking methods (top 1% and 5% of features).

TABLE VI. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR LUNG CANCER DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	12600	Accuracy	0.808	**0.897**	0.685	0.882	0.872
		Recall	0.808	**0.897**	0.685	0.882	0.872
Information Gain 1 %:	126	Accuracy	0.951	**0.956**	0.685	0.941	0.916
		Recall	0.951	**0.956**	0.685	0.941	0.916
Info Gain & Wrapper (KNN & Best First)	10	Accuracy	0.867	**0.970**	0.685	0.921	0.897
		Recall	0.867	**0.970**	0.685	0.921	0.897
Info Gain & Wrapper (NB & Best First)	15	Accuracy	**0.990**	0.902	0.685	0.951	0.946
		Recall	**0.990**	0.901	0.685	0.951	0.946
Info Gain & Wrapper (KNN & GreadyStepwise)	8	Accuracy	0.906	**0.966**	0.685	0.926	0.897
		Recall	0.906	**0.966**	0.685	0.926	0.897
Info Gain & Wrapper (NB & GreadyStepwise)	13	Accuracy	**0.985**	0.916	0.685	0.931	0.926
		Recall	**0.985**	0.916	0.685	0.931	0.926
Info Gain & Wrapper (KNN & RankSearch)	119	Accuracy	0.941	**0.966**	0.685	0.946	**0.966**
		Recall	0.941	**0.966**	0.685	0.946	**0.966**
Info Gain & Wrapper (NB & RankSearch)	126	Accuracy	0.951	**0.956**	0.685	0.941	0.916
		Recall	0.951	**0.956**	0.685	0.941	0.916

TABLE VII. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR LUNG CANCER DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	12600	Accuracy	0.808	**0.897**	0.685	0.882	0.872
		Recall	0.808	**0.897**	0.685	0.882	0.872
Information Gain 5 %:	630	Accuracy	0.941	**0.956**	0.685	0.941	**0.956**
		Recall	0.941	**0.956**	0.685	0.941	**0.956**
Info Gain & Wrapper (KNN & Best First)	11	Accuracy	0.892	**0.990**	0.685	0.936	0.926
		Recall	0.892	**0.990**	0.685	0.936	0.926
Info Gain & Wrapper (NB & Best First)	12	Accuracy	**0.985**	0.931	0.685	0.936	0.970
		Recall	**0.985**	0.931	0.685	0.936	0.970
Info Gain & Wrapper (KNN & GreadyStepwise)	11	Accuracy	0.892	**0.990**	0.685	0.936	0.921
		Recall	0.892	**0.990**	0.685	0.936	0.926
Info Gain & Wrapper (NB & GreadyStepwise)	12	Accuracy	**0.985**	0.931	0.685	0.936	0.970
		Recall	**0.985**	0.931	0.685	0.936	0.970
Info Gain & Wrapper (KNN & RankSearch)	213	Accuracy	0.936	**0.970**	0.685	0.936	0.961
		Recall	0.936	**0.970**	0.685	0.936	0.961
Info Gain & Wrapper (NB & RankSearch)	232	Accuracy	0.946	**0.961**	0.685	0.936	0.941
		Recall	0.946	**0.961**	0.685	0.936	0.941

TABLE VIII. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR CNS DATASET USING THE TOP 1% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	7129	Accuracy	0.617	0.567	0.650	**0.667**	0.550
		Recall	0.617	0.567	0.650	**0.667**	0.550
Information Gain 1 %:	71	Accuracy	0.717	0.817	0.650	**0.833**	0.767
		Recall	0.717	0.817	0.650	**0.833**	0.767
Info Gain & Wrapper (KNN & Best First)	7	Accuracy	0.733	**0.900**	0.650	0.833	0.867
		Recall	0.733	**0.900**	0.650	0.833	0.867
Info Gain & Wrapper (NB & Best First)	12	Accuracy	**0.900**	0.783	0.650	0.883	0.833
		Recall	**0.900**	0.783	0.650	0.883	0.833
Info Gain & Wrapper (KNN & GreadyStepwise)	3	Accuracy	0.600	**0.883**	0.650	0.750	0.767
		Recall	0.600	**0.883**	0.650	0.750	0.767
Info Gain & Wrapper (NB & GreadyStepwise)	6	Accuracy	**0.850**	0.583	0.650	0.800	0.700
		Recall	**0.850**	0.583	0.650	0.800	0.700
Info Gain & Wrapper (KNN & RankSearch)	40	Accuracy	0.767	0.883	0.650	**0.900**	0.817
		Recall	0.767	0.883	0.650	**0.900**	0.817
Info Gain & Wrapper (NB & RankSearch)	55	Accuracy	0.750	0.850	0.650	**0.867**	0.767
		Recall	0.750	0.850	0.650	**0.867**	0.767

