Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection

Jun Chin Ang, Andri Mirzal, Habibollah Haron, and Haza Nuzly Abdull Hamed

Abstract—Recently, feature selection and dimensionality reduction have become fundamental tools for many data mining tasks, especially for processing high-dimensional data such as gene expression microarray data. Gene expression microarray data comprises up to hundreds of thousands of features with relatively small sample size. Because learning algorithms usually do not work well with this kind of data, a challenge to reduce the data dimensionality arises. A huge number of gene selection are applied to select a subset of relevant features for model construction and to seek for better cancer classification performance. This paper presents the basic taxonomy of feature selection, and also reviews the state-of-the-art gene selection methods by grouping the literatures into three categories: supervised, unsupervised, and semi-supervised. The comparison of experimental results on top 5 representative gene expression datasets indicates that the classification accuracy of unsupervised and semi-supervised feature selection is competitive with supervised feature selection.

Index Terms—Feature selection, gene expression, semi-supervised, supervised, unsupervised

1 Introduction

Feature selection is a dimensionality reduction technique that is commonly used in the fields of machine learning, pattern recognition, statistics, and data mining. This technique aims to select a subset of relevant features from the original set of features according to some criteria. Some examples of feature selection techniques include Information Gain, Relief, Chi Squares, Fisher Score, and Lasso. Feature selection usually is used in the domains where the datasets comprise of thousands of features but with relatively small sample size (e.g., gene expression data). Feature selection that is applied to gene expression data is also known as gene selection [1]. Gene selection is necessary as the data usually contains many irrelevant, redundant, and noisy expressions, and also is effective for early tumor detection and cancer discovery as it leads to a more reliable cancer diagnosis or prognosis and better clinical treatment [2].

The gene expression data can either be fully labeled, unlabeled, or partially labeled. This leads to the development of supervised, unsupervised and semi-supervised gene selection to discover the biological patterns and class prediction [3]. Typically, unlabeled data is composed of samples and their features without the presence of labels indicating any explanation or information. Whereas, labeled data uses a set of data that are marked with some meaningful labels or classes. Supervised feature selection is the process of selecting a feature subset based on some criteria for measuring importance and relevance of the features by using labeled data. Unsupervised feature selection, where there is no prior knowledge about the true functional classes, evaluates feature relevance by exploiting the innate structures of the data, such as data variance, separability, and data distribution. A semi-supervised feature selection integrates a small amount of labeled data into unlabeled data as additional information to improve the performance of an unsupervised feature selection. Even though there are a lot of review papers on gene selection in the literatures [4], [5], [6], [7], [8], but to the best of our knowledge, there is no detailed discussion on the methods and separates them in such three categories as in this paper.

This paper is divided into six sections. Section 2 presents an overview of feature selection and Section 3 gives a review on some gene selection approaches especially that have been proposed over the past five years and further group them in three categories: supervised, unsupervised and semi-supervised approaches. Section 4 describes the challenges inherent in gene selection task. Section 5 discusses the gene selection approaches reviewed in the previous sections and collates the experimental results of gene selection methods on several benchmark datasets. The overall discussion with a few recommendations for future directions is presented in the last section.

2 Overview of Feature Selection

Feature selection aims to select a feature subset from the original set of features based on feature's relevance and redundancy. Yu and Liu [9] classify the feature subsets into four categories: (a) completely irrelevant and noisy features, (b) weakly relevant and redundant features, (c) weakly relevant and non-redundant features, and (d) strongly relevant features. An optimal subset principally contains all the features in the category (c) and (d). Strongly relevant features are indispensable for enhancement of discriminative power

- J.C. Ang, H. Haron, and H.N.A. Hamed are with the Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, Johor, Malaysia.
E-mail: jcang2@live.utm.my, {habib, haza}@utm.my.
- A. Mirzal is with the College of Graduate Studies, Arabian Gulf University, Manama, Bahrain. E-mail: andrim@agu.edu.bh.

Manuscript received 8 July 2014; revised 20 Jan. 2015; accepted 11 Aug. 2015. Date of publication 14 Sept. 2015; date of current version 4 Oct. 2016.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2015.2478454

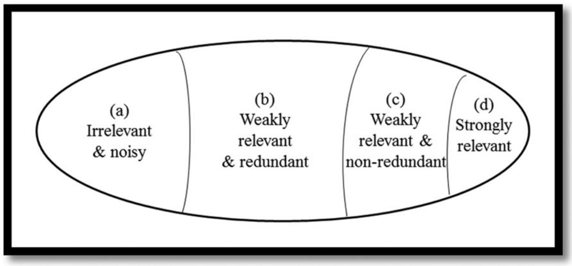


Fig. 1. Feature classifications based on relevancy and redundancy.

and prediction accuracy. Sometimes, weakly relevant features can be useful in improving prediction accuracy if the feature is non-redundant and compatible with evaluation measures. An irrelevant feature indicates that the feature does not contribute to prediction accuracy. Thus, ideally all strongly relevant features and some weakly relevant features should be selected, and irrelevant, redundant or noisy features should be eliminated in order to build a good model prediction. The main reason for eliminating the redundant features is because they may have significant statistical relations with other features, not because they contain worthless information. As a single entity, a feature may be irrelevant, but can be highly relevant when combined with other features [10]. A method proposed by Ding and Peng [11], called Minimal redundancy and maximum relevancy (mRMR) is a novel feature selection approach that is based on mutual information (MI) as a measure of both relevancy and redundancy where the redundancy of a selected feature subset S is an aggregate MI measure between each pair of features in S and the relevancy to a class c is an aggregate MI measure between each feature with respect to c. This approach has then been improved and integrated in many others literatures, such as [12], [13], [14], [15], [16], [17], [18]. Fig. 1 shows feature classifications based on relevancy and redundancy.

Feature selection provides a lot benefits as it improves prediction performance, understandability, scalability, and generalization capability of the classifier. It also reduces computational complexity and storage, provides faster and more cost-effective model [19], and plays an important role in knowledge discovery. Moreover, it offers new insights for determining the most relevant or informative features. However, feature selection consists of several complex stages that usually are costly. And even the optimal model parameters of full feature set might need to be redefined for a few times in order to obtain the optimal model parameters for selected feature subsets.

In machine learning, a feature vector is an n-dimensional vector that represents the expression values of the feature over all samples. The space associated with these vectors is often called the feature space. In order to reduce dimensionality of the feature space, feature selection or feature extraction techniques can be employed. Feature selection is a subset of a general field known as feature extraction. Feature extraction is a technique that transforms the original feature space into a distinct space with different set of axes in order to reduce the dimensionality of the data [20], whereas feature selection reduces the original feature space into a subspace without transformation. Some common feature extraction techniques include Principle Component Analysis (PCA), Factor Analysis (FA), Linear Discriminant Analysis (LDA), and Singular Value Decomposition (SVD). And some examples of feature selection techniques include Information Gain, Relief, Chi Squares, Fisher Score, and Lasso. Compared to feature selection, feature extraction is more general and the transformation mapping may provide a better discriminatory ability. But the transformed distinct space is problematic since there may be no physical meaning for better interpretation [21]. Hence, this paper will discuss feature selection since it is considered to be more superior in terms of readability and interpretability [5].

There are three types of feature selection method, supervised, unsupervised and semi-supervised. Supervised feature selection is the earliest and most common practice approach [48]. Supervised feature selection utilizes the labeled data in the process of feature evaluation. However, data is abundant and continues to accumulate at an unprecedented rate. The labeled data given by external knowledge is costly to obtain and may be unreliable and mislabeled [49]. This fact aggravates the risk of over-fitting the learning process in the supervised feature selection by either unintentionally removing many relevant features or selecting irrelevant features. There are some literature reviews on feature selection, for example [4], [5], [6], [7]. Most of them discussed the supervised feature selection.

Unsupervised feature selection is more challenging than supervised and semi-supervised approaches because it is unassisted by labeled data. Nevertheless it has several advantages, e.g., it is unbiased since there is no need to utilize experts or data analysts to categorize the samples, and it still can perform well even when no prior knowledge is available. Unsupervised feature selection is essential for exploratory analysis of biological data and it provides an effective way to discover the unknown meaningful insights for classification of disease types [124]. The main drawbacks of unsupervised approach are that it neglects the possible correlation between different features, thus the produced subsets might be suboptimal for the actual discrimination task, and it relies on some mathematical principles without guarantee that the principles are universally valid for all data [125]. A good survey about the unsupervised wrapper feature selection approaches can found in Ref. [126].

Semi-supervised and semi-unsupervised feature selection are the extensions of supervised and unsupervised feature selection that work on both labeled data and unlabeled data. The term semi-supervised feature selection is used when the majority of data is labeled and semi-unsupervised feature selection is used when the majority of data is unlabeled. Usually the labeled data is used to maximize the margin between data points of different classes, and the unlabeled data is used to discover the geometrical structure of the space [145].

2.1 Development of Feature Selection

The process of selecting a subset of relevant and informative features from the original set of features can be divided into five main stages as shown in Fig. 2. The decision made at each stage influences the feature selection performance [22].

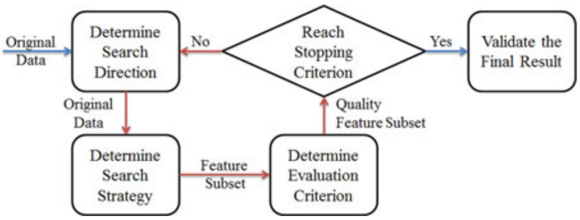

Original Data
Determine Search Direction
No
Reach Stopping Criterion
Yes
Validate the Final Result
Original Data
Quality Feature Subset
Determine Search Strategy
Feature Subset
Determine Evaluation Criterion

Fig. 2. The main stages in feature selection process.

Stage 1: Determine search direction. The first stage is to determine the starting point and the search direction. Search may start with an empty set and successively adds new features in each iteration, called forward search. In contrast, the search can be started with a full set and then the features are eliminated consecutively in each iteration, called backward elimination search. Another alternative is to begin with both ends by simultaneously adding and removing the features in each iteration, called bi-directional search. Search may also begin somewhere in the middle by randomly selecting the features to form the subset.

Stage 2: Determine search strategy. According to Gheyas and Smith [10], a good search strategy should provide good global search capability, rapid convergence to near optimal solution, good local search ability, and high computational efficiency. Search strategies can be categorized into three groups: exponential, sequential, and randomized.

Exponential search, also called complete search, is the most exhaustive global search strategy. It starts from the original feature set and guarantees to find the optimal result. However, this strategy is generally impractical and computationally intensive especially for high dimensional data sets, and prohibitive and intractable for all but a small initial number of features. An example of this strategy is exhaustive search [24], a search that evaluates all possible subsets to find the optimal subset.

Sequential search, also called greedy hill-climbing search, adds or removes one feature at a time. The most common sequential strategies are sequential forward selection (SFS) and sequential backward selection (SBS). It is relatively simple to implement, its complexity is polynomial with respect to the number of features, and it is robust to multi collinearity problems. However, these methods perform poorly on non-monotonic indices and may cause nesting effect [25] because once a feature is added (or deleted), it is not allowed to be deleted (or added) latter. Moreover, they are sensitive to feature interaction, hence they can easily be trapped into local minima [10]. Sequential forward floating selection (SFFS) and sequential backward floating selection (SBFS) [25] were developed to overcome these problems by providing mechanisms to re-select the deleted features and delete the already added features. Some other examples of sequential search strategies are best first search, beam search—an optimized solution of best first search, and plus *l* take-away *r* algorithm (PTA) [26].

Randomized search strategy starts by randomly selecting the features and then proceeds with two different search strategies. The first uses the classical sequential or bi-directional search, e.g., simulated annealing [23] and random hill-climbing [27]. And the second uses strategies that have no regular movements, e.g., genetic algorithm (GA) [28], Las Vegas algorithm [29], and Tabu search [30]. The second strategies can escape local optima in the search space, but they have a greater chance of producing incorrect results due to the nonavailability of mechanism to capture the relationship between the features.

