Journal of Biomedical Informatics 117 (2021) 103764

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Classification of breast cancer using microarray gene expression data: A survey

Muhammed Abd-Elnaby [a,*], Marco Alfonse [a], Mohamed Roushdy [b]

[a] Faculty of Computers and Information Science, Ain Shams University, Cairo, Egypt
[b] Faculty of Computers and Information Technology, Future University, New Cairo, Egypt

ARTICLE INFO

Keywords:
Feature selection
Machine learning
Cancer classification
Microarray data

ABSTRACT

Cancer, in particular breast cancer, is considered one of the most common causes of death worldwide according to the world health organization. For this reason, extensive research efforts have been done in the area of accurate and early diagnosis of cancer in order to increase the likelihood of cure. Among the available tools for diagnosing cancer, microarray technology has been proven to be effective. Microarray technology analyzes the expression level of thousands of genes simultaneously. Although the huge number of features or genes in the microarray data may seem advantageous, many of these features are irrelevant or redundant resulting in the deterioration of classification accuracy. To overcome this challenge, feature selection techniques are a mandatory preprocessing step before the classification process. In the paper, the main feature selection and classification techniques introduced in the literature for cancer (particularly breast cancer) are reviewed to improve the microarray-based classification.

1. Introduction

All cells have a nucleus that contains deoxyribonucleic acid (DNA). DNA is carrying genetic information of the organism to develop, function, grow and reproduce. The coding segments of DNA are called genes, which are re sponsible for making proteins. Proteins do the essential work in every organism and they are synthesized in two steps. Firstly, DNA is transcribed into mRNA, then mRNA is translated into proteins. Genetic technologies such as DNA microarrays measure the simultaneous expression of genes, offering us a global view of the cell, which helps in differentiating between normal and diseased states. Cancer can be described as a group of diseases associated with uncontrollable cell growth that invades and metastasizes to other tissues. It is considered the second main cause of death globally, about 9.6 million in 2018, 1 out of 6 dies due to cancer. The most common types for men are; Lung, prostate, colorectal, stomach, and liver cancer while breast, colorectal, lung, cervical, and thyroid cancer are popular among women [1]. Breast cancer is a heterogeneous disease having different histological and biological properties and various treatment responses [2]. It can be traced back to genetic, epigenetic, or transcriptome changes. It appears as a lump, nipple discharge, or a change of skin texture around the nipple region. Early treatment of cancer increases the possibility of the cure and reduces the fatality rate and probability of recurrence [3]. Recurrence may happen after months or years from an initial treatment and it can be local where cancer affects the same place or can be distant where cancer returns to different areas in the body [4]. Breast Cancer is detected using traditional methods, e.g., physical detection, blood test, and X-ray scan, but they are time-consuming and subject to human errors [5]. Medical errors are considered the third-leading cause of death in the US [6]. Therefore, an effective tool for the diagnosis of breast cancer is necessary, and for this purpose microarray technology is extensively used. Gene expression data of DNA microarray represents the state of a cell at a molecular level [7]. It has a great perspective as a medical diagnosis. They either analyzed to determine whether the patient is oncological or not (two-class problems), distinguish between different types of cancer (multi-class problems) [8], predict the response to a drug based on the gene signature, or identify tumors [9] by finding groups of similarly expressed genes. They effectively analyzed by machine learning (ML). ML is an automatic and intelligent learning technique that gives machines the ability to learn without being explicitly programmed. Ml techniques are widely employed in solving many complex real-world problems and have proven to be efficient in

* Corresponding author.
E-mail addresses: muhammed.abdelnaby1704@gmail.com (M. Abd-Elnaby), marco_alfonse@cis.asu.edu.eg (M. Alfonse), mohamed.roushdy@fue.edu.eg (M. Roushdy).

https://doi.org/10.1016/j.jbi.2021.103764
Received 8 July 2020; Received in revised form 9 March 2021; Accepted 26 March 2021
Available online 6 April 2021

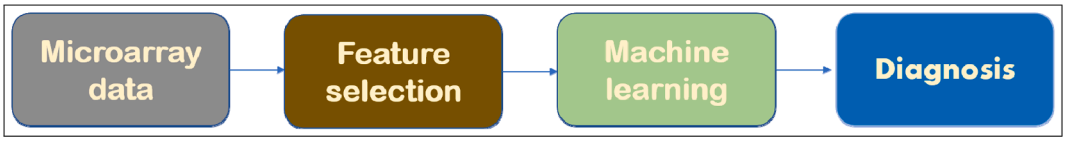

Fig. 1. The pipeline of microarray analysis.

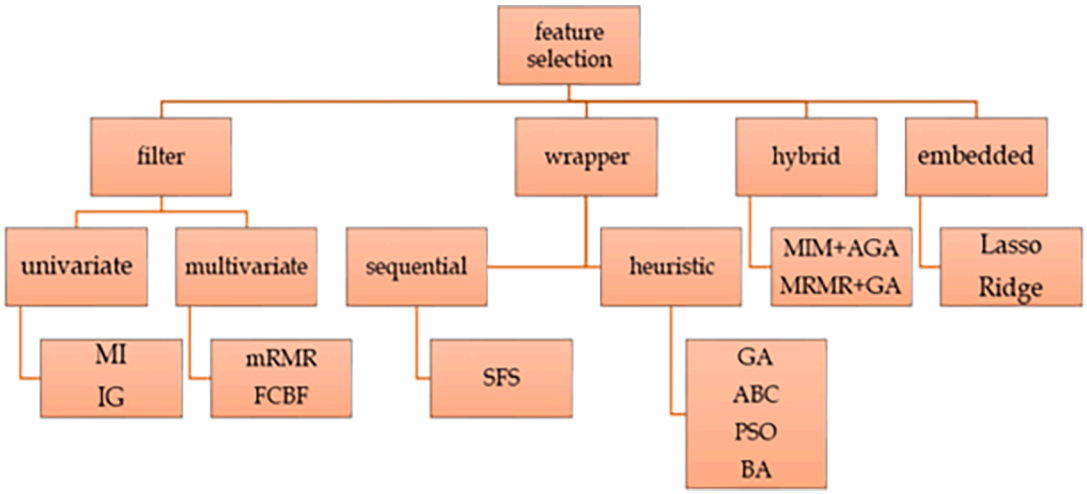

Fig. 2. The taxonomy of the feature selection techniques.

analyzing gene expression data. Applying ML improves the accuracy of predicting cancer vulnerability, mortality, and recurrence by about 15:25 [10]. While the increase in information as a result of increasing feature numbers should improve the differentiating capacity, the accuracy deteriorates when the number of features exceeds a specific limit, especially with a small number of samples, which is known as the "curse of dimensionality. This deterioration occurs as not all features are informative; many are irrelevant, redundant, or noisy. To overcome this problem feature selection techniques are utilized to select the most informative genes. Another issue with microarray data, they are unbalanced; the number of available samples in each class is not equal that makes the classification biased toward the class having the majority of samples also ranking of features is considered as a challenge [11]. To address this issue oversampling techniques can be used. The paper is organized as follows: Section 2: defines the methodology of classifying breast cancer. Section 3: presents the state of the art of breast cancer classification. Section 4 presents the discussion and finally, section 5 offers the conclusion and possible future work.