TABLE IX. THE FEATURE SELECTION AND MACHINE LEARNING METHODS FOR CNS DATASET USING THE TOP 5% FEATURES

	No. of Features	Measure	NB	KNN	SVM	RF	Stacking
All Features:	7129	Accuracy	0.617	0.567	0.650	**0.667**	0.550
		Recall	0.617	0.567	0.650	**0.667**	0.550
Information Gain 5 %:		Accuracy	0.667	0.617	0.650	**0.783**	0.733
		Recall	0.667	0.617	0.650	**0.783**	0.733
Info Gain & Wrapper (KNN & Best First)	13	Accuracy	0.583	**0.967**	0.650	0.767	0.800
		Recall	0.583	**0.967**	0.650	0.767	0.800
Info Gain & Wrapper (NB & Best First)	11	Accuracy	**0.883**	0.800	0.650	0.800	0.833
		Recall	**0.883**	0.800	0.650	0.800	0.833
Info Gain & Wrapper (KNN & GreadyStepwise)	2	Accuracy	0.467	**0.800**	0.650	0.717	0.700
		Recall	0.467	**0.800**	0.650	0.717	0.700
Info Gain & Wrapper (NB & GreadyStepwise)	11	Accuracy	**0.883**	0.800	0.650	0.800	0.833
		Recall	**0.883**	0.800	0.650	0.800	0.833
Info Gain & Wrapper (KNN & RankSearch)	37	Accuracy	0.750	0.850	0.650	**0.883**	0.750
		Recall	0.750	0.850	0.650	**0.883**	0.750
Info Gain & Wrapper (NB & RankSearch)	55	Accuracy	0.750	0.850	0.650	**0.867**	0.833
		Recall	0.750	0.850	0.650	**0.867**	0.833

V. CONCLUSION AND FUTURE WORK

The investigation of high dimensionality issue in microarray datasets has been conducted in this paper. Several combinations of ranking methods (using information gain with threshold of 1% and 5%) and wrapper methods (using KNN and NB with Best First, Greedy Stepwise, and Rank Search) were used to select the most important genes for microarray datasets. These datasets included Breast Cancer, Brain Cancer, Lung Cancer and CNS datasets. The experimental results showed the consistent good performance of applying all feature selection methods comparing with the case when all features were used (no feature selection methods). Among these used methods, the KNN with Information Gain & Wrapper (KNN & Best First) and NB with Info Gain & Wrapper (NB & Best First) obtained the best performance and overcame all other methods. Therefore, this study recommends to use one of these methods on high dimensionally microarray methods with the aim of obtaining better cancer classification accuracy. Future works will investigate other hybrid and intelligent feature selection methods for cancer classification using microarray datasets.

ACKNOWLEDGMENT

The authors would like to thank Deanship of Scientific Research at Al Imam Mohammad ibn Saud Islamic university, Saudi Arabia, for financing this project under the grant no. (18-11-09-015).

REFERENCES

[1] Y. Lu, and J. Han, "Cancer Classification Using Gene Expression Data," Information Systems, vol. 28, pp. 243-268, 2003.

[2] L.Yu, and H. Liu, "Redundancy based Feature Selection for Microarray Data," in Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, pp. 737-742, 2004.

[3] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods", Information Sciences, 2014;282:111-35.

[4] G. Cosma, D. Brown, M. Archer, M. Khan, and A. G. Pockley, "A survey on computational intelligence approaches for predictive modeling in prostate cancer", Expert Systems with Applications, 70:1-19, 2017.

[5] R. K. Singh, and M. Sivabalakrishnan , "Feature selection of gene expression data for cancer classification: a review", Procedia Computer Science, 50:52-7, 2015.

[6] L.Wang, Feature selection in bioinformatics. 2012.

[7] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data", IEEE Transactions on Knowledge and Data Engineering, 25(1):1-14, 2013.

[8] P. C. Conilione, and D. Wang., "A comparative study on feature selection for E. coli promoter recognition", Intertional Journal of Information Technology,11(8):54-66, 2005.

[9] Z. M. Hira, and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data", Advances in bioinformatics, 2015.

[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data", Knowledge and Information Systems, 34(3):483-519, 2013.