Stage 3: Determine evaluation criterion. Originally evaluation methods of feature selection are classified into four types: filter, wrapper, embedded, and hybrid [31]. In recent years, another kind of evaluation method is developed, called ensemble feature selection [32].

Filter, or also called open-loop method, is the earliest method. It examines the features based on the intrinsic characteristics prior to the learning tasks. Filter algorithms principally measure the feature characteristics based on four types of evaluation criteria, i.e., dependency, information, distance, and consistency [33]. Most filter methods in literature are univariate. They are known to be very efficient and computationally faster hence more easily scale up to huge databases than wrapper methods. Filter method is independent of any learning algorithm therefore it can provide general solutions for various classifiers. Also the bias in the feature selection does not correlate with the bias in the learning algorithm, so it has a better generalization property [11]. However, filter method ignores the interactions between classifiers and the possible interaction among features (combined features may have net effects that are not necessarily reflected by the individual features in that group). It also leads to varying prediction performance when the same selected features are used in different learning algorithms [34]. For reference, Lazar [32] reviewed the filter method for feature selection in the gene microarray analysis.

Wrapper, or close-loop method, wraps the feature selection around the learning algorithm and utilizes classification error rate or performance accuracy as feature evaluation criterion. It selects the most discriminative feature subset by minimizing the prediction error of a particular classifier. This method often gives better performance results than the filter method because it takes into accounts the feature dependencies and directly incorporates bias in the learning algorithm. However, it is less general than the filter method because the selection process must be re-executed if another learning algorithm is utilized since there is no guarantee that the selected features is optimal for other learning algorithms. Furthermore, wrapper method is more prone to over-fitting than the filter method because the classifier is repeatedly called to evaluate each subset. The majority of wrapper methods are multivariate, hence they require extensive computation resources to achieve the convergences and can be intractable for large datasets.

Embedded method is a built-in feature selection mechanism that embeds the feature selection in the learning algorithm and uses its properties to guide feature evaluation. Embedded method is more efficient and computationally more tractable than wrapper method while maintaining similar performance. This is because the embedded method avoids the repetitive execution of classifier and examination of every feature subset. Moreover, this method has lower risk to over-fitting compared to wrapper method. Like wrapper, embedded method takes into account the dependencies among features, but is only specific to a given learning algorithm [35]. However the computational complexity is a major issue, especially in high-dimensional data.

Hybrid and ensemble methods represent latest developments in feature selection. Hybrid method can be either formed by combining two different methods (e.g., filter and wrapper), two methods of the same criterion, or two feature selection approaches. Hybrid method attempts to inherit the advantages of both methods by combining their complementary strengths [36]. It uses different evaluation criteria in different search stages to improve the efficiency and prediction performance with better computational performance. The most common hybrid method is the combination of filter and wrapper methods [34].

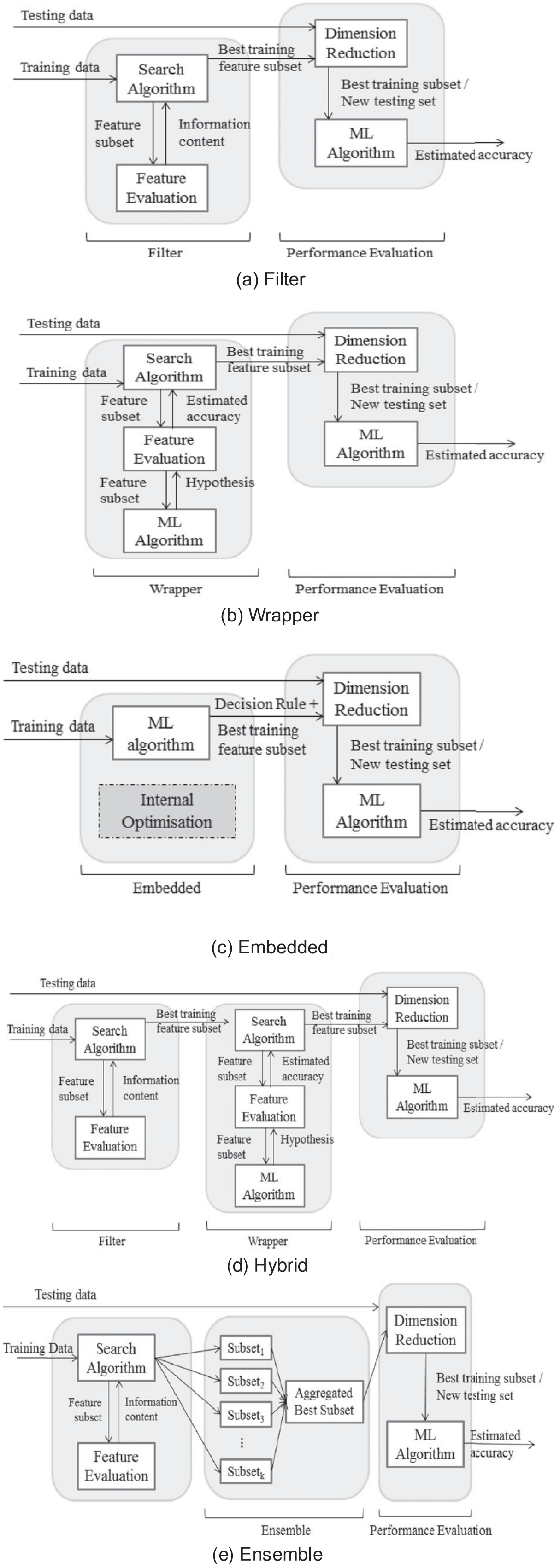
Ensemble method is a method that aims to construct a group of feature subsets and then produce an aggregate result out of the group [37]. It is purposely designed to tackle the instability and perturbation issues in many feature selection algorithms. This method is based on different subsampling strategies where a particular feature selection method is run on a number of subsamples and the obtained features are merged to form a more stable subset. The performance of feature selection is no longer depending on a single selected subset, thus it is more flexible and robust when dealing with high dimensional data. Moreover, ensemble method provides a better approximation to the optimal subset or ranking of features by aggregating the outputs of several feature selectors. A detailed discussion on ensemble feature selection can be found in [38].

Fig. 3 shows the taxonomy of feature evaluation methods and Table 1 indicates the advantages and disadvantages of each evaluation method.

Stage 4: Define stopping criteria. A stopping criterion determines when the feature selection process should halt. A suitable stopping criterion can avoid over-fitting and thus leads to a more efficient process in producing an optimal feature subset with lower computational complexity. The decisions made in the previous stages will influence the choice of stopping criterion. The common stopping criteria are:

- Predefined number of features,
- Predefined number of iterations,
- Percentage of improvement over two consecutive iteration steps, and
- Obtaining an optimal feature subset according to some evaluation function.

Stage 5: Validate the result. To evaluate the effectiveness of potential feature sets for classification and prediction, various error estimation or validation techniques have been proposed. The most common error estimation methods are cross validation (CV) and performance measurements based on confusion matrix. In addition, some other validation and analysis have also been performed in previous studies. For example, Rand index [39] and Jaccard index [40] for similarity measures; Kuncheva Index (KI) [41] for stability measure; analysis of variance (ANOVA) for complexity analysis; and Boolean threshold functions for representing gene expression signatures [42]. This paper discusses the usage of cross validation method and some common expressions derived from confusion matrix.



(a) Filter

(b) Wrapper

(c) Embedded

(d) Hybrid

(e) Ensemble

Fig. 3. Taxonomy of feature evaluation methods.

Cross validation is the most common and popular validation method. In this method, the original data sets are split into two parts: training and testing sets. The training set is used to train the classifier, and then the test set is used for the final evaluation. CV has the advantage of producing an effectively unbiased error estimate. This is because the CV procedure is repeated for different samples drawn from a

TABLE 1
Advantages and Disadvantages of Feature Evaluation Methods

	Advantages	Disadvantages
Filter	■ Faster than wrapper ■ Scalable ■ Classifier independent ■ Better computational complexity than wrapper ■ Better generalizable property	■ Ignore interaction between classifiers ■ Ignore dependency among features
Wrapper	■ Interact with classifier ■ Consider the dependence among features ■ Higher performance accuracy than filter	■ More prone to over-fitting ■ Classifier specific ■ Require expensive computation
Embedded	■ Interact with classifier ■ Better computational complexity than wrapper ■ Higher performance accuracy than filter ■ Less prone to over-fitting than wrapper ■ Consider the dependence among features	■ Classifier specific
Hybrid	■ Higher performance accuracy than filter ■ Better computational complexity than wrapper ■ Less prone to over-fitting than wrapper	■ Classifier specific
Ensemble	■ Less prone to over-fitting ■ More flexible and robust upon high dimensional data	■ Difficult to understand an ensemble of classifiers

population, the average error estimate will approximate the expected error for the designed classifiers across all possible equal-sized samples [43]. The main drawback of the CV is its error estimate is highly variable. Three common types of CV are k-fold, leave-one-out CV (LOOCV), and hold-out CV. The CV error rate (E) is defined as the average error rate on test subsamples (E_i) with the formula:

$$E = \frac{1}{k}\sum_{i=1}^{k} E_i$$

The results for a classifier can also be evaluated using the confusion matrix for two possible outcomes (Table 2).

The quality of a classification can be evaluated by computing the number of correctly recognized class examples (true positives), the number of correctly recognized examples that do not belong to the class (true negatives), and the number of examples that neither were correctly assigned to the class (false positives) nor recognized as class examples (false negative) [44]. Several standard measurements have been defined for the confusion matrix, the detailed information can be found in Ref. [45], [46], [47], for example:

- Error rate
- Classification
- TP rate / recall / sensitivity
- Specificity
- Precision
- F1-score/ F-score / F-measure
- Receiver operating characteristic (ROC) curve
- Area under curve (AUC)

3 A Review on Gene Selection

3.1 Supervised Gene Selection

Supervised gene selection utilizes labeled data to select relevant features in gene expression data. In this study, we focus only on the recent five years literatures. Table 3 summarizes the recent works on supervised gene selection.

The majority of researchers focus on the development of supervised feature selection with filter evaluation framework. For examples, Sun et al. [50] proposed a Local Learning based Feature Selection (LLBFS) method to handle the problems of complex distributed data and high data dimensionality. LLBFS is conceived as an extension of RELIEF and relies on kernel density estimation and margin maximization concepts. Lan and Vucetic [51] proposed a novel filter approach based on multi-task learning, which aims to improve the accuracy of target classifier by exploiting the auxiliary data. The multi-task filter selection method was shown to be very successful when applied in conjunction with both single-task and multi-tasks classifiers.

Aforementioned, wrapper framework needs more expensive computational cost than filter framework, hence it receives less attention. One of the wrapper example is successive feature selection (SFS) proposed by Sharma et al. [52]. Sharma et al. attempt to overcome the drawback of conventional feature selection algorithm whereby a weakly ranked gene that could beneficial for classification accuracy with an appropriate subset of genes has been left out from the selection.