2. Methodology of cancer classification

Analysis of DNA microarray data is done through the following steps. Firstly, data are preprocessed using feature selection techniques to remove noisy, redundant features and get only informative ones. Then the resultant subset is used to train the learning model to diagnose cancer subtypes as illustrated in Fig. 1.

2.1. Feature selection

Feature selection is the process of automatically or manually select the features that have an impact on the prediction to:

- Reduce Overfitting: overfitting means the model doesn't generalize well from our training data to unseen data due to noise and redundancy in the data. The model will be well generalized when removing such data.
- Improves Accuracy: train the model with less misleading data will improve the accuracy.
- Reduce Training Time: The smaller the number of features, the less computation time required for training.
- Offer biologists with insight about the mechanism between gene signature and diseases [12].

Feature selection can be classified based on the integration between the selection algorithm and the implemented model into four main categories, as shown in Fig. 2. the pros and cons of feature selection

Table 1
The pros and cons of feature selection techniques.

	Pros	Cons
Filter	• Not dependent on any particular algorithm • Fast and are computationally simple	• Do not consider the interaction with the classifier • Low of performance
Wrapper	• It always selects a near perfect subset. • Error rate in this method is less compared to other methods.	• It has higher risk of over fitting than filter techniques. • It is computationally very intensive compared to other methods. • It is meant for the particular learning machine on which it has been tested.
Embedded	• Computationally less intensive than wrapper methods. • They include the interaction with the classification model. • They make better use of the available data by not needing to split the training data into a training and validation set. • They reach a solution faster by avoiding retraining a predictor from scratch for every variable subset investigated.	• Specific to a learning machine. • Problem of over-fitting compared to filters.
Hybrid	• combines the advantages of various approaches.	• Time complexity may increase

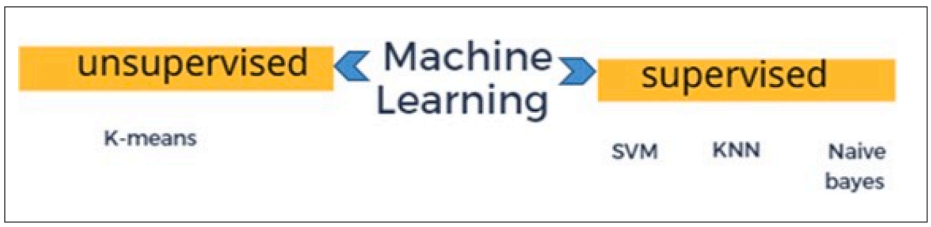

Fig. 3. The taxonomy of the machine learning techniques.

techniques are presented in table 1.

2.2. Filter approach

It evaluates the features based on the intrinsic properties of the data like distance, correlation, and consistency independently of the classifier. Although they are computationally faster and have a strong generalization ability [5] their performance is lower. Filter approaches are divided into univariate and multivariate. Univariate feature selection examines each feature individually to measure the strength of an association between the features and the outcome variable. Common types are mutual information (MI) and information gain (IG).

- MI measures the correlation between the two variables. In other words, measures how much information one variable (X) knows about another one (Y). In gene selection, it measures the correlation between gene and classification category. The larger the value of MI, the more informative the genes are [13].
- IG is a statistical property that measures how infor mative a feature is. Highly related features to class are those with the information, while unrelated features give no information. To determine the value of IG, entropy value which is the impurity of the given samples is used. Then a threshold is set and features which value higher than the threshold are selected [14].

Multivariate evaluates features in the context of others, the most typically used techniques are:

- Minimum Redundancy and Maximum Relevance (mRMR) tends to select highly correlated features with the class and lowly between themselves. It may not proper to select both features that are highly relevant and highly correlated, as they wouldn't add more information due to high correlation, but they would increase the model complexity and make it susceptible to overfitting [15].
- Correlation-based Feature Selection (CFS) ranks features based on the correlation due to the heuristic evaluation functions. Features are evaluated according to the hypothesis "Good feature subset contains features that are highly correlated with the classification and yet uncorrelated to each other" [13], which means a low correlation with the class refers to irrelevant features while informative features are strongly correlated [15].
- Fast Correlation Based Filter (FCBF) is a multivariate algorithm that bases on symmetrical uncertainty (SU) to select highly correlated features with the class. Then it applies heuristics to remove the redundant features and maintain relevant ones to the class [6].

2.2.1. Wrapper approach

Wrapper methods are using learning algorithms to select the optimal subset of features. They have better accuracy than filter methods, but they are intended for a particular learning algorithm, tend to overfit and they are very computationally intensive, as the model has to train each subset. This approach can be categorized into sequential selection algorithms and heuristic search algorithms. Sequential selection algorithms remove or add one feature at a time based on the classifier performance until a subset of the desired k features is reached that gives maximum accuracy. Common techniques are sequential forward selection and sequential backward selection. Heuristic search algorithms: the most utilized algorithms are:

- Genetic Algorithm (GA) is a heuristic search algorithm that is inspired by natural evolution. The main principle of GA is randomly generating a population through three operations. Firstly, the selection operation chooses individuals whose fitness functions are better. Then, in the crossover operation, each pair of individuals are selected with a random crossover point to generate new offspring. Finally, the Mutation process makes diversity in the population [16].
- Artificial Bee Colony (ABC) is a swarm-based algorithm that simulates how honeybees search for food. The colony of bees consists of employed bees, on– lookers, and scouts. Employed bees: its numbers are equal to the number of food sources. Each employed bee goes to a food source and evaluates it. Based on information shared by employed bees, onlookers elect the food source. An employed bee becomes a scout when the food source is depleted and begins to randomly search for a new food source around [17].
- Particle Swarm Optimization (PSO) is a swarmbased algorithm, that mimics how members in groups such as birds or fishes interact to share information. In PSO, a candidate solution is represented by a particle that has fitness values and velocity to direct the fly. Through updating the position of the particle due to its own and of other particle experience, an optimum solution can be reached [18].
- Bat Algorithm (BA) is a swarm-based algorithm, based on the mechanism bats use to situate their prey, echolocation. Bats are randomly fly based on the distance to the target. They automatically alter the frequency and rate of the emitted pulse. The solution is elected from among the best solutions [19].