[11] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, et al., "A survey on filter techniques for feature selection in gene expression microarray analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4):1106-19, 2012.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics", bioinformatics, 23(19):2507-17, 2007.

[13] M. A. Hall, Correlation-based feature selection for machine learning, 1999.

[14] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers", Genome Research, 11(11):1878-8.7, 2001.

[15] L. E. A. d. S. Santana, A. M. de Paula Canuto, "Filter-based optimization techniques for selection of feature subsets in ensemble systems", Expert Systems with Applications, 41(4):1622-31, 2014.

[16] P. M. Narendra, and K. Fukunaga, "A branch and bound algorithm for feature subset selection," IEEE Transactions on Computers, vol. C-26, no. 9, pp. 917-922, Sep. 1977.

[17] E. Alba, J. Garcia-Nieto, L. Jourdan, E-G Talbi, editors, "Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms", 2007 IEEE Congress on Evolutionary Computation, 2007.

[18] S. S. Hameed, R. Hassan, and F. F. Muhammad, "Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm", PLOS ONE, 12(11): 2017, e0187371. doi: 10.1371/journal.pone.0187371.

[19] L. Huijuan, C. Junying, Y. Ke, J. Qun, X. Yu, and G. Zhigang, "A hybrid feature selection algorithm for gene expression data classification", Neurocomputing, 256: 56-62, 2017.

[20] L-F Chen, C-T Su, K-H Chen, P-C Wang, "Particle swarm optimization for feature selection with application in obstructive sleep apnea diagnosis", Neural Computing and Applications, 21(8):2087-96, 2012.

[21] T. Latkowski, and S. Osowski, "Data mining for feature selection in gene expression autism data", Expert Systems with Applications, 42(2):864-72, 2015.

[22] Y. Chen, D. Miao, and R. Wang, "A rough set approach to feature selection based on ant colony optimization" , Pattern Recognition Letters, 31(3):226-33, 2010.

[23] F. González, and L. A. Belanche, "Feature selection for microarray gene expression data using simulated annealing guided by the multivariate joint entropy", arXiv preprint arXiv:13021733, 2013.

[24] F. Ardjani, K. Sadouni, and M. Benyettou, "Optimization of SVM multiclass by particle swarm (PSO-SVM)", 2nd IEEE International Workshop on Database Technology and Applications (DBTA), 2010.

[25] B. Tran, B. Xue, and M. Zhang, "Improved PSO for feature selection on high-dimensional datasets', Asia-Pacific Conference on Simulated Evolution and Learning, Springer ,2014.

[26] B. Ghaddar, and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines", European Journal of Operational Research, 265(3):993-1004, 2018.

[27] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach", Genomics, 110(1):10-7, 2018.

[28] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification", Computational Biology and Chemistry, 56:49-60, 2015.

[29] E. Pashaei, and N. Aydin, "Binary black hole algorithm for feature selection and classification on biological data", Applied Soft Computing, 56:94-106, 2017.

[30] R. Aziz, C. K. Verma, M. Jha, and N. Srivastava, "Artificial neural network classification of microarray data using new hybrid gene selection method", International Journal of Data Mining and Bioinformatics, 17(1):42-65, 2017.

[31] S. A. Ludwig, S. Picek, and D. Jakobovic, "Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm", In: Kahraman C, Topcu YI, editors. Operations Research Applications in Health Care Management, Cham: Springer International Publishing, p. 327-47, 2018.

[32] C. S. R. Annavarapu, S. Dara, and H. Banka, "Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm", EXCLI journal,15:460, 2016.

[33] S. Kar, K. D. Sharma, M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique", Expert Systems with Applications, 42(1):612-27, 2015.

[34] I. Jain, V. K. Jain, andR. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification", Applied Soft Computing, 62:203-15, 2018.

[35] T. Almutiri, and F. Saeed, "Chi Square and Support Vector Machine with Recursive Feature Elimination for Gene Expression Data Classification", In 2019 IEEE First International Conference of Intelligent Computing and Engineering (ICOICE), December. (pp. 1-6), 2019.

[36] M. Dettling, and P. Bühlmann, "Supervised clustering of genes. Genome biology", 3(12), 1-15, 2002.

[37] Y. Sun, V. Urquidi, and S. Goodison, "Derivation of molecular signatures for breast cancer recurrence prediction using a two-way validation approach", Breast cancer research and treatment, 119(3), 593-599, 2010.

[38] http://www.biolab.si/supp/bi-cancer/projections/info/lung.html, last accessed 2010/05/20.