Many embedded-based feature selection algorithms are designed by integrating regression as a constraint of existing learning models to achieve a sparse solution. For example, Nie et al. [53] and Xiang et al. [54] implemented a robust feature selection with $L_{2,1}$-norms, Du et al. [55] applied L_2-norm penalty with augmented data technique, and Liang et al. [56] established a regularized sparse multinomial

TABLE 2
Confusion Matrix Representation

Truth	Prediction	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

TABLE 3
Literature Details for Supervised Gene Selection

	Literature	Feature Selection	Microarray Data Set	Classifier	Validation Method
Filter	(Sun et al., 2010) [50]	Local-learning based feature selection	■ Prostate[60]; ■ Breast [61]; ■ DLBCL [62];	SVM; KNN (1-NN)	Complexity analysis; LOOCV
	(Wang and Gotoh, 2010) [63]	Canonical α depended degree-based feature selection approach	■ Colon [64]; ■ CNS (Central Nervous System) [65]; ■ DLBCL [62]; ■ Leukemia [66], [67]; ■ Lung [68]; ■ Prostate [69]; ■ Breast [61];	NB; DT; SVM; KNN	LOOCV
	(Zhu et al., 2010) [70]	Hierarchical Bayesian model-based entropy- D-optimality and A-optimality	■ Leukemia [71]; ■ Multi-tissue [72]; ■ Breast [73]; ■ DLBCL [74]; ■ NCI60 [75]; ■ SRBCT [76];	SVM; NB	k-fold CV (k = 10)
	(Lan and Vucetic, 2011) [51]	Multi-task feature selection filter	■ Multi-tissue [72];	Lasso; SVM; Penalized Logistic Regression	Kruskal-Wallis test (p-value); Accuracy;
	(Mishra and Sahu, 2011) [77]	Signal-to-noise ratio (SNR)	■ Leukemia [67];	SVM; KNN (3-NN)	k-fold CV (k = 10); Holdout validation; LOOCV
	(Chandra and Gupta, 2011) [78]	Effective range based gene selection (ERGS)	■ Leukemia [66], [67]; ■ Colon [64]; ■ DLBCL [74]; ■ Lung [68]; ■ Prostate [69]	NB; SVM	LOOCV
	(Zheng and Kwoh, 2011) [79]	Discrete function learning (DFL) algorithm	■ Leukemia [66], [67]; ■ Ovarian [80] ■ DLBCL [62];	DT C4.5; NB; linear SVM	Complexity analysis; Correctness analysis; k-fold CV (k = 10)
	(Mao and Tang, 2011) [81]	Regularized recursive Mahalanobis separability measure	■ Leukemia [67]; ■ Lung [68]; ■ DLBCL [62]; ■ Prostate [69]; ■ Colon [64];	-	Bootstrap
	(Wei et al., 2012) [82]	Feature score based recursive feature elimination (FS-RFE), Subset level score based recursive feature elimination (SL-RFE)	■ Leukemia [67]	SVM; KNN	Complexity analysis; LOOCV
	(Liu et al., 2013) [83]	Robust principal component analysis (RPCA)	■ Colon [64]	-	Recognition Accuracy (Accs)
	(Rajapakse and Mundra, 2013) [84]	F-score / Kruskal-Wallis score + Pareto fronts analysis (F-PFA) (KW-PFA)	■ Multi-tissue [72], [85]; ■ Lung [86]; ■ Leukemia [66]	Linear SVM	Stability analysis: Kuncheva index; Pearson's correlation coefficient; k-fold CV (k = 4)
	(Hajiloo et al., 2013a) [87]	MeanDiff feature selection	■ Breast	KNN	LOOCV; Classification accuracy; Sensitivity analysis; Specificity analysis; Precision
	(Li and Yin, 2013) [88]	Multiobjective binary biogeography based optimization (MBBBO)	■ Multi-tissue ■ Brain ■ Leukemia ■ Lung ■ SRBCT ■ Prostate ■ DLBCL	SVM	Computational Complexity; LOOCV
	(Mandal and Mukhopadhyay, 2013) [12]	Improved minimum redundancy maximum relevance approach	■ Prostate; ■ Leukemia [67]; ■ Ovarian [80];	-	k-fold CV (k = 10); Sensitivity; Specificity; Accuracy; F-score; AUC;
	Liao et al., (2014) [89]	Locality sensitive Laplacian score (LSLS)	■ Leukemia [66], [67]; ■ Lung [68]; ■ DLBCL [74]; ■ Prostate [69]; ■ SRBCT [76];	SVM	Computational complexity; Accuracy; Precision; Recall; F-score; AUC; LOOCV
	(Maulik and Chakraborty, 2014) [90]	Fuzzy preference based rough set (FPRS)	■ Leukemia [66], [67]; ■ SRBCT [76]; ■ DLBCL [62]; ■ Prostate [69]; ■ Childhood Leukemia	Transductive SVM (TSVM)	k-fold CV (k = 5);

TABLE 3
(*Continued*)

	Literature	Feature Selection	Microarray Data Set	Classifier	Validation Method
Wrapper	(Ai-Jun and Xin-Yuan, 2010) [91]	Generalized singular g-prior with Bayesian stochastic search variable selection (gsg-SSVS)	■ Leukemia [67]; ■ Colon [64];	-	LOOCV
	(Perez et al., 2010) [92]	Population-based incremental learning (PBIL) - evolutionary algorithm	■ Leukemia [67];	-	ANOVA
	(Ji et al., 2011) [93]	Partial least square (PLS)-based global gene selection	■ Leukemia [67]; ■ SRBCT [76];	Linear SVM	k-fold CV (k = 10)
	(Luo et al., 2011) [94]	Improved SVM-recursive cluster elimination (ISVM_RCE)	■ Leukemia; ■ Colon [64]; ■ Brain [95]; ■ DLBCL [74]; ■ Prostate [69]; ■ Pancreas;	Linear SVM, KNN	k-fold CV
	(Li et al., 2011) [96]	Margin influence analysis (MIA)	■ Colon [64]; ■ Breast [97];	SVM	LOOCV
	(Sharma et al., 2012a) [52]	Top-r feature selection called successive FS (SFS) and block reduction	■ SRBCT [76]; ■ Prostate [69]; ■ Leukemia [66];	LDA with nearest centroid classifier (LDA-NCC); Bayes classifier; Nearest neighbor classifier (NNC)	Sensitivity analysis; k-fold CV (k = 3)
	(Sharma et al., 2012b) [98]	Null LDA technique	■ SRBCT [76]; ■ Lung [68]; ■ Leukemia [66], [67];	NNC	Sensitivity analysis; k-fold CV
	Yu et al., (2012) [99]	Sample weighting SVM-RFE	■ Colon [64]; ■ Leukemia [67]; ■ Prostate [69]; ■ Lung [68];	Linear SVM; KNN (1-NN)	Stability analysis: Kuncheva Index, and Normalized Percentage Of Overlapping genes-related (nPOGR); AUC;
	(Liu et al., 2013) [100]	Sample selection method (FS-SSM) - Fuzzy interactive self-organizing data algorithm (ISODATA)	■ Multiple myeloma [101]; ■ Leukemia [67]; ■ Colon [64]; ■ DLBCL [62]; ■ Prostate [69];	Linear SVM; KNN (5-NN)	AUC; Recognition rate
Embedded	(Nie et al., 2010) [53]	Robust feature selection based on $l_{2,1}$-norms	■ Glioma [95]; ■ Lung [86]; ■ Multi-tissue [85], [95]; ■ Prostate [69]; ■ Leukemia [102];	SVM	k-fold CV (k = 5)
	(Cai et al., 2011) [103]	Multi-class $\mathrm{L}_{2,1}$-norm support vector machine	■ Brain [95]; ■ Lung; ■ Leukemia [66]; ■ Multi-tissue[85]; ■ Prostate [69];	SVM, KNN	k-fold CV (k = 5);
	(Maldonado et al., 2011) [104]	Kernel penalized SVM (KP-SVM)	■ Colon [64]; ■ DLBCL [74];	SVM	k-fold CV (k = 10); Standard Deviation
	(Xiang et al., 2012) [54]	Discriminative least square regression (DLSR)	■ Multi-tissue [95]; ■ Glioma [95]; ■ Lung [86]; ■ MLL (Mixed-lineage Leukemia); ■ SRBCT [76]; ■ CLL-SUB-111; ■ GLA-BRA- 180; ■ TOX-171;	KNN (1-NN)	k-fold CV (k = 10); Complexity analysis; Paired Students' t tests
	Pang et al., (2012) [105]	Random survival forests iterative feature elimination	■ Lymphoma + survival data; ■ DLBCL + survival data; ■ Breast + survival data;	-	k-fold CV (k = 10); Computational time
	(Anaissi et al., 2013) [106]	Balanced iterative random forest (BIRF)	■ Childhood Leukemia; ■ NCI60 [75]; ■ Colon [64]; ■ Lung [68];	RF	Out-of-bag (OOB) error rate; AUC; Independent Sub-sample method
	(Du et al., 2013) [55]	Two-stage gene selection (TSGS) with L_2-norm penalty and augmented data technique	■ Arthritis [107];	-	k-fold CV;
	(Liang et al., 2013) [56]	Regularized sparse multinomial logistics regression (SLogReg) with a $L_{1/2}$ penalty	■ Leukemia [67]; ■ Prostate [69]; ■ Colon [64]; ■ DLBCL [62];	KNN (3-NN; 5-NN)	LOOCV

TABLE 3
(*Continued*)

	Literature	Feature Selection	Microarray Data Set	Classifier	Validation Method
Hybrid	(Hu et al., 2010) [13]	Neighborhood mutual information based efficient Minimum-redundancy maximum relevancy (NMI-EmRMR)	■ Breast [108]; ■ DLBCL [74]; ■ Leukemia [66], [67]; ■ Lung [109]; ■ SRBCT [76];	Linear SVM; KNN (5-NN); Classification and Regression Trees (CART)	k-fold CV (k = 10)
	(Saengsiri et al., 2010) [110]	Information gain (INFO) or gain ratio (GR) or correlation based FS (CFS) (filter) > greedy search (GS-SVM) or genetic algorithm (GA-SVM) (wrapper)	■ Colon [64]; ■ DLBCL; ■ Leukemia [67];	Radial SVM	Precision; Recall; F-measure; Accuracy rate
	(Tong and Mintram, 2010) [111]	GA-evaluation > ANN-activation function (GANN)	■ Leukemia [67]; ■ SRBCT [76];	NB; RF; SVM; Classification Tree (J48)	Fitness evaluation; TP rate; FP rate;
	(Shi et al., 2011) [112]	k top scoring pair (k-TSP) (filter) > SVM	■ Breast [61], [113]; ■ Lung [109]; ■ CNS [65];	SVM; KNN	Classification error; LOOCV; K-fold CV (K = 5)
	(Leung and Hung, 2010) [31]	Multiple-filter multiple wrapper (MFMW) feature selection	■ Leukemia [67]; ■ Breast [97]; ■ Colon [64]; ■ DLBCL [62]; ■ Prostate [69]; ■ Lung [68];	Weighted voting (WV); KNN; SVM	LOOCV
	(Mundra and Rajapakse, 2010) [15]	SVM-RFE > minimum redundancy–maximum relevance (mRMR)	■ Colon [64]; ■ Leukemia [67]; ■ Liver [114]; ■ Prostate [69];	Linear SVM	Accuracy; Sensitivity; Specificity; Matthew's Correlation coefficient (MCC)
	(Akadi et al., 2011) [14]	Minimum redundancy–maximum relevance (mRMR) (filter) > genetic algorithm (wrapper)	■ NCI [75]; ■ DLBCL [74]; ■ Lung [115]; ■ Leukemia [67]; ■ Colon [64];	SVM; NB	LOOCV
	(Lee and Leu, 2011) [116]	Between group to within group sum square (BW) ratio > GA with dynamic parameter setting (GADP)	■ Colon [64]; ■ SRBCT [76]; ■ Breast [73]; ■ Leukemia [67]; ■ DLBCL [74]; ■ Multi-tissue [72];	SVM	Prediction accuracy;
	(Liu et al., 2012) [117]	Bhattacharyya distance (filter) > fuzzy interactive self-organizing data algorithm (ISODATA-RFE) (wrapper)	■ Multiple myeloma [101]; ■ Leukemia [67]; ■ Colon [64]; ■ DLBCL [62]; ■ Prostate [69];	SVM; KNN; Hierarchical clustering	Sensitivity analysis
	(Shreem et al., 2012) [16]	ReliefF, mRMR (filter), GA (wrapper) (R-m-GA)	■ CNS [65]; ■ DLBCL [74]; ■ Prostate [69];	Instance-based learner (IB1)	k-fold CV (k = 10)
	(Hajiloo et al., 2013b) [118]	Signal-to-noise (SNR) (filter) > fuzzy support vector machine (FSVM) (wrapper)	■ Leukemia [67]; ■ Colon [64]; ■ Prostate [69]	-	k-fold CV (k = 10)
	(Chang et al., 2013) [119]	ReliefF-GA-ANFIS (ANFIS- Adaptive neuro-fuzzy inference system)	■ Oral cancer prognosis dataset –clinicopathologic data & genomic data;	ANFIS; ANN; SVM; Logistics regression	k-fold CV, AUC
Ensemble	(Abeel et al., 2010) [57]	Ensemble FS: linear SVM+RFE	■ Leukemia [67]; ■ Colon [64];; ■ DLBCL [74]; ■ Prostate [69]	SVM	Stability measure : Kuncheva index; ROC
	(Huawen Liu et al., 2010) [120]	Ensemble gene selection by grouping (EGSG)	■ Breast [61]; ■ CNS [65]; ■ Leukemia [67]; ■ Colon [64]; ■ Prostate [69];	NB; KNN (3-NN)	LOOCV
	(Yang et al., 2010b) [121]	Multiple filter – genetic ensemble based gene selection (MF-GE)	■ Leukemia [66], [67]; ■ Colon [64]; ■ Liver [122];	DT; RF; KNN (3-NN; 7-NN); NB	k-fold CV (k = 3); Mean; Majority Voting
	(Yang and Mao, 2011) [59]	Multicriterion-fusion-based recursive feature elimination (MCF-RFE)	■ Colon [64]; ■ Leukemia [67]; ■ Prostate [69]; ■ CNS [65]; ■ DLBCL [62];	Linear SVM; KNN (3-NN)	Classification error; Standard deviation of error estimation; AUC; Stability measure