2.2.2. Embedded approach

Embedded methods learn which features best contribute to the accuracy of the model while the model is created. The most common types of embedded feature selection methods are regularization methods.

2.2.3. Hybrid approach

The hybrid approach can be any combination of any number of same or different methods of feature selection to combine the advantages of both approaches and overcome or handle the drawback of each approach individually. A combination is usually a filter-wrapper approach that gets the benefit of fast computational of filter approach to remove redundant features and high performance of wrapper approach. It also less prone to overfitting than wrapper but it is classifier specific.

2.3. Machine learning

Machine Learning techniques are mainly classified into two categories as shown in Fig. 3.

2.3.1. Supervised approach

Supervised machine learning algorithms need labeled data. The commonly used classifier in microarray analysis are:

- K-nearest neighbor (KNN) is a lazy learner that builds no model. It is used for classification and regression tasks. For classification it classifies data based on the classification of its neighbors, "birds of feather flocks together", an object is classified to the major class among its k nearest neighbors. For regression, the output is the average of the values of k nearest neighbors [20,21].
- Naïve Bayes (NB) is a probabilistic machine learning algorithm based on Bayes' theorem and widely used in classification tasks. Naïve means the features are independent of each other and changing the value of one feature does not directly change the value of any of the other features. NB classifies data by calculating the posterior probability for each class using the probability of the features belonging to the class. However, the simple assumption of NB, it is fast and effective in real problems. Bayesian belief networks are used to deal with the features' dependency [20].
- Support Vector Machine (SVM) is commonly used for classifying gene expession data due to the sparseness of solution sparseness of solution and it's ability to handle large feature space [22]. Firstly, data items are plotted in n-dimensional space. Then SVM finds the hyperplane that best differentiate between classes.

2.3.2. Unsupervised approach

Unsupervised is a form of learning that requires no labeled data. One of the common techniques is K-means where data with similar features are grouped in the same cluster.k-means in microarray analysis can be used to remove redundant genes by grouping similar data [11].

3. Different methods for classifying cancer

3.1. Filter approach

Purbolaksono et al. [5] introduced a system of 3 stages for classifying microarray data, the first was discretization which used k-means for transforming continuous data into discrete and dividing data into clusters. Then the second stage was feature selection, mutual information was used for dimensional reduction and obtaining informative genes. Finally, the Bayes theorem was implemented on five datasets. The Best result was obtained with k = 10 and the result showed that Bayesian Network methods have better performance than Naïve Bayes in classifying the microarray data.

Cilia et al. [8] compared the performance of various feature selection and classification techniques on six datasets. For feature selection, the authors focused on feature ranking techniques, which evaluate each feature singularly. The datasets were evaluated using a decision tree (DT), Random Forest (RF), KNN, and multilayer perceptron classifiers with 10-fold cross-validation (CV). The result of utilized filter techniques was compared with the Sequential Forward Floating Search, the Fast Correlation-Based Filter, and the Minimum Redundancy Maximum Relevance. The ranking techniques obtained high results with three datasets. While FCBF and SFFS obtained high results for the other three. However, with the high result obtained, there was a need for further reduction in Ovarian, Lymphoma, and Lung datasets.

Aydadenta and Adiwijaya [11] utilized k-means and IG for feature selection. Initially, k-means was used to group similar features in one cluster, so a redundant one is removed. Then Relief algorithm was used to rank the clusters' elements and top-ranking features of each cluster were combined to train RF. The proposed model was evaluated on three datasets and showed a higher result than the model using RF only without clustering [23].

V. Bolón.et.al. [12] reviewed state of art techniques applied in the domain of microarray classification. Then a practical evaluation was done to compare the performance of the different techniques. different feature selection techniques eg. ReliefF, SVM-RFE, mRMR, IG, and FCBF were used for gene selection. Then C4.5, NB, and SVM were tested to get the accuracy of the model. The result showed the efficiency of SVM to obtain high accuracy. Al-Batah et al.[24] used the filter method, CFS to remove redundant genes and get the informative ones, then for classification process Decision Table, JRip, and OneR were used. The proposed approach can achieve high accuracy and fast computational speed with just a few numbers genes.

Gao et al. [25] proposed PA-SVM that combines PSO with ABC named (PA) to optimize the classification of SVM. FCBF was initially used to obtain informative genes. Then PA-SVM evaluated 9 datasets. The result was compared with other classifiers. According to the result, PA-SVM achieved good results with just a few numbers of genes.

Baliarsingh et al [26] proposed Jaya optimized extreme learning machine (JELM) for breast cancer classification. Jaya is used to select the optimal input weights and hidden biases for ELM. The authors used Wilcoxon rank sum test to select relevant genes. The performance of JELM was compared by the performance of SVM, KNN, NB, and c4.5 and achieved a higher result about 90.91%. although the proposed model achieved high accuracy, it selected a huge subset of about 505 genes so it needs a further reduction in the genes subset.

Su et al. [27] introduced a gene selection method based on Kolmogorov-Smirnov (K-S) test and CFS. Firstly, K-S test removed redundant and noise genes by comparing the distribution of two sample types. Then, the filtered subset was evaluated by CFS. Only genes that are highly correlated with the class and have low redundancy remained. Finally, the proposed method the evaluation of proposed method was done using SVM classifier with 10-fold CV. It's the result was compared with other FS techniques. K-S test-CFS had superior performance but optimization in running time was needed.