TABLE 3
(*Continued*)

Literature	Feature Selection	Microarray Data Set	Classifier	Validation Method
(Ghorai et al., 2011) [17]	mRMR > nonparallel plane proximal classifier (NPPC) ensemble by GA	■ Leukemia [67]; ■ Colon [64]; ■ Lung [68]; ■ Breast [97]; ■ DLBCL [62]; ■ Liver [122]; ■ Prostate [69];	SVM	k-fold CV (k = 10); Majority voting; P-value
(Tan et al., 2011) [58]	Modified two-stage linear SVM- RFE	■ Lung [86];	SVM	k-fold CV (k = 10)
(Gaafar et al., 2012) [18]	Maximum relevance minimum redundancy (mRMR) > ensemble GA	■ Breast [97]; ■ Colon [64]; ■ Leukemia [67]; ■ Lung [68]; ■ DLBCL [62]; ■ Prostate [69];	KNN	LOOCV; Diversity measure: Kohavi-Wolpert Variance (KW); Similarity Analysis
(Song et al., 2013) [123]	Fast clustering-based feature selection (FAST)	14 Microarray dataset, e.g. ■ TOX-171;; ■ Leukemia; ■ CLL-SUB-111; ■ SMK-CAN-187; ■ GLA-BRA-180	Naïve bayes; C4.5; IB1; Inductive rule learner Repeated Incremental Pruning to Produce Error Reduction (IRIPPER)	k-fold CV (k = 10); Runtime; Sensitivity;

logistics regression with $L_{1/2}$ penalty. They measured features with sparsity gap between the high and low weight, and if the sparsity gap is high, the weight could be used for selecting relevant features. This method has significantly reduced the amount of space needed to store the vectors, which usually used to represent the large amounts of data.

As stated previously, most of the hybrid feature selections are designed by combining filter and wrapper methods. Minimal redundancy and maximum relevancy is the most common method used as a part of combination. For example, Hu et al. [13] employed the search strategy of mRMR for constructing neighbourhood mutual information (NMI) for improving the efficiency of mRMR gene selection, Akadi et al. [14] proposed a two-stage gene selection by combining mRMR as filter and genetic algorithm as wrapper, Mundra and Rajapakse [15] incorporated a mutual information based mRMR filter in SVM-RFE to minimize the gene redundancy, and Shreem et al. [16] used ReliefF and mRMR as filters to minimize gene redundancy and GA with classifier to choose the most discriminating genes.

Ensemble method aims to combine the multiple output of experts. For example, Abeel et al. [57] and Tan et al. [58] implemented ensemble feature selection methods by using linear SVM-RFE as the selection mechanism. The method proposed in [57] has improved the biomarker stability and accuracy. Yang and Mao [59] proposed an ensemble method called multi-criterion fusion-based recursive feature elimination (MCF-RFE) where the experimental results showed that the proposed method outperformed SVM-RFE in term of classification accuracy and stability.

3.2 Unsupervised Gene Selection

Unsupervised gene selection approach selects the feature subset unassisted by labeled data. Table 4 shows the detail of this approach and its framework.

Xu et al. [124] proposed an unsupervised gene selection with filter-based evaluation framework by applying diffusion maps to address the multi-dimensionality problem and using the eigenfunctions of Markov matrices as a coordinate system on the original data set in order to obtain efficient representation of data geometric descriptions. The optimal feature subset is then clustered with a neural network and fuzzy ART which learn arbitrary input patterns in a stable, fast and self-organizing way to form a partition of cancer samples. Loscalzo et al. [127] studied the sample size dependency for the stability of feature selection. The authors identified consensus feature groups from subsampling of training samples and performed feature selection by treating each consensus feature group as a single entity.

Chuang et al. [128], Shen et al. [129], and Chuang et al. [130] utilized particle swarm optimization (PSO) to implement their unsupervised feature selection algorithms (the first two works used wrapper framework and the latter used hybrid framework). Chuang et al. [128] improved the binary PSO (IBPSO) to avoid getting trapped in local optima and search for superior classification results in an area with a lower number of genes. Shen et al. [129] identified and removed the redundant genes and samples simultaneously by applying the modified PSO. Chuang et al. [130] proposed a hybrid of Tabu search and binary PSO for feature selection. Filippone et al. [131] proposed a gene selection that makes use simulated annealing as the combinatorial search method and fuzzy C-means as the learning algorithm. Maugis et al. [132] selected relevant features using backward stepwise selection for Gaussian mixture models and an integrated likelihood criterion approximated by the Bayesian information criterion to guide the search for features and to determine the number of clusters.

Witten and Tibshirani [133], and Luss and Aspremont [134] proposed an embedded-based feature selection of sparse clustering. Witten and Tibshirani [133] introduced the application of sparse K-means and sparse hierarchical clustering that use a lasso-type penalty to select the features. Luss and Aspremont [134] studied the method for sparse PCA that seeks sparse factors or linear combinations of data variables and describes the maximum amount of variance in the data while including limited nonzero coefficients.

TABLE 4
Literature Details of Unsupervised Gene Selection

	Literature	Feature Selection	Microarray Dataset	Classifier	Validation Method
Filter	(Ferreira and Figueiredo, 2012) [137]	Algorithm-relevance-redundancy unsupervised feature selection	■ Colon; ■ SRBCT; ■ Lymphoma; ■ Leukemia; ■ DLBCL; ■ Tox-171; ■ Brain; ■ Prostate; ■ Multi-tissue; ■ CLL-SUB-111; ■ SMK-CAN-187; ■ GLI-85;	Linear SVM	Complexity analysis; Similarity analysis; Cumulative relevance measure; k-fold CV (k = 10); Difficulty measure [138]
	(Xu et al., 2010) [124]	Gene selection with correlation coefficient and diffusion map	■ SRBCT [76];	Neural network clustering theory; Fuzzy ART (Adaptive resonance theory) (FA)	Rand index
	(Shen et al., 2009a) [129]	Simultaneous gene and sample selection by modified particle swarm optimization (SSPSO)	■ Bipolar disorder; ■ Gliomas of grades III and IV; ■ Sarcoma;	Linear SVM	k-fold CV (k = 5)
	(Lin and Chien, 2009) [139]	Statistical clustering based on linear relationship and coefficient correlation	Breast cancer cDNA microarray data from Stanford Microarray Database (SMD);	-	-
	(Loscalzo et al., 2009) [127]	Consensus group stable (CGS) feature selection	■ Colon [64]; ■ Leukemia [67]; ■ Lung; ■ Prostate [69]; ■ DLCBL [74]; ■ SRBCT [76];	Linear SVM; KNN (1-NN)	Similarity measure; Stability measure; K-fold CV (K = 10)
	(Chuang et al., 2008) [128]	Improved binary particle swarm optimization (PSO)	■ Multi-tissue; ■ Brain; ■ Leukemia; ■ Lung; ■ SRBCT; ■ Prostate; ■ DLBCL; http://www.gems-system.org/	KNN	LOOCV
Wrapper	(Maugis et al., 2009) [132]	Multivariate Gaussian models and clustering	■ Transcriptome dataset of Arabidopsis thaliana [140];	-	-
	(Filippone et al., 2006) [131]	Simulated annealing with fuzzy C-means	■ Leukemia [67];	-	Representation error
Embedded	(Witten and Tibshirani, 2010) [133]	Sparse K-means; Sparse hierarchical clustering	■ Breast [108]; ■ Single nucleotide polymorphism;	-	Classification error rate
	(Luss and Aspremont, 2010) [134]	Sparse principle component analysis (DSPCA)	■ Colon [64]; ■ DLCBL [74];	K-means	Rand index
	(Niijima and Okuno, 2009) [141]	Laplacian linear discriminant analysis-recursive feature elimination (LLDA-RFE)	■ Colon [64]; ■ Leukemia [66], [67]; ■ Brain [65]; ■ Breast [61]; ■ Lung [109]; ■ SRBCT [76];	Nearest Mean Classifier (NMC)	Classification Error
Hybrid	(Chuang et al., 2009) [130]	Binary particle swarm optimization (BPSO) embedded in Tabu search (TS)	■ Multi-tissue[72], [85]; ■ Brain [65], [95]; ■ Leukemia [66], [67]; ■ Lung [86]; ■ SRBCT [76]; ■ Prostate [69]; ■ DLBCL [62];	KNN (1-NN); SVM	LOOCV; one-versus-rest (SVM-OVR)
	(Boutsidis et al., 2008) [135]	Principle component analysis + column subset selection problem	■ Subject-by-Single nucleotide polymorphism;	-	Classification accuracy
	(Kim and Gao, 2006) [136]	Principle components through least-square estimation (LSE) forward selection + boost expectation maximization (BEM)	■ Leukemia [67];	EM; BEM; K-means	K-fold CV (K = 10)
Ensemble	(Zhang et al. 2012) [142]	Feature ranking based on the consensus matrix (FRCM)	■ Leukemia [66]; ■ Brain [95]; ■ Prostate [143]; ■ Gliomas [144];	K-means	Computational complexity: time complexity, space complexity; Normalized Mutual Information (NMI); Adjusted Ranked Index (ARI)

TABLE 5
Literature Details of Semi-Supervised Gene Selection

	Literature	Feature Selection	Microarray Dataset	Classifier	Validation Method
Filter	(Kalakech et al., 2011) [49]	Pairwise constraint score (Must Link and Cannot Link) with Laplacian score	■ Colon [64]; ■ Leukemia [67];	KNN	Kendall's coefficient; Rank sum; accuracy rate
	(Benabdeslem and Hindawi, 2011) [147]	Constrained Laplacian score (CLS)	■ Colon [64]; ■ Leukemia [67];	KNN (1-NN)	Classification accuracy
	(Benabdeslem and Hindawi, 2013) [146]	Constrained semi-supervised feature selection with redundancy elimination (CSFSR)	■ TOX-171; ■ CLL-SUB-111;	SVM; K-means	Redundancy analysis; Rand index;
Wrapper	(Barkia et al., 2011) [151]	Semi-supervised feature importance evaluation method (SSFI)	■ Colon [152]; ■ Leukemia [67]; ■ Ovarian [153];	-	Confidence measure; Relevance measure;
	(Ren et al., 2008) [154]	Forward semi-supervised feature selection (FW-SemiFS)	■ Colon [64];	NB; Nearest neighbor; KNN	Classification accuracy; Mean; Standard Deviation
Embedded	(Helleputte and Dupont, 2009) [148]	Partial supervised AROM (PS-l2-AROM)	■ Leukemia [67]; ■ DLBCL [62]; ■ Prostate [69]; ■ Colon [64];	Linear SVM	Stability measure: KI; BCR
	(Hindawi and Benabdeslem, 2013) [149]	Local-to-global feature selection (L2GFS) with K-means	■ TOX-171; ■ CLL-SUB-111;	Linear SVM	Rand index; Classification accuracy
Hybrid	(Liu et al., 2006) [150]	Spectral biclustering + cosine measure with SVM	■ DLBCL [74] ■ Liver [122]	SVM	Similarity measure: Cosine measure; Paired t-test (p-value); k-fold CV (k = 10)
Ensemble	(Hindawi et al., 2013a) [155]	Ensemble Laplacian constraint score (EnsCLS)	■ DLCBL [74]; ■ Leukemia [67]; ■ Prostate [69];	-	Classification accuracy
	Yu et al., (2014) [156]	Modified double selection based semi-supervised clustering ensemble (MDS-SSCE)	■ Bladder [157]; ■ Brain [65]; ■ DLCBL [74]; ■ SRBCT [76]; ■ Leukemia [66], [67]; ■ Endometrial [158]; ■ Breast [61]; ■ Multi-tissue [159];	K-means	Means and Standard deviation of Normalized Mutual Information

Boutsidis et al. [135] proposed a hybrid framework by combining the PCA and the Column Subset Selection Problem (CSSP). Kim and Gao [136] firstly retrieved the gene subsets with the original physical meaning based on their capacities to reproduce sample projections on principal components (PCs) by applying the Least-Square-Estimation (LSE) based evaluation. They also applied the boost-expectation-maximization (BEM) clustering to improve the quality of the partitioning.

3.3 Semi-Supervised Gene Selection

Semi-supervised gene selection approach utilizes both labeled and unlabeled data in the process of selecting a feature subset. Table 5 shows the detail of this approach and its framework.

Benabdeslem and Hindawi [146], [147], Helleputte and Dupont [148], and Kalakech et al. [49] proposed semi-supervised feature selection methods based on constraint scores for local properties of unlabeled data. Constraints provide guidance to partition data samples where the similar samples must be grouped together and the dissimilar samples cannot be grouped together. Hindawi and Benabdeslem [149] provided a locally weighting metric model based on constrained K-means clustering in order to perform a global semi-supervised feature selection with filter evaluation framework. Helleputte and Dupont [148] extended the previous works on embedded-based feature selection by adding partial supervision on the dimensions to be selected. This embedded Approximation of zero-norm Minimization (AROM) approach is proposed based on the regularized linear models and it makes use of partial supervision on the features a priori assumed to be more relevant. In addition, Liu et al. [150] introduced a semi-unsupervised gene selection that can find much smaller and informative gene subsets without a priori class information. The authors first used the spectral bi-clustering to obtain the best two classes partitioning eigenvectors to pre-select the genes. Then the best gene combinations among the genes are selected based on the similarity between the genes and the best eigenvectors.

4 Problems Faced in Gene Selection

DNA microarray technology is vastly applied in biomedical research and various studies purposely analyze the gene expression to discover certain diseases or classify disease subtypes, and further predict the responses to therapies or survival times. There are several types of DNA microarrays, e.g., complementary DNA (cDNA), oligonucleotide, bacterial artificial chromosomes (BAC), and single nucleotide polymorphism (SNP) microarrays. There are currently two main trends in microarray technology, cDNA bi-color glass slide and the high-density oligonucleotide array manufactured by Affymetrix GeneChip, and it seems that these techniques are the most commonly used techniques for profiling cancer gene expression data sets.

However, many challenges or problems need to be solved to reveal new knowledge from gene expression data. The curse of dimensionality is always the major concern in microarray analysis; the error in scanning may cause mislabeled problems; and the differences in the numbers of microarray collected have caused class imbalance problems. Moreover, it is difficult to determine the gene relevancy/ redundancy and retrieve useful biological information from the gene expressions. The process of microarray analysis may also lead to unexpected erroneous conclusions and biases. In addition, the different standards of analyzing the results create the problem of cross-platform comparisons. The following discusses these problems concisely.

Curse of dimensionality. Microarray data usually contains large number of gene expressions (up to several hundreds of thousands of features) but with only a limited number of samples (a few dozen of patients). This is a major obstacle in microarray data analysis since high dimensional data always lead to higher risk of over-fitting in many machine learning methods.

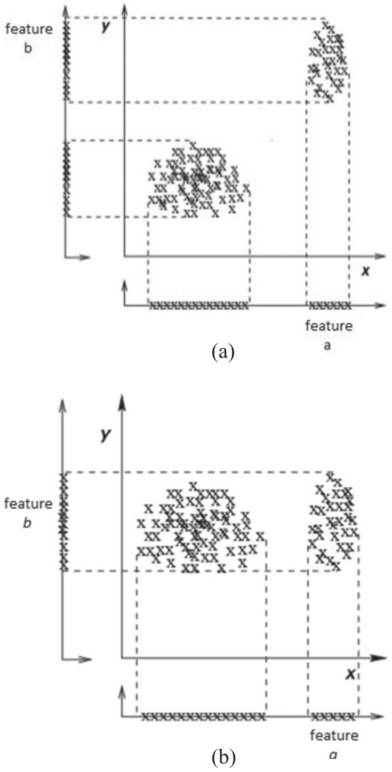
Mislabeled data or imbalanced data issue. Mislabeled data or missing gene expression values due to improper scanning could affect the experimental accuracy and lead to imprecise conclusion about gene expression pattern. Microarray datasets are typically noisy and most of them have class imbalance problems; one class can dominate the data set, e.g., 65 percent ALL versus 35 percent AML in leukemia dataset published in [67]. In the case where the classifier learns from mislabeled or imbalanced data, it may influence the generalization ability of the classifier.

Gene relevancy and redundancy issue. Feature relevancy and redundancy are the main concerns in determining the usefulness or efficacy of a feature or feature subset. Figs. 4a and 4b show the examples of redundant and irrelevant features [160]. As shown, these features may reduce the discriminative power.

Difficulty in biological information retrieval. In microarray studies, substantial gene expression levels are revealed simultaneously in a small fraction of samples. It is significantly important to identify genes that are relevant to the biological phenomenon of interest and to characterize their expression profiles. Currently, most of the microarray analyses only focus on obtaining high classification accuracy results even though actually revealing biological information from gene expression is also important to assist domain experts in designing or planning more appropriate treatments based on specific patient condition [6].

Erroneous and bias problem. The processes of microarray data analysis, including study, experimental design, data accessibility, and platform selection can lead to erroneous conclusions. Technical factors such as differences in physical, batch of reagents used, and various levels of skill of the technicians could possibly cause the biases [6]. The unexpected erroneous and bias raise the difficulties in analyzing microarray data.

Problem of cross-platform comparisons. Cross-platform comparisons of gene expression studies are difficult to perform since microarray data analyses were usually constructed by different standards and the results may not be reproducible. Few researchers [161], [162] have seriously dealt with this problem and conducted more validation tests on the reproducibility, sensitivity, specificity, robustness in the gene expression analysis.



(a)

(b)

Fig. 4 (a) Feature *a* and *b* are redundant, because either of them is sufficient to discriminate the two clusters, (b) Feature *b* is irrelevant, because it does not contribute to cluster discrimination. On the other hand, if feature *a* is omitted, then only one cluster can be recognized.

5 Analysis of Gene Selection Reviewed

As discussed in the previous sections, every gene selection approach has its own pros and cons. Table 6 summarizes the relevant works in supervised, unsupervised and semi-supervised gene selection based on the evaluation framework used.

From the literature study, semi-supervised feature selection seems to be the better approach. This is because in microarray data, there is large supply of unlabeled data but limited number of labeled data since it can be expensive (or even impossible) to determine labels for all data. Semi-supervised feature selection makes use of unsupervised algorithms to find the low dimensional embedding from the unlabeled data and make use of supervised algorithms to learn reasonably accurate classifiers from the labeled data. Semi-supervised feature selection takes advantages of the strengths of both and often can find the most discriminative and informative features [145].

For evaluation framework, hybrid method seems to be better than the other methods. Hybrid method inherits the strength of two methods by using their different evaluation criteria in different search stages. It improves the efficiency by narrowing the search space, thus effectively reduces the computational complexity. Moreover, hybrid method has lower risk of over-fitting and is able to provide higher performance accuracy. Hence, hybrid method will be the most suitable evaluation framework to deal with the curse of dimensionality issue in microarray data.

TABLE 6
A Summary of Relevant Works in Supervised, Unsupervised, and Semi-Supervised Gene Selection

	Supervised	Unsupervised	Semi-supervised
Filter	[50]; [63]; [70]; [51]; [77]; [78]; [79]; [81]; [82]; [83]; [84]; [87]; [89]; [88]; [12]; [90];	[137]; [124]; [129]; [139]; [127]; [128];	[49]; [147]; [146];
Wrapper	[91]; [92]; [93]; [94]; [96]; [52]; [98]; [100]; [99];	[132];[131];	[151]; [154];
Embedded	[53]; [103]; [104]; [54]; [106]; [55]; [56]; [105];	[133]; [134]; [141];	[148]; [149]
Hybrid	[13]; [110]; [111]; [112]; [15]; [31]; [14]; [116]; [117]; [16]; [118]; [119];	[130]; [135]; [136];	[150];
Ensemble	[57]; [120]; [121]; [59]; [17]; [58]; [18];	[142];	[155]; [156];

As shown in the Table 6, only one work combined semi-supervised learning approach with hybrid evaluation method, i.e., [150]. As reported, the author was able to greatly reduce the number of genes needed for producing high prediction accuracy. However, only similarity test and prediction accuracy were measured in their analysis. So, more validation tests on its sensitivity, specificity, reproducibility and robustness are expected for cross-platform comparisons.

5.1 Dataset Analysis

Based on Tables 3, 4, and 5, the five most commonly used gene microarray expression datasets in the literatures are leukemia [67], colon [64], prostate [69], Diffuse Large B-Cell Lymphoma (DLBCL) [74], and Small round blue cell tumor (SRBCT) of childhood datasets [76]. The comparison of validation result using highest prediction accuracy based on CV and the number of selected genes in each literature are shown in Table 7, 8, 9, 10, and 11.

TABLE 7
Highest Prediction Accuracy and the Number of Selected Genes for Leukemia Dataset

	Evaluation method	Literature	CV analysis	Num. of Selected Gene
Supervised	Filter	[63]	100% (LOOCV)	-
		[77]	97.5% (10-fold)	5
		[78]	100% (LOOCV)	80
		[79]	94.1% (10-fold)	-
		[82]	≈ 95% (LOOCV)	20
		[89]	98.61%(LOOCV)	17
		[90]	97.22% (5-fold)	22
		[12]	100% (10-fold)	-
	Wrapper	[91]	97.37%	6
		[93]	≈ 94% (10-fold)	10
		[98]	100%	100
	Embedded	[56]	98.3%(LOOCV)	-
	Hybrid	[118]	98.57% (10-fold)	-
		[116]	100%	5
		[14]	100% (LOOCV)	15
		[31]	100%(LOOCV)	4
		[15]	98.35%	37
	Ensemble	[17]	94.52% (10-fold)	-
		[121]	96.27% (3-fold)	-
		[120]	100%(LOOCV)	30
Unsupervised	Filter	[127]	≈ 100% (10-fold)	30
	Hybrid	[130]	100%(LOOCV)	-
		[136]	90.2% (10-fold)	50

The comparison results indicate that unsupervised and semi-supervised gene selection are capable to produce higher CV accuracy with fewer number of selected genes.

6 Discussion

This paper provides a review on the current and relevant feature selection researches in gene expression microarray analysis. It also discusses the challenges and problems faced in order to achieve better diseases prediction or new diseases discovery. To effectively deal with these problems, the decisions made in each stage of feature selection are important. A plenitude of gene selection approaches have been designed by researchers, yet this paper implies that there are still many open opportunities for further improvement.