Ahmad, F. K [28] utilized different filter feature selection techniques namely SNR, FC, IG, and t-Test to select the informative genes. Gene selection techniques were applied on three datasets. Finally, SVM was used to evaluate the proposed methods. IG was effective to select a minimum set of attributes and SVM had high accuracy with IG and SNR techniques.

3.2. Hybrid approach

WU et al. [23] proposed, a hybrid improved binary quantum particle swarm optimization algorithm HI-BQPSO for feature selection, combining the advantages of filtering and a random heuristic search. Firstly, the maximum information coefficient (MIC) was used to calculate the correlation between features and class to obtain an initial feature subset. Then the improved BQPSO was used to obtain the optimized feature subset. The proposed model was evaluated using 9 gene datasets with SVM classifier. However, HI-BQPSO had good overall performance and strong searchability, it still needs improvement especially for CNS dataset.

Medjaheda et al. [29] proposed an approach to diagnosis cancer. In the first phase, Support Vector Machines based on Recursive Feature Elimination (SVM-RFE) was used to eliminate 40 percent of features. The remaining subset was processed via Binary Dragonfly (BDF) to retain informative genes only. The proposed method was evaluated on 6 microarray cancer datasets. However, the model achieved comparable results but for breast, it was not satisfying as it achieved high accuracy but with a very huge number of features.

Jain et al. [30] proposed a hybrid feature selection method that combined CFS and IBPSO. The IBPSO enhanced the early convergence to the local optimum of BPSO. The proposed method was utilized on 11 microarray datasets and evaluated by NB with stratified 10 k-CV. The model was compared with 7 classifiers and outperformed them in terms of accuracy and number of selected genes in most cases.

Shahbeig et al [31] proposed a hybrid TLBO-PSO that combined teaching learning-based optimization (TLBO) algorithm and mutated fuzzy adaptive particle swarm optimization (PSO) algorithm. The mutated PSO is used to overcome PSO possibility to be trapped in the local optimum solutions. A constant or even linearly changed value of

inertia weight may prevent the PSO algorithm from reaching the optimum result. Fuzzy tuning of the inertia weight based on the proposed total normalized function value can enhance The convergence speed of PSO and avoid trapping at the local optimum. The proposed method was evaluated using SVM and achieved 91.88% accuracy with 195 features.

Lu et al. [32] proposed MIMAGA, a hybrid feature selection algorithm combining mutual information maximization (MIM) and the adaptive genetic algorithm (AGA). Initially, MIM was applied as a preprocessing step to obtain a subset contains only 300 genes. Then wrapper technique, AGA, was applied. Finally, extreme learning machine was applied as a classifier on the data set. MIMAGA was compared with other FS techniques. The result showed that while MIMAGA takes a long time, it was efficient and had the best result.

Alomari et al. [33] proposed a hybrid filter-wrapper gene selection method using the filter approach, Minimum Redundancy Maximum Relevancy, and wrapper approach flower pollination algorithm (FPA). Initially, MRMR was employed to obtain important genes, that have the minimum redundancy for input genes and the maximum relevancy to the target class, from the gene expression data. Then these genes were used by FPA to get the most informative ones. The proposed model was evaluated on three datasets and compared with MRMR-GA. the proposed method showed comparative results regarding the accuracy and a low number of features.

Turgut et al.[34] used Recursive Feature Elimination(RFE) and Randomized Logistic Regression (RLR) feature elimination methods to select informative genes. The proposed method selected the top 50 features. Performance of the proposed methods was evaluated using 8 classifiers: SVM, KNN, Multi-Layer Perceptron, DT, RF, LR, AdaBoost, and Gradient Boosting Machines with k- CV on two different breast cancer datasets. The best result was achieved with SVM as a classifier for both datasets.

Mufassirin and Ragel. [35] proposed a novel filter- wrapper based feature selection approach. Initially, a filter method gain ratio was used to determine the importance of genes, by measuring the gain ratio concerning the relevant class to eliminate irrelevant and redundant genes. The second phase wrapper subset evaluator was used to evaluate the subset produced after using gain ratio. Finally, the proposed approach was evaluated using J48, DT, NB, Sequential Minimal Optimization on five datasets. The proposed model had time efficiency and gave high results.

Sreepada et al. [36] proposed a hybrid of filter-wrapper approach for gene selection to combine the fast computation of the filter approach and the accuracy of the wrapper approach. Firstly, Filter techniques are computation- ally faster, and the wrapper approach is more efficient for classification accuracy. Each of F-Score and IG was separately used to produce a subset for each, then both sub- sets were combined. Then wrapper methods, Sequential Backward Elimination (SBE) and Sequential Forward Selection (SFS) with SVM were used to get the informative genes. The proposed method was evaluated using three datasets and achieved good results of more than 97% for two datasets.

Hameed et al. [37] proposed a hybrid approach to elect the informative genes. Firstly, Pearson correlation coefficient (PCC) was ran 10 times to select the top 100 ranked features. Then either binary PSO or GA was used for further reduction. Different classifiers were employed to test the accuracy of eleven datasets based on 10-fold CV. The result showed that SVM had higher accuracy and BPSO performing faster and have high result than GA with a smaller number of selected genes.

Salem et al. [38] proposed a hybrid approach named (IG- SGA). Initially, IG was used with various thresholds to reduce the feature set. Then the reduced subset was passed to GA to obtain the most informative gene. Finally, genetic programming was used to classify seven datasets. The performance was assessed using 10-fold CV. However, the proposed model showed a higher result, needs further improvement was needed specifically for Lung- Ontario datasets and there was a limitation in terms of the time complexity.

Utami and Rustama [39] proposed a hybrid method PSO–SVM and ABC–SVM to obtain an accurate result in diagnosing breast cancer. Initially, PSO and ABC were used as feature selection techniques. Then SVM was used as a classifier. The result showed that ABC–SVM had accurate results and it was effective to deal with high dimensional data like microarray.

Zhongxin et al. [40] proposed a Feature Selection Algorithm based on Mutual Information and Lasso(FSMIL). In the first stage, MI was used to filter irrelevant features. Then in the next stage, an improved version of lasso was trained in the candidate subset to produce the most informative genes. the produced methods were applied on five datasets. To test the accuracy of the methods SVM classifier was utilized. The proposed method achieved high accuracy, especially for lung and Lymphoma datasets.