In general, we observe that many researchers put huge and fruitful efforts in supervised gene selection approach, and majority of them used filter evaluation framework. However, the comparison results on a few common microarray data sets revealed that unsupervised and semi-supervised approaches are also able to produce good prediction performances with only partially involving some labeled data or even without any labeled data. Thus, the development or

TABLE 8
Highest Prediction Accuracy and the Number of Selected Genes for Colon Dataset

	Evaluation Method	Literature	CV analysis	Num. of Selected Gene
Supervised	Filter	[63]	91.93% (LOOCV)	2
		[78]	83.87% (LOOCV)	100
	Wrapper	[91]	88.71% (LOOCV)	10
		[94]	81.95%	51
		[96]	100% (LOOCV)	100
	Embedded	[106]	96%	19
		[56]	95.1% (LOOCV)	-
		[104]	96.57% (10-fold)	20
	Hybrid	[118]	93.75% (10-fold)	-
		[116]	100%	8
		[14]	98.39% (LOOCV)	15
		[31]	95.16% (LOOCV)	6
		[15]	91.68%	78
	Ensemble	[17]	82.77% (10-fold)	-
		[121]	77.01% (3-fold)	-
		[120]	93.55% (LOOCV)	30
Unsupervised	Filter	[127]	≈ 90% (10-fold)	20

TABLE 9
Highest Prediction Accuracy and the Number of Selected Genes for Prostate Dataset

	Evaluation Method	Literature	CV analysis	Num. of Selected Gene
Supervised	Filter	[63]	98.04% (LOOCV)	-
		[78]	94.12% (LOOCV)	10
		[89]	88.24%(LOOCV)	17
		[90]	91.56% (5-fold)	20
	Wrapper	[52]	100% (3-fold)	4
		[94]	93.10%	97
	Embedded	[56]	95.1% (LOOCV)	-
		[53]	95.09% (5-fold)	20
		[103]	100% (5-fold)	80
	Hybrid	[118]	95.18% (10-fold)	-
		[31]	98.04%(LOOCV)	6
		[15]	98.29%	10
		[16]	100%(10-fold)	-
	Ensemble	[17]	90.16% (10-fold)	-
		[120]	99.02%(LOOCV)	30
Unsupervised	Filter	[127]	≈ 90% (10-fold)	10
	Hybrid	[130]	92.16%(LOOCV)	-

advancement of unsupervised and semi-supervised approaches can be considered as promising future directions in gene selection research.

Another promising direction for gene selection is the development of hybrid and ensemble frameworks to enhance the robustness of the selected feature subsets. Hybrid method is developed by combining two or more evaluation criteria. And ensemble method works by aggregating the results out of the groups. The characteristics of these two methods are specifically more flexible and efficient in dealing with high dimensional data. Unfortunately, there are not many theoretical or empirical works that study the hybrid or ensemble approach in gene expression analysis. Hence, further developments of such approaches are necessary.

Joint analysis of two or more data sets can be another interesting opportunity for future gene selection research. It can be done by involving more than two gene expression data sets in one joint analysis, or perhaps by combining gene expression data set with other prognosis or clinical reports. This promotes the consideration of various aspects and thus enhancing the confidence level. The integration of features and samples selection in joint analysis is certainly a complex and exhausting task but that would be a major breakthrough in feature selection research.

TABLE 10
Highest Prediction Accuracy and the Number of Selected Genes for DLBCL Dataset

	Evaluation Method	Literature	CV analysis	Num. of Selected Gene
Supervised	Filter	[70]	100%	10
		[78]	96.88% (LOOCV)	60
		[89]	94.81%(LOOCV)	18
	Wrapper	[13]	99.9% (10-fold)	9
		[94]	94.97%	97
	Embedded	[56]	94.8% (LOOCV)	-
		[104]	99.73% (10-fold)	8
	Hybrid	[116]	100%	6
		[14]	100% (LOOCV)	15
		[16]	100%(10-fold)	-
Unsupervised	Filter	[127]	≈ 100%(10-fold)	30
Semi-supervised	Hybrid	[150]	99.92% (10-fold)	2

TABLE 11
Highest Prediction Accuracy and the Number of Selected Genes for SRBCT Dataset

	Evaluation Method	Literature	CV analysis	Num. of Selected Gene
Supervised	Filter	[54]	96.47% (10-fold)	80
		[89]	100%(LOOCV)	18
		[90]	95.61% (5-fold)	20
	Wrapper	[13]	84% (10-fold)	9
		[93]	100% (10-fold)	85
		[52]	100% (3-fold)	4
		[98]	100%	500
	Hybrid	[116]	100%	8
Unsupervised	Filter	[127]	≈ 100% (10-fold)	20
	Hybrid	[130]	100% (LOOCV)	-

Most studies treat the highest classification accuracy as the ultimate goal. However, as mentioned above, the lack of ground truth in sample data (due to the potential of mislabeling or misclassifying the samples) limits the basis of judgment regarding the error rate. Thus, more research efforts in evaluation and validation of feature selection, like measurement of specificity, sensitivity, similarity, and stability of signature, should be devoted.

ACKNOWLEDGMENTS

The authors would like to thank editors, reviewers, and Dr. Lim S.P. for the valuable comments. Ang Jun Chin was supported by The Universiti Teknologi Malaysia (UTM) under Zamalah Scholarship. Andri Mirzal was supported by Dhofar University under Seed Grant 2015 program.

REFERENCES

[1] K. Yang, Z. Cai, J. Li, and G. Lin, "A stable gene selection in microarray data analysis," *BMC Bioinf.*, vol. 7, no. 1, p. 228, Apr. 2006.

[2] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer, "Evaluating microarray-based classifiers: An Overview," *Cancer Informat.*, vol. 6, pp. 77–97, Feb. 2008.

[3] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," *J. Mach. Learn. Res.-Proc. Track*, vol. 10, pp. 4–13, 2010.

[4] F. K. Ahmade, N. M. Norwawi, S. Deris, and N. H. Othman, "A review of feature selection techniques via gene expression profiles," in *Proc. Int. Symp. Inf. Technol.*, 2008, vol. 2, pp. 1–7.

[5] S. Alelyani, J. Tang, and H. Liu, "Feature selection for clustering: A review," in *Data Clustering: Algorithms and Applications*, vol. 29. Boca Raton, FL, USA: CRC Press, 2013.

[6] G. V. S. George and V. C. Raj, "Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile," *Int. J. Comput. Sci. Eng. Surv.*, vol. 2, no. 3, pp. 16–27, Aug. 2011.

[7] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," 2013.

[8] V. Bolon-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, 2014.

[9] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, Dec. 2004.

[10] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognit.*, vol. 43, no. 1, pp. 5–13, Jan. 2010.

[11] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 03, no. 02, pp. 185–205, Apr. 2005.

[12] M. Mandal and A. Mukhopadhyay, "An improved minimum redundancy maximum relevance approach for feature selection in gene expression data," *Procedia Technol.*, vol. 10, pp. 20–27, 2013.

[13] Q. Hu, W. Pan, S. An, P. Ma, and J. Wei, "An efficient gene selection technique for cancer recognition based on neighborhood mutual information," *Int. J. Mach. Learn. Cybern.*, vol. 1, no. 1–4, pp. 63–74, Dec. 2010.

[14] A. E. Akadi, A. Amine, A. E. Ouardighi, and D. Aboutajdine, "A two-stage gene selection scheme utilizing MRMR filter and GA wrapper," *Knowl. Inf. Syst.*, vol. 26, no. 3, pp. 487–500, Mar. 2011.

[15] P. A. Mundra and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection," *IEEE Trans. NanoBiosci.*, vol. 9, no. 1, pp. 31–37, Mar. 2010.

[16] S. S. Shreem, S. Abdullah, M. Z. A. Nazri, and M. Alzaqebah, "Hybridizing ReliefF, mRMR filters and GA wrapper approaches for gene selection," *J. Theor. Appl. Inf. Technol.*, vol. 46, no. 2, 2012.

[17] S. Ghorai, A. Mukherjee, S. Sengupta, and P. K. Dutta, "Cancer classification from gene expression data by NPPC ensemble," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 3, pp. 659–671, May-June 2011.

[18] M. A. Gaafar, N. A. Yousri, and M. A. Ismail, "A novel ensemble selection method for cancer diagnosis using microarray datasets," in *Proc. IEEE 12th Int. Conf. BioInformat. BioEng.*, 2012, pp. 368–373.

[19] M. Gutkin, R. Shamir, and G. Dror, "SlimPLS: A method for feature selection in gene expression-based disease classification," *PLoS ONE*, vol. 4, no. 7, p. e6416, Jul. 2009.

[20] R. Varshavsky, A. Gottlieb, M. Linial, and D. Horn, "Novel unsupervised feature filtering of biological data," *Bioinf.*, vol. 22, no. 14, pp. e507–e513, Jul. 2006.

[21] P. Krızek, "Feature selection: Stability, algorithms, and evaluation," Ph.D. dissertation, Dept. Cybern., Faculty of Elect. Eng., Czech Technical Univ., Praha, Czech Republic, 2008.

[22] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artif. Intell.*, vol. 97, no. 1–2, pp. 245–271, Dec. 1997.

[23] J. Doak, *An Evaluation of Feature Selection Methods and Their Application to Computer Security*. Santa Cruz, CA, USA: Comput. Sci., Univ. California, 1992.

[24] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Englewood Cliffs, NJ, USA: Prentice-Hall, 1982.

[25] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett.*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994.

[26] F. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large scale feature selection," in *Pattern Recognition in Practice IV*, E. S. Gelsema, L. N. Kanal Eds. Amsterdam, The Netherlands: North Holland, 1994, pp. 403–413.

[27] D. B. Skalak, "Prototype and feature selection by sampling and random mutation hill climbing algorithms," in *Proc. 11th Int. Conf. Mach. Learn.*, 1994, pp. 293–301.

[28] D. E. Goldberg et al., *Genetic Algorithms in Search, Optimization, and Machine Learning*, vol. 412. Reading, MA, USA: Addison-Wesley, 1989.

[29] G. Brassard and P. Bratley, *Fundamentals of Algorithmics*. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[30] F. Glover, "Tabu Search—Part I," *ORSA J. Comput.*, vol. 1, no. 3, pp. 190–206, Aug. 1989.

[31] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 1, pp. 108–117, Jan.-Feb. 2010.

[32] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1106–1119, July-Aug. 2012.

[33] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, Jan. 1997.

[34] Y. Peng, Z. Wu, and J. Jiang, "A novel feature selection approach for biomedical data classification," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 15–23, Feb. 2010.

[35] Y. Saeys, "Feature selection for classification of nucleic acid sequences," Ph.D. dissertation, Ghent Univ., Ghent, Belgium, 2004.

[36] M. Monirul Kabir, M. Monirul Islam, and K. Murase, "A new wrapper feature selection approach using neural network," *Neurocomputing*, vol. 73, nos. 16–18, pp. 3273–3283, Oct. 2010.

[37] Q. Shen, R. Diao, and P. Su, "Feature selection ensemble," in *Proc. Turing-100*, 2012, vol. 10, pp. 289–306.

[38] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *Proc. IEEE 13th Int. Conf. Inf. Reuse Integr.*, 2012, pp. 356–363.

[39] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J. Am. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, Dec. 1971.

[40] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytol.*, vol. 11, no. 2, pp. 37–50, 1912.

[41] L. I. Kuncheva, "A stability index for feature selection," in *Proc. 25th Conf Proc. 25th IASTED Int. Multi-Conf.: Artificial Intell. Appl.*, Anaheim, CA, USA, 2007, pp. 390–395.

[42] M. Muselli, A. Bertoni, M. Frasca, A. Beghini, F. Ruffino, and G. Valentini, "A mathematical model for the validation of gene selection methods," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 5, pp. 1385–1392, Sep.-Oct. 2011.

[43] U. Braga-Neto, R. Hashimoto, E. R. Dougherty, D. V. Nguyen, and R. J. Carroll, "Is cross-validation better than resubstitution for ranking genes?" *Bioinf.*, vol. 20, no. 2, pp. 253–258, Jan. 2004.