Sardana et.al. [41] proposed a hybrid approach Cluster quantum Genetic Algorithm(ClusterQGA) to accurately classify cancer. Initially, a cluster was used to remove irrelevant and redundant data, then the computer power of quantum and genetic algorithm were effectively used to select only relevant features. The proposed method was applied to four datasets and evaluated using SVM and KNN classifiers. However, ClusterQGA was successfully reduced the number of genes, the accuracy of classifying needs further improvement.

Singh and Sivabalakrishnan [42] presented a hybrid selection technique that comprised mRMR with Adaptive Genetic Algorithm (AGA). In the first phase, mRMR was effectively used to reduce the dimensions and the redundancy in the data. Subsequently produced subset was further processed via AGA to get the most relevant genes. The mRMR-AGA approach was evaluated by four classifiers on four benchmarked datasets and achieved comparable results.

Nagpala and Singhb [43] proposed qualitative mutual information (QMI) for feature selection. Initially, RF was used to obtain the importance of each gene which was used to calculate the preference score (PS). PS reduces the redundancy in the subset. Then MI was used to obtain the informative genes. The proposed method evaluated four datasets, and for classification, NB, C4.5, and IB1 were used with 10-fold CV. The result showed that the proposed method along with NB obtained an accurate result of more than 98% for two datasets.

Loey et al. [44] presented an intelligent decision support system for diagnosing microarray cancer data. Initially, IG was used to select relative genes. Then the selected subset was reduced via Grey Wolf Optimization (GWO). Finally, SVM was utilized for breast and colon cancer classification. However, the IG-GWO approach achieved high accuracy but with a huge number of features (about 240) for the breast cancer dataset.

Hamim et al. [45] combined a filter approach fisher score(F) with C5.0 to select relevant breast cancer genes. Initially, Fisher score removed redundant genes and reduced the subset to only 10% of genes. Then, C5.0 selects only 5 relevant genes. The proposed FC5 was assessed by C5.0, ANN, SVM, and LR with stratified 10-fold CV. C5.0 achieved higher accuracy about 93.28%.

3.3. Other approaches

Jinthanasatian et al. [46] utilized a neuro-fuzzy with firefly algorithm to classify microarray data. A neurofuzzy algorithm was used to select informative genes, and rule set generation as a classifier. firefly algorithm was utilized to optimize the parameters. The proposed method was evaluated on seven datasets and the accuracy was assessed with 10 k-fold method. The proposed algorithm provided comparable results with other techniques, but further improvement is needed especially for the colon dataset achieved only 76.94%.

Li et al. [47] proposed random value-based oversampling (RVOS) and an improved version of SVM-RFE to effectively analyze microarray data. Firstly, RVOS was utilized to balance the distribution of two samples. Then an improved version of linear SVM (LLSVM) with the improved RFE strategy was used to get the informa- tive genes. Finally, the proposed model was evaluated using four classifiers with stratified

Table 2
Different Methods for classifying breast cancer.

Ref	Feature selection	Classifier	Dataset[ref]	Classificationaccuracy	No.Genes
Purbolaksono et al. [5]	MI	BN	Colon [48]	86.7%	NA
		BN	Ovarian [49]	98.4%	
		BN	Leukemia [50]	88.2%	
		BN	Breast [51]	84%	
		NB	Lung [52]	98.66%	
Cilia et al.[8]	GR	KNN	Breast [53]	91.96%	50
	GR	KNN	Colon [48]	91.94%	10
	FCBF	NN	Leukemia [50]	99.44%	51
	FCBF	NN	Lymphoma	100%	128
	SFFS	NN	Lung [52]	97.92%	308
	GR	NN	Ovarian [49]	96.85%	500
V.Bolón.et.al. [12]	ReliefF	NB	Breast [51]	89%	50
	ReliefF	SVM	Prostate [54]	97%	50
	IG	SVM	Colon [48]	85%	50
Al-Batah et al.[24]	CFS	JRip	Breast [51]	88.7%	138
		Decision Table, JRip	Colon [48]	96.8%	26
		Decision Table	CNS	90.0%	9
		Decision Table	Leukemia [50]	98.6%	79
		JRip	Lung	97.5%	548
Gao et al. [25]	FCBF	PA-SVM	Breast	88.66%	92
			Lung	79.49%	6
			NervSys	91.67%	28
			Prostate	100%	27
			Colon	93.55%	14
			Leukemia	100%	97
			Ovarian	100%	30
			DLBCL1	100%	73
			DLBCL2	86.21%	27
Baliarsingh et al. [26]	Wilcoxon rank sum test	JELM	Breast[51]	90.91%	505
Su et al. [27]	K-S test + CFS	SVM	Breast [51]	87.4%	11.7
			Lung [52]	91.6%	23
			Colon [48]	90.1%	10.7
			Ovarian [49]	98.5%	33.2
			Leukemia [50]	79.6%	25.2
Ahmad, F. K. [28]	IG	SVM	Breast	80%	200
	IG		Colon	87%	100
	SNR		Lung	91%	100
Medjaheda el at. [29]	SVM-RFE+BDF	SVM	Breast[51]	89.47%	7237
Jain el at. [30]	CFS-iBPSO	NB	Breast[51]	92.75%	32.7
Shahbeig et al. [31]	TLBO-PSO	SVM	Breast[51]	91.88%	195
Lu et al. [32]	MIMAGA	ELM	Leukemia	95.95%	60
			Colon [48]	85.45%	77
			Prostate	97.12%	60
			Lung	93.75%	74
			Breast	87.12%	59
			SRBCT	90.11%	78
Alomari et al. [33]	MRMR-GA		Colon	88.01%	6.73
	MRMR-FPA		Ovarian	100%	4
	MRMR-FPA		Breast	85.88%	16.8
Turgut et al. [34]	RFE + RLR	SVM	Breast [51]	87.87%	50
Mufassirin and Ragel. [35]	GR + Wrapper	NB	Breast [51]	89.69%	NA
	Subset	Bayes Net	Colon [48]	95.16%	
	Evaluator	NB	Lung	97.04%	
		NB	Leukemia [50]	100%	
		NB	Ovarian [49]	100%	
Hameed et al. [37]	PCC-BPSO	SVM	Brain	97.62%	13
	PCC-BPSO	SVM	Breast [51]	90.72%	41
	PCC-BPSO	SVM	CNS	98.33%	39
	PCC-BPSO	Bayes net	Colon [48]	93.55%	23
	PCC-BPSO	KNN||Bayes net	Leukemia	100%	17
	PCC-BPSO	NB	Lung [52]	98.03%	39
	PCC-BPSO	NB	Lymphoma	100%	19
	PCC-GA	SVM	MLL	100%	22
	PCC-BPSO	KNN	Ovarian	100%	15
	PCC-BPSO	SVM	Prostate	97.06%	33
	PCC-BPSO	RF||SVM	SRBCT	100%	19
Utami and Rustama. [39]	ABC	SVM	Breast	88%	NA
Sardana et.al[41]	ClusterQGA	SVM	Breast [51]	86.6%	21
		KNN	Melanoma	94.74%	6
		KNN	Colon[48]	100%	11
		SVM	Prostate	100%	14
Singh andSivabalakrishNA[42]	mRMRAGA	ELM	Breast[51]	86.73%	140
Nagpala andSinghb[43]	QMI	IB1	Colon	87.09%	68
			Breast [51]	90.72%	98

(continued on next page)

Table 2 (*continued*)

Ref	Feature selection	Classifier	Dataset[ref]	Classificationaccuracy	No.Genes
			Leukemia	98.61%	93
Loey el al.[44]	IG + GWO	SVM	Colon[48]	95.9%	16
			Breast [51]	94.87%	240
Hamim et al. [45]	FC5	C5.0	Breast[51]	93.28%	5
Jinthanasatian et al. [46]	a neuro-fuzzy algorithm	rule set generation	Lung [52]	93.42%	4
			Ovarian [49]	96.13%	12
			Prostate [54]	87.43%	5
			Leukemia [50]	82.27%	7
			Breast [51]	82.37%	7
			Colon	76.94%	11
			DLBCL	83.81%	13
Li et al. [47]		SVM-RFE	Prostate [54]	92.20%	NA
		SVM-RFE	Breast [51]	86.09%	
		SVM-VSSRFE	CNS [55]	88.39%	
		SVM-RFE	Colon [48]	93.75%	
		SVM-VSSRFE	Ovarian [49]	100%	
		SVM-VSSRFE	Leukemia [50]	100%	

Table 3
Different Methods for classifying other cancer.

Ref	Feature selection	Classifier	Dataset[ref]	Classificationaccuracy	No.Genes
Aydadenta and Adiwijaya.[11]	K-means -Relief	RF	Colon [48]	85.87%	NA
			Lung [52]	98.90%	
			Prostate[54]	88.97%	
WU et al. [23]	MIC- BQPSO	SVM	Leukemia	97.81%	NA
			Colon	88.36%	
			Prostate	91.60%	
			DLBCL	96.8%	
			CNS	74.64%	
Sreepada et al. [36]	(F-Score, IG)And (SBE and SFS)	SVM	Colon [48]	87.5%	12
			DLBCL	100%	14
			Leukemia	97.37%	16
Salem et al. [38]	IG-SGA	GP	Leukemia [50]	97.06%	3
			Colon [48]	85.48%	60
			CNS [55]	86.67%	38
			Lung-Ontario	74.4%	11
			Lung –Michigan	100%	9
			DLBCL	94.80%	110
			Prostate cancer	100%	26
Zhongxin et al.[40]	FSMIL	SVM	Colon	90.86%	NA
			Prostate	96.52%	
			Lymphoma	98.68%	
			Leukemia	98.63%	
			Lung	100%	

5-fold CV. Results showed that SVM-VSSRFE had better results for three datasets, and also the efficiency of LLSVM-VSSRFE in reducing time consumption, especially with high dimensional datasets.

Different Methods for classifying breast cancer,other cancer types are presented in tables 2,3 respectively. The accuracy of state of art methods are presented in Fig. 4.

4. Discussion

Although microarray data are proven to be efficient for diagnosing cancer, the huge number of its features with respect to small sample size, for example, breast datasets Van't Veer [51] and wang [53] have 24,482 and 18,000 features with only 97 samples, cause a so-called curse of dimensionality problem. To avoid it hybrid and filter selection techniques are commonly used. Filter approach is fast and isn't computationally extensive so it is used in [5,8,12,24–28] and recommended to be initially used in hybrid approach in [29–35,37,39,41–46]. Applying filter approach on Van't Veer, K-S test-CFS in [27] generated a small subset in comparison to subset generated by other filter technique but with lower accuracy about 87.4%.While in [26] Wilcoxon rank-sum test achieved high accuracy about 90.91% but with a large gene subset about 505 genes. On the other hand applying GR on wang achieved higher accuracy with only 50 genes. While the filter approach may achieve a high result but it selects a larger number of features on the other hand the ability of evolutionary wrapper feature selection techniques to find optimal or near-optimal subset help hybrid approach to achieve higher accuracy with just a small subset. In [45], FC5 could generate the smallest subset about 5 genes, while the highest result achieved in [41] using IG-GWO but with a large subset of about 240 genes. However, the hybrid approach can achieve better performance than a filter, SVM-RFE, and BDF in [30] had the worst performance in terms of the Number of selected genes it selected about 7237 genes but with acceptable accuracy. For other cancer types, while GR in [8] generated a small subset for colon [48], it produced a large subset for ovarian [49] and lung [52]. PCC-BPSO in [37] produced a the small subset for ovarian and lung that led to the best accuracy using KNN and NB respectively. Applying ClusterQGA in [41] with KNN classifier led to the best performance for colon about 100% accuracy. SVM has a high accuracy of for breast, KNN has better accuracy for colon and ovarian and NB for lung. Another issue due to few samples for accurate validation, 10-fold CV is commonly used.