[44] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manage.*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[45] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, "Understanding and using sensitivity, specificity and predictive values," *Indian J. Ophthalmol.*, vol. 56, no. 1, pp. 45–50, 2008.

[46] A. Yousefpour, R. Ibrahim, H. N. A. Hamed, and M. S. Hajmohammadi, "Feature reduction using standard deviation with different subsets selection in sentiment analysis," in *Proc. 6th Asian Conf. Intell. Inf. Database Syst.*, 2014, pp. 33–41.

[47] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," *Mach. Learn.*, vol. 31, pp. 1–38, 2004.

[48] T. Bo and I. Jonassen, "New feature subset selection procedures for classification of expression profiles," *Genome Biol.*, vol. 3, no. 4, pp. 1–0017, 2002.

[49] M. Kalakech, P. Biela, L. Macaire, and D. Hamad, "Constraint scores for semi-supervised feature selection: A comparative study," *Pattern Recognit. Lett.*, vol. 32, no. 5, pp. 656–665, Apr. 2011.

[50] Y. Sun, S. Todorovic, and S. Goodison, "Local-learning-based feature selection for high-dimensional data analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1610–1626, Sep. 2010.

[51] L. Lan and S. Vucetic, "Improving accuracy of microarray classification by a simple multi-task feature selection filter," *Int. J. Data Mining Bioinf.*, vol. 5, no. 2, pp. 189–208, Jan. 2011.

[52] A. Sharma, S. Imoto, and S. Miyano, "A Top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 3, pp. 754–764, May-June 2012.

[53] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst. 23*, 2010, pp. 1813–1821.

[54] S. Xiang, F. Nie, G. Meng, C. Pan, and C. Zhang, "Discriminative least squares regression for multiclass classification and feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 11, pp. 1738–1754, Nov. 2012.

[55] D. Du, K. Li, and J. Deng, "An efficient two-stage gene selection method for microarray data," in *Proc. 2nd Int. Conf. Intell. Comput. Sustainable Energy Environ.*, 2013, pp. 424–432.

[56] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang, "Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification," *BMC Bioinf.*, vol. 14, no. 1, p. 198, Jun. 2013.

[57] T. Abeel, T. Helleputte, Y. V. de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinf.*, vol. 26, no. 3, pp. 392–398, Feb. 2010.

[58] P. L. Tan, S. C. Tan, C. P. Lim, and S. E. Khor, "A modified two-stage SVM-RFE model for cancer classification using microarray data," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 668–675.

[59] F. Yang and K. Z. Mao, "Robust feature selection for microarray data based on multicriterion fusion," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 4, pp. 1080–1092, July-Aug. 2011.

[60] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald, "Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy," *Cancer*, vol. 104, no. 2, pp. 290–298, Jul. 2005.

[61] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[62] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nat. Med.*, vol. 8, no. 1, pp. 68–74, Jan. 2002.

[63] X. Wang and O. Gotoh, "A robust gene selection method for microarray-based cancer classification," *Cancer Informat.*, vol. 9, pp. 15–30, Feb. 2010.

[64] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, Jun. 1999.

[65] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, Jan. 2002.

[66] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nat. Genet.*, vol. 30, no. 1, pp. 41–47, Jan. 2002.

[67] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[68] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res.*, vol. 62, no. 17, pp. 4963–4967, Sep. 2002.

[69] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, Mar. 2002.

[70] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature selection for gene expression using model-based entropy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 7, no. 1, pp. 25–36, Jan.-Mar. 2010.

[71] E.-J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, Mar. 2002.

[72] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 26, pp. 15149–15154, Dec. 2001.

[73] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, M. Raffeld, Z. Yakhini, A. Ben-Dor, E. Dougherty, J. Kononen, L. Bubendorf, W. Fehrle, S. Pittaluga, S. Gruvberger, N. Loman, O. Johannsson, H. Olsson, B. Wilfond, G. Sauter, O.-P. Kallioniemi, Å. Borg, and J. Trent, "Gene-expression profiles in hereditary breast cancer," *N. Engl. J. Med.*, vol. 344, no. 8, pp. 539–548, 2001.

[74] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, Feb. 2000.

[75] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. C. F. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein, and P. O. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nat. Genet.*, vol. 24, no. 3, pp. 227–235, Mar. 2000.

[76] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nat. Med.*, vol. 7, no. 6, pp. 673–679, Jun. 2001.

[77] D. Mishra and B. Sahu, "Feature selection for cancer classification: a signal-to-noise ratio approach," *Int. J. Sci. Eng. Res.*, vol. 2, no. 4, pp. 1–7, 2011.

[78] B. Chandra and M. Gupta, "An efficient statistical feature selection approach for classification of gene expression data," *J. Biomed. Inform.*, vol. 44, no. 4, pp. 529–535, Aug. 2011.

[79] Y. Zheng and C. K. Kwoh, "A feature subset selection method based on high-dimensional mutual information," *Entropy*, vol. 13, no. 4, pp. 860–901, Apr. 2011.

[80] 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, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002.

[81] K. Z. Mao and W. Tang, "Recursive mahalanobis separability measure for gene subset selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 1, pp. 266–272, Jan.-Feb. 2011.

[82] D. Wei, S. Li, and M. Tan, "Graph embedding based feature selection," *Neurocomputing*, vol. 93, pp. 115–125, Sep. 2012.

[83] J.-X. Liu, Y.-T. Wang, C.-H. Zheng, W. Sha, J.-X. Mi, and Y. Xu, "Robust PCA based method for discovering differentially expressed genes," *BMC Bioinf.*, vol. 14, no. Suppl. 8, p. S3, May 2013.

[84] J. C. Rajapakse and P. A. Mundra, "Multiclass gene selection using pareto-fronts," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 1, pp. 87–97, Jan.-Feb. 2013.

[85] A. I. Su, J. B. Welsh, L. M. Sapinoso, S. G. Kern, P. Dimitrov, H. Lapp, P. G. Schultz, S. M. Powell, C. A. Moskaluk, H. F. Frierson, and G. M. Hampton, "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Res.*, vol. 61, no. 20, pp. 7388–7393, Oct. 2001.

[86] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13790–13795, Nov. 2001.

[87] M. Hajiloo, B. Damavandi, M. HooshSadat, F. Sangi, J. R. Mackey, C. E. Cass, R. Greiner, and S. Damaraju, "Breast cancer prediction using genome wide single nucleotide polymorphism data," *BMC Bioinf.*, vol. 14, no. Suppl. 13, p. S3, Oct. 2013.

[88] X. Li and M. Yin, "Multiobjective binary biogeography based optimization for feature selection using gene expression data," *IEEE Trans. NanoBiosci.*, vol. 12, no. 4, pp. 343–353, Dec. 2013.

[89] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, and Z. Cao, "Gene selection using locality sensitive laplacian score," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 6, pp. 1146–1156, Nov.-Dec. 2014.

[90] U. Maulik and D. Chakraborty, "Fuzzy preference based feature selection and semisupervised SVM for cancer classification," *IEEE Trans. NanoBiosci.*, vol. 13, no. 2, pp. 152–160, Jun. 2014.

[91] Y. Ai-Jun and S. Xin-Yuan, "Bayesian variable selection for disease classification using gene expression data," *Bioinf.*, vol. 26, no. 2, pp. 215–222, Jan. 2010.

[92] M. Perez, D. M. Rubin, T. Marwala, L. E. Scott, and W. Stevens, "A Population-Based Incremental Learning approach to microarray gene expression feature selection," in *Proc. IEEE 26th Conv. Electr. Electron. Eng. Israel*, 2010, pp. 10–14.

[93] G. Ji, Z. Yang, and W. You, "PLS-based gene selection and identification of tumor-specific genes," *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 41, no. 6, pp. 830–841, Nov. 2011.

[94] L.-K. Luo, D.-F. Huang, L.-J. Ye, Q.-F. Zhou, G.-F. Shao, and H. Peng, "Improving the computational efficiency of recursive cluster elimination for gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 1, pp. 122–129, Jan.-Feb. 2011.

[95] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, P. M. Black, A. von Deimling, S. L. Pomeroy, T. R. Golub, and D. N. Louis, "Gene Expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Res.*, vol. 63, no. 7, pp. 1602–1607, Apr. 2003.

[96] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, D.-S. Cao, B.-B. Tan, B.-C. Deng, and C.-C. Lin, "Recipe for uncovering predictive genes using support vector machines based on model population analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 6, pp. 1633–1641, Nov.-Dec. 2011.

[97] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 20, pp. 11462–11467, Sep. 2001.

[98] A. Sharma, S. Imoto, S. Miyano, and V. Sharma, "Null space based feature selection method for gene expression data," *Int. J. Mach. Learn. Cybern.*, vol. 3, no. 4, pp. 269–276, Dec. 2012.

[99] L. Yu, Y. Han, and M. E. Berens, "Stable gene selection from microarray data via sample weighting," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 1, pp. 262–272, Jan.-Feb. 2012.

[100] Q. Liu, Z. Zhao, Y. Li, X. Yu, and Y. Wang, "A novel method of feature selection based on SVM," *J. Comput.*, vol. 8, no. 8, pp. 2144–2149, Aug. 2013.

[101] F. Zhan, J. Hardin, B. Kordsmeier, K. Bumm, M. Zheng, E. Tian, R. Sanderson, Y. Yang, C. Wilson, M. Zangari, E. Anaissie, C. Morris, F. Muwalla, F. van Rhee, A. Fassas, J. Crowley, G. Tricot, B. Barlogie, and J. Shaughnessy, "Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells," *Blood*, vol. 99, no. 5, pp. 1745–1757, Mar. 2002.

[102] S. P. A. Fodor, "DNA SEQUENCING: Massively parallel genomics," *Science*, vol. 277, no. 5324, pp. 393–395, Jul. 1997.

[103] X. Cai, F. Nie, H. Huang, and C. Ding, "Multi-class L2,1-norm support vector machine," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 91–100.

[104] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Inf. Sci.*, vol. 181, no. 1, pp. 115–128, Jan. 2011.

[105] H. Pang, S. L. George, K. Hui, and T. Tong, "Gene selection using iterative feature elimination random forests for survival outcomes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 5, pp. 1422–1431, Sep.-Oct. 2012.

[106] A. Anaissi, P. J. Kennedy, M. Goyal, and D. R. Catchpoole, "A balanced iterative random forest for gene selection from microarray data," *BMC Bioinf.*, vol. 14, no. 1, p. 261, Aug. 2013.

[107] N. Sha, M. Vannucci, P. J. Brown, M. K. Trower, G. Amphlett, and F. Falciani, "Gene selection in arthritis classification with large-scale microarray expression profiles," *Comput. Funct. Genomics*, vol. 4, no. 2, pp. 171–181, Apr. 2003.

[108] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000.

[109] D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nat. Med.*, vol. 8, no. 8, pp. 816–824, Aug. 2002.

[110] P. Saengsiri, S. N. Wichian, P. Meesad, and U. Herwig, "Comparison of hybrid feature selection models on gene expression data," in *Proc. 8th Int. Conf. ICT Knowl. Eng.*, 2010, pp. 13–18.

[111] D. L. Tong and R. Mintram, "Genetic algorithm-neural network (GANN): A study of neural network activation functions and depth of genetic algorithm search applied to feature selection," *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 75–87, Dec. 2010.

[112] P. Shi, S. Ray, Q. Zhu, and M. A. Kon, "Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction," *BMC Bioinf.*, vol. 12, no. 1, p. 375, Sep. 2011.

[113] 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, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, no. 9460, pp. 671–679, Feb. 2005.

[114] N. Iizuka, M. Oka, H. Yamada-Okabe, M. Nishida, Y. Maeda, N. Mori, T. Takao, T. Tamesa, A. Tangoku, H. Tabuchi, K. Hamada, H. Nakayama, H. Ishitsuka, T. Miyamoto, A. Hirabayashi, S. Uchimura, and Y. Hamamoto, "Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection," *Lancet*, vol. 361, no. 9361, pp. 923–929, Mar. 2003.