5. Conclusion and future work

Microarray data analysis deepens your understanding of cancer pathogenesis and also having diagnostic value. It accurately diagnoses

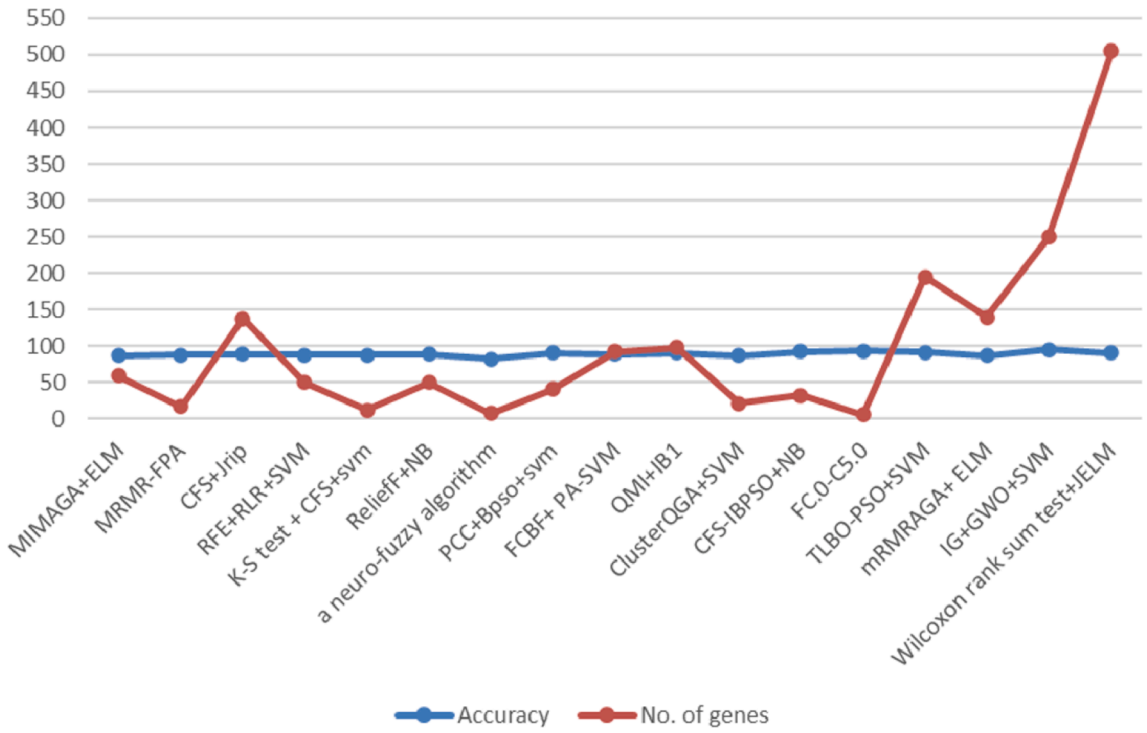

Fig. 4. The performance result of breast cancer dataset.

cancer. However, The accuracy influenced by a large number of features and the limited number of samples. Dimensionality reduction techniques, mainly feature selection approaches are utilized to overcome this deterioration inaccuracy. The survey reviewed the state of the art of feature selection and classification techniques. The review showed that SVM is the most applied classification algorithm and achieved a high result of about 94.87% with hybrid feature selection (IG-GWO). As future work, a hybrid feature selection technique based on a heuristic search algorithm will be examined to obtain a more accurate result.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: a cancer journal for clinicians 68 (2018) 394–424.
[2] N. Eliyatkın, E. Yalçın, B. Zengel, S. Aktaş, E. Vardar, Molecular classification of breast carcinoma: from traditional, old- fashioned way to a new age, and a new way, The journal of breast health 11 (2015) 59.
[3] Lindsey A. Torre, Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, Ahmedin Jemal, Global cancer statistics, 2012: Global Cancer Statistics, 2012, CA: A Cancer Journal for Clinicians 65 (2) (2015) 87–108, https://doi.org/10.3322/caac.21262.
[4] R. Priya, P.S. Vadivu, A Review on Data Mining Techniques for Prediction of Breast Cancer Recurrence, International Journal of Engineering and Management Research (IJEMR) 9 (2019) 142–146.
[5] M.D. Purbolaksono, K.C. Widiastuti, M.S. Mubarok, F.A. Ma'ruf, Implementation of mutual information and bayes theorem for classification microarray data, Journal of Physics: Conference Series, IOP Publishing (2018), 012011.
[6] M.A. Makary, M. Daniel, Medical error—the third leading cause of death in the US, Bmj 353 (2016).
[7] H.J. Hong, W.S. Koom, W.-G. Koh, Cell microarray technologies for high-throughput cell-based biosensors, Sensors 17 (2017) 1293.
[8] N.D. Cilia, C. De Stefano, F. Fontanella, S. Raimondo, A. Scotto di Freca, An experimental comparison of feature- selection and classification methods for microarray datasets, Information 10 (2019) 109.
[9] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, G. Han, Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles, IEEE/ACM transactions on computational biology and bioinformatics 11 (2014) 727–740.
[10] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and structural biotechnology journal 13 (2015) 8–17.
[11] H. Aydadenta, A Clustering Approach for Feature Selection in Microarray Data Classification Using Random forest, Journal of Information Processing Systems 14 (2018).
[12] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Information Sciences 282 (2014) 111–135.
[13] J.R. Vergara, P.A. Estévez, A review of feature selection methods based on mutual information, Neural computing and applications 24 (2014) 175–186.
[14] B. Azhagusundari, A.S. Thanamani, Feature selection based on information gain, International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2 (2013) 18–21.
[15] M.A. Hall, L.A. Smith, Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper, FLAIRS conference (1999) 235–239.
[16] N. Almugren, H. Alshamlan, A survey on hybrid feature selection methods in microarray gene expression data for cancer classification, IEEE Access 7 (2019) 78533–78548.
[17] M.S. Hossain, A. El-Shafie, Application of artificial bee colony (ABC) algorithm in search of optimal release of Aswan High Dam, Journal of Physics: Conference Series, IOP Publishing (2013), 012001.
[18] D. Wang, D. Tan, L. Liu, Particle swarm optimization algorithm: an overview, Soft Computing 22 (2018) 387–408.
[19] X.-S. Yang, A new metaheuristic bat-inspired algorithm, Nature inspired cooperative strategies for optimization (NICSO, Springer 2010 (2010) 65–74.
[20] S.A. Abdulrahman, W. Khalifa, M. Roushdy, A.M. Salem, Comparative study for 8 computational intelligence algorithms for human identification, Comput. Sci. Rev. 36 (2020), 100237.
[21] Widiawati, I.F., Nugrahapraja, H., Fajriyah, R. (2018). K-Nearest Neighbor (KNN) Analysis on Genes Expression Datasets of Maize Nested Association Mapping (NAM) Showed Confident Classification on Organ-specific Expression. 2018 1st International Conference on Bioinformatics, Biotechnology, and Biomedical Engineering - Bioinformatics and Biomedical Engineering, 1, 1-3.
[22] B. Sahu, S. Dehuri, A.K. Jagadev, Feature selection model based on clustering and ranking in pipeline for microarray data, Informatics in Medicine Unlocked 9 (2017) 107–122.
[23] Q. Wu, Z. Ma, J. Fan, G. Xu, Y. Shen, A feature selection method based on hybrid improved binary quantum particle swarm optimization, IEEE Access 7 (2019) 80588–80601.
[24] M.S. Al-Batah, B.M. Zaqaibeh, S.A. Alomari, M.S. Alz-boon, Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers, International Journal of Online and Biomedical Engineering (iJOE) 15 (2019) 62–73.

[25] L. Gao, M. Ye, C. Wu, Cancer classification based on support vector machine optimized by particle swarm optimization and artificial bee colony, Molecules 22 (2017) 2086.
[26] S.K. Baliarsingh, C. Dora, S. Vipsita, Jaya Optimized Extreme Learning Machine for Breast Cancer Data Classification, Springer Singapore, Singapore, 2021, pp. 459–467.
[27] Q. Su, Y. Wang, X. Jiang, F. Chen, W.-C. Lu, A cancer gene selection algorithm based on the KS test and CFS, BioMed research international 2017 (2017).
[28] F.K. Ahmad, A comparative study on gene selection methods for tissues classification on large scale gene expression data, Jurnal Teknologi 78 (2016) 116–125.
[29] S.A. Medjahed, T.A. Saadi, A. Benyettou, M. Ouali, Kernel-based learning and feature selection analysis for cancer diagnosis, Applied Soft Computing. 51 (2017) 39–48.
[30] I. Jain, V.K. Jain, R. Jain, Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification, Appl Soft Comput. 62 (2018) 203–215.
[31] S. Shahbeig, M.S. Helfroush, A. Rahideh, A fuzzy multi-objective hybrid TLBO-PSO approach to select the associated genes with breast cancer, Signal Process. 131 (2017) 58–65.
[32] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, Z. Gao, A hybrid feature selection algorithm for gene expression data classification, Neurocomputing 256 (2017) 56–62.
[33] O.A. Alomari, A.T. Khader, M.A. Al-Betar, Z.A.A. Alyasseri, A hybrid filter-wrapper gene selection method for cancer classification, 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS), IEEE, 2018, pp. 113- 118.
[34] S. Turgut, M. Dağtekin, T. Ensari, Microarray breast cancer data classification using machine learning methods, 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), IEEE, 2018, pp. 1-3.
[35] M.M. Mufassirin, R.G. Ragel, A novel filter-wrapper based feature selection approach for cancer data classification, 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), IEEE, 2018, pp. 1-6.
[36] R.S. Sreepada, S. Vipsita, P. Mohapatra, An efficient approach for microarray data classification using filter wrapper hybrid approach, 2015 IEEE International Advance Computing Conference (IACC), IEEE, 2015, pp. 263-267.
[37] S.S. Hameed, F.F. Muhammad, R. Hassan, F. Saeed, Gene Selection and Classification in Microarray Datasets using a Hybrid Approach of PCC-BPSO/GA with Multi Classifiers, JCS. 14 (2018) 868–880.
[38] H. Salem, G. Attiya, N. El-Fishawy, Classification of human cancer diseases by gene expression profiles, Applied Soft Computing 50 (2017) 124–134.
[39] D. Utami, Z. Rustam, Gene selection in cancer classification using hybrid method based on Particle Swarm Optimization (PSO), Artificial Bee Colony (ABC) feature selection and support vector machine, AIP Conference Proceedings, AIP Publishing LLC (2019), 020047.
[40] W. Zhongxin, S. Gang, Z. Jing, Z. Jia, Feature selection algorithm based on mutual information and Lasso for microarray data, The Open Biotechnology Journal 10 (2016).
[41] M. Sardana, R. Agrawal, B. Kaur, A hybrid of clustering and quantum genetic algorithm for relevant genes selection for cancer microarray data, International Journal of Knowledge-based and Intelligent Engineering Systems 20 (2016) 161–173.
[42] R.K. Singh, M. Sivabalakrishnan, Microarray Gene Expression Data Classification using a Hybrid Algorithm: MRMRAGA, International Journal of Innovative Technology and Exploring Engineering (IJITEE) August (8) (2019).
[43] A. Nagpal, V. Singh, A feature selection algorithm based on qualitative mutual information for cancer microarray data, Procedia computer science, 132 (2018) 244–252, Biotechnology Journal 10 (2016).
[44] Loey M, Jasim MW, EL-Bakry HM, Taha MHN, Khalifa NEM. Breast and Colon Cancer Classification from Gene Expression Profiles Using Data Mining Techniques. Symmetry. 2020;12:408.
[45] M. Hamim, I. El Moudden, H. Moutachaouik, M. Hain, Decision Tree Model Based Gene Selection and Classification for Breast Cancer Risk Prediction, Springer International Publishing, Cham, 2020, pp. 165–177.
[46] P. Jinthanasatian, S. Auephanwiriyakul, N. Theera-Umpon, Microarray data classification using neuro-fuzzy classifier with firefly algorithm, 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017, pp. 1-6.
[47] Z. Li, W. Xie, T. Liu, Efficient feature selection and classification for microarray data, PloS one 13 (2018).
[48] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of the National Academy of Sciences 96 (1999) 6745–6750.
[49] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, Use of proteomic patterns in serum to identify ovarian cancer, The lancet 359 (2002) 572–577.
[50] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasen-beek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, science 286 (1999) 531–537.
[51] L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. Van Der Kooy, M.J. Marton, A.T. Witteveen, Gene expression profiling predicts clinical outcome of breast cancer, nature, 415 (2002) 530-536.
[52] G.J. Gordon, R.V. Jensen, L.-L. Hsiao, S.R. Gullans, J.E. Blu-menstock, S. Ramaswamy, W.G. Richards, D.J. Sugarbaker, R. Bueno, Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, Cancer research 62 (2002) 4963–4967.
[53] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E. Meijer-van Gelder, J. Yu, Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, The Lancet 365 (2005) 671–679.
[54] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, W. Sellers, Gene Expression Correlates of Clinical Prostate Cancer Behavior, Cancer cell 1 (2002) 203–209.
[55] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M. E. McLaughlin, J.Y. Kim, L.C. Goumnerova, P.M. Black, C. Lau, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (2002) 436–442.