[115] M. E. Garber, O. G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G. D. Rosen, C. M. Perou, R. I. Whyte, R. B. Altman, P. O. Brown, D. Botstein, and I. Petersen, "Diversity of gene expression in adenocarcinoma of the lung," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13784–13789, Nov. 2001.

[116] C.-P. Lee and Y. Leu, "A novel hybrid feature selection method for microarray data analysis," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 208–213, Jan. 2011.

[117] Q. Liu, Z. Zhao, Y.-X. Li, and Y. Li, "Feature selection based on sensitivity analysis of fuzzy ISODATA," *Neurocomputing*, vol. 85, pp. 29–37, May 2012.

[118] M. Hajiloo, H. R. Rabiee, and M. Anooshahpour, "Fuzzy support vector machine: an efficient rule-based classification technique for microarrays," *BMC Bioinf.*, vol. 14, no. Suppl 13, p. S4, Oct. 2013.

[119] S.-W. Chang, S. Abdul-Kareem, A. F. Merican, and R. B. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinf.*, vol. 14, no. 1, p. 170, May 2013.

[120] H. Liu, L. Liu, and H. Zhang, "Ensemble gene selection by grouping for microarray data classification," *J. Biomed. Inform.*, vol. 43, no. 1, pp. 81–87, Feb. 2010.

[121] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data," *BMC Bioinf.*, vol. 11, no. Suppl. 1, p. S5, Jan. 2010.

[122] X. Chen, S. T. Cheung, S. So, S. T. Fan, C. Barry, J. Higgins, K.-M. Lai, J. Ji, S. Dudoit, I. O. L. Ng, M. van de Rijn, D. Botstein, and P. O. Brown, "Gene expression patterns in human liver cancers," *Mol. Biol. Cell*, vol. 13, no. 6, pp. 1929–1939, Jun. 2002.

[123] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.

[124] R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch II, "Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps," *Artif. Intell. Med.*, vol. 48, nos. 2/3, pp. 91–98, Feb. 2010.

[125] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2010, pp. 333–342.

[126] H. Liu, E. R. Dougherty, J. G. Dy, K. Torkkola, E. Tuv, H. Peng, C. Ding, F. Long, M. Berens, L. Parsons, Z. Zhao, L. Yu, and G. Forman, "Evolving feature selection," *IEEE Intell. Syst.*, vol. 20, no. 6, pp. 64–76, Nov. 2005.

[127] S. Loscalzo, L. Yu, and C. Ding, “Consensus group stable feature selection,” in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2009, pp. 567–576.

[128] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, “Improved binary PSO for feature selection using gene expression data,” *Comput. Biol. Chem.*, vol. 32, no. 1, pp. 29–38, Feb. 2008.

[129] Q. Shen, Z. Mei, and B.-X. Ye, “Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification,” *Comput. Biol. Med.*, vol. 39, no. 7, pp. 646–649, Jul. 2009.

[130] L.-Y. Chuang, C.-H. Yang, and C.-H. Yang, “Tabu search and binary particle swarm optimization for feature selection using microarray data,” *J. Comput. Biol.*, vol. 16, no. 12, pp. 1689–1703, 2009.

[131] M. Filippone, F. Masulli, and S. Rovetta, “Unsupervised gene selection and clustering using simulated annealing,” in *Proc. 6th Int. Workshop Fuzzy Logic Appl.*, 2006, pp. 229–235.

[132] C. Maugis, G. Celeux, and M.-L. Martin-Magniette, “Variable selection for clustering with Gaussian mixture models,” *Biometrics*, vol. 65, no. 3, pp. 701–709, Sep. 2009.

[133] D. M. Witten and R. Tibshirani, “A framework for feature selection in clustering,” *J. Am. Stat. Assoc.*, vol. 105, no. 490, pp. 713–726, Jun. 2010.

[134] R. Luss and A. d’ Aspremont, “Clustering and feature selection using sparse principal component analysis,” *Optim. Eng.*, vol. 11, no. 1, pp. 145–157, Feb. 2010.

[135] C. Boutsidis, M. W. Mahoney, and P. Drineas, “Unsupervised feature selection for principal components analysis,” in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2008, pp. 61–69.

[136] Y. B. Kim and J. Gao, “Unsupervised gene selection for high dimensional data,” in *Proc. 6th IEEE Symp. BioInformat. BioEng.*, 2006, pp. 227–234.

[137] A. J. Ferreira and M. A. T. Figueiredo, “Efficient feature selection filters for high-dimensional data,” *Pattern Recognit. Lett.*, vol. 33, no. 13, pp. 1794–1804, Oct. 2012.

[138] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, “Conditional likelihood maximisation: A unifying framework for information theoretic feature selection,” *J. Mach. Learn. Res.*, vol. 13, pp. 27–66, Jan. 2012.

[139] K.-S. Lin and C.-F. Chien, “Cluster analysis of genome-wide expression data for feature extraction,” *Expert Syst. Appl.*, vol. 36, no. 2, Part 2, pp. 3327–3335, Mar. 2009.

[140] S. Gagnot, J.-P. Tamby, M.-L. Martin-Magniette, F. Bitton, L. Taconnat, S. Balzergue, S. Aubourg, J.-P. Renou, A. Lecharny, and V. Brunaud, “CATdb: A public access to Arabidopsis transcriptome data from the URGV-CATMA platform,” *Nucleic Acids Res.*, vol. 36, no. suppl. 1, pp. D986–D990, Jan. 2008.

[141] S. Niijima and Y. Okuno, “Laplacian linear discriminant analysis approach to unsupervised feature selection,” *IEEE/ACM Trans. Comput. Biol. Bioinf.* , vol. 6, no. 4, pp. 605–614, Oct.-Dec. 2009.

[142] S. Zhang, H.-S. Wong, Y. Shen, and D. Xie, “A new unsupervised feature ranking method for gene expression data based on consensus affinity,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1257–1263, July-Aug. 2012.

[143] S. A. Tomlins, R. Mehra, D. R. Rhodes, X. Cao, L. Wang, S. M. Dhanasekaran, S. Kalyana-Sundaram, J. T. Wei, M. A. Rubin, K. J. Pienta, R. B. Shah, and A. M. Chinnaiyan, “Integrative molecular concept modeling of prostate cancer progression,” *Nat. Genet.*, vol. 39, no. 1, pp. 41–51, Jan. 2007.

[144] M. Bredel, C. Bredel, D. Juric, G. R. Harsh, H. Vogel, L. D. Recht, and B. I. Sikic, “Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas,” *Cancer Res.*, vol. 65, no. 19, pp. 8679–8689, Oct. 2005.

[145] J. Zhao, K. Lu, and X. He, “Locality sensitive semi-supervised feature selection,” *Neurocomputing*, vol. 71, nos. 10–12, pp. 1842–1849, Jun. 2008.

[146] K. Benabdeslem and M. Hindawi, “Efficient semi-supervised feature selection: Constraint, relevance and redundancy,” *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1131–1143, May 2013.

[147] K. Benabdeslem and M. Hindawi, “Constrained laplacian score for semi-supervised feature selection,” in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2011, pp. 204–218.

[148] T. Helleputte and P. Dupont, “Partially supervised feature selection with regularized linear models,” in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, New York, NY, USA, 2009, pp. 409–416.

[149] M. Hindawi and K. Benabdeslem, “Local-to-global semi-supervised feature selection,” in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage.*, New York, NY, USA, 2013, pp. 2159–2168.

[150] B. Liu, C. Wan, and L. Wang, “An efficient semi-unsupervised gene selection method via spectral biclustering,” *IEEE Trans. NanoBiosci.*, vol. 5, no. 2, pp. 110–114, Jun. 2006.

[151] H. Barkia, H. Elghazel, and A. Aussem, “Semi-supervised feature importance evaluation with ensemble learning,” in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 31–40.

[152] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, “Tissue Classification with Gene Expression Profiles,” *J. Comput. Biol.*, vol. 7, nos. 3/4, pp. 559–583, Aug. 2000.

[153] M. Schummer, W. V. Ng, R. E. Bumgarner, P. S. Nelson, B. Schummer, D. W. Bednarski, L. Hassell, R. L. Baldwin, B. Y. Karlan, and L. Hood, “Comparative hybridization of an array of 21 500 ovarian cDNAs for the discovery of genes overexpressed in ovarian carcinomas,” *Gene*, vol. 238, no. 2, pp. 375–385, Oct. 1999.

[154] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P. S. Yu, “Forward semi-supervised feature selection,” in *Proc. 12th Asia-Pacific Conf. Adv. Knowl. Discovery Data Mining*, 2008, pp. 970–976.

[155] M. Hindawi, H. Elghazel, and K. Benabdeslem, “Efficient semi-supervised feature selection by an ensemble approach,” in *Proc. Int. Workshop Complex Mach. Learn. Problems Ensemble Methods*, 2013, pp. 41–55.

[156] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, and G. Han, “Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles,” *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 4, pp. 727–740, July-Aug. 2014.

[157] L. Dyrskjot, T. Thykjaer, M. Kruhoffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Orntoft, “Identifying distinct classes of bladder carcinoma using microarrays,” *Nat. Genet.*, vol. 33, no. 1, pp. 90–96, Jan. 2003.

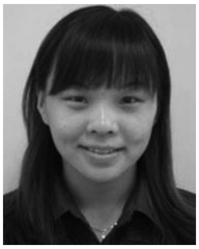
[158] J. I. Risinger, G. L. Maxwell, G. V. R. Chandramouli, A. Jazaeri, O. Aprelikova, T. Patterson, A. Berchuck, and J. C. Barrett, “Microarray analysis reveals distinct gene expression profiles among different histologic types of endometrial cancer,” *Cancer Res.*, vol. 63, no. 1, pp. 6–11, Jan. 2003.

[159] Y. Hoshida, J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Subclass mapping: Identifying common subtypes in independent disease data sets,” *PLoS ONE*, vol. 2, no. 11, p. e1195, Nov. 2007.

[160] J. G. Dy and C. E. Brodley, “Feature selection for unsupervised learning,” *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, Dec. 2004.

[161] J. Garcia-Nieto, E. Alba, L. Jourdan, and E. Talbi, “Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis,” *Inf. Process. Lett.*, vol. 109, no. 16, pp. 887–896, Jul. 2009.

[162] Y. Liu, “Wavelet feature extraction for high-dimensional microarray data,” *Neurocomputing*, vol. 72, nos. 4–6, pp. 985–990, Jan. 2009.



Jun Chin Ang received the BSc and MSc degrees in computer science from the Universiti Teknologi Malaysia (UTM). She is currently working toward the PhD degree in UTM. Her research interests include cancer classification and clustering, gene expression pattern analysis, application intelligence, knowledge discovery, and data mining.

Andri Mirzal received the BEng degree in electrical engineering from the Institut Teknologi Bandung, and the MSc and PhD degrees in information science and technology from Hokkaido University. His research interests include machine learning, bioinformatics, optimization methods, web search engine, and linear inverse problems. From 2011 to 2014, he was with the Faculty of Computing, UTM. He is currently an associate professor in the College of Graduate Studies, Arabian Gulf University.

Habibollah Haron received the diploma and bachelor degrees in computer science from the Universiti Teknologi Malaysia (UTM) in 1987 and 1989, respectively, the master's of science (computer technology in manufacture) degree from the University of Sussex, East Sussex, United Kingdom, in 1995, and the PhD degree in computer-aided geometric design from UTM in 2004. He is currently a professor of soft computing techniques with the Faculty of Computing, UTM. His current research interests include soft computing (SC) techniques in prediction, optimization, and planning. He is a senior member of the IEEE.

Haza Nuzly Abdull Hamed received the bachelor's degrees in information technology from Universiti Utara Malaysia, the master's degree in computer science from the Universiti Teknologi Malaysia (UTM), and the PhD degree from Knowledge Engineering and Discovery Research Institute, Auckland University of Technology. He is currently a senior lecturer at the Faculty of Computing, UTM. His current research interests include integrative connectionist environment for spatial and spatiotemporal data processing using computational intelligence methods with evolutionary algorithm optimization.