

RESEARCH

Open Access

# Stable bagging feature selection on medical data



Salem Alelyani\* 

\*Correspondence:  
s.alelyani@kku.edu.sa  
Department of Computer  
Science, Center for Artificial  
Intelligence, King Khalid  
University, P.O. Box 9004,  
Abha 61413, Saudi Arabia

## Abstract

In the medical field, distinguishing genes that are relevant to a specific disease, let's say colon cancer, is crucial to finding a cure and understanding its causes and subsequent complications. Usually, medical datasets are comprised of immensely complex dimensions with considerably small sample size. Thus, for domain experts, such as biologists, the task of identifying these genes have become a very challenging one, to say the least. Feature selection is a technique that aims to select these genes, or features in machine learning field with respect to the disease. However, learning from a medical dataset to identify relevant features suffers from the *curse-of-dimensionality*. Due to a large number of features with a small sample size, the selection usually returns a different subset each time a new sample is introduced into the dataset. This selection instability is intrinsically related to data variance. We assume that reducing data variance improves selection stability. In this paper, we propose an ensemble approach based on the bagging technique to improve feature selection stability in medical datasets via data variance reduction. We conducted an experiment using four microarray datasets each of which suffers from high dimensionality and relatively small sample size. On each dataset, we applied five well-known feature selection algorithms to select varying number of features. The proposed technique shows a significant improvement in selection stability while at least maintaining the classification accuracy. The stability improvement ranges from 20 to 50 percent in all cases. This implies that the likelihood of selecting the same features increased 20 to 50 percent more. This is accompanied with the increase of classification accuracy in most cases, which signifies the stated results of stability.

**Keywords:** Feature selection, Ensemble technique, Bagging, Dimensionality reduction, Medical data, Microarray, Variance, Bias

## Introduction

In the growth of data mining and collection technologies, data learning and understanding are a tedious task due to a large number of features present that are known as variables or attributes. Usually, data harvesting is conducted in relation to a specific problem, such as collecting human genomes from patients for a particular disease, gathering social media data for gender identification, or crawling websites for offensive materials to name just a few. In this paper, we call this problem-specific a class. When we know

the class of the dataset, the learning is called *supervised learning*. Otherwise, it is called *unsupervised learning* [1–3].

Most of the collected data suffer from high dimensionality the includes a high number of features. Most of these features are irrelevant and noisy [4, 5]. The noise in the dataset should be eliminated before the learning segment and only depend on a relevant feature set to accomplish reasonable learning accuracy leading to a reduction in the computational cost [6]. Assume  $X = \{x_1, x_2, \dots, x_m\}$  to be the dataset *m-by-n* matrix where  $m$  is the number of samples and the vector  $x_i$  is the  $i^{th}$  sample with  $n$  features. Therefore, the feature selection algorithm  $f()$  with respect to the class  $y$  could be represented in the following mathematical equation:

$$f(X, y) \rightarrow X',$$

where  $X'$  is a *m-by-r* and  $r \leq n$ , in real cases  $r \ll n$ . The set of features that are in  $X$  but not in  $X'$  are the irrelevant features to  $y$ . Moreover,  $f()$  maximizes learning performance while minimizing  $r$ .

There are few indicators of how good the feature selection algorithm  $f()$  is. Learning performance, such as: classification accuracy, is the most utilized indicator of the quality of  $f()$ . Usually, the learning performance can be described as  $\psi(X') \geq \psi(X)$ . Another indicator is the computational cost which is markedly smaller in  $X'$  than  $X$  due to the removal of irrelevant features. The third indicator has gained increasing attention in recent times which is due to selection stability. In machine learning tasks, we assume to build a learning model that is applicable to new samples from the same domain. However, this is not always the case; it reduces occurrence of the domain experts within our learning model. Feature selection stability  $\varphi()$  is the measure of how robust the feature selection algorithm  $f()$  is with respect to a certain level of permutation on data set  $X$  during the selection and learning processes.

In the realm of medicine, machine learning has been widely adopted and applied during the last few decades to expedite diagnosis and to deliver profoundly accurate inferences [7, 8]. Computerized Tomography Scan (CT-Scans), for instance, generate a large number of images for each patient. The physician's decision depends heavily upon analyzing these images and other related results [9]. However, it is time-consuming and difficult to identify diseases in a timely manner, especially, in case of early disease. Therefore, machine learning might play a critical role in supporting a doctor's decision [5]. One form of medical diagnosis and classification stem from *microarrays* which is a representational array of genome expression levels. Microarrays are burdened with vital issues when it comes to machine learning tasks, which we will discuss in more details in the sections below. The most important issue in microarrays is known as *curse-of-dimensionality* [7, 8, 10]. Typically, it has a large number of features, i.e., genes, with a small sample size, i.e., number of patients. Most of these features are irrelevant to the class causing lower learning performance and a sharp rise in computational costs. Thus, we need to select the most relevant subset features. However, unlike other domains, feature selection in medical field is not a preprocessing technique only, but a bio-marker discovery tool as well [11].

Relevant features set to a specific class is found to drastically differ with an introduction of new samples [12–14]. The measurement of this change is called *feature selection stability*. The stability of feature selection algorithm  $\varphi(f())$  means how sensitive the algorithm is to a perturbation on sample sets [15]. For illustration, assume we have microarray  $X$  with  $n$  genes for  $m$  patients who have been preliminarily diagnosed with colon cancer. Biologically, scientists believe that the number of genes related to this kind of cancer is very small comparison to the total number of genomes in the human DNA; thus, a small subset of  $n$  features should only be used to build the learning model. If feature selection algorithms select different subsets of genes each time we introduce new patients, domain experts would have less trust in our algorithms.

In this paper, we aim to propose an ensemble approach based on the bagging technique to improve stability while maintaining or improving learning performance, i.e., classification accuracy. The remainder of the paper is organized as follows: (i) we introduce feature selection algorithms in "Feature selection algorithm" section, (ii) we give a literature review of the stability and how to evaluate it in "Feature selection stability" section, (iii) we provide the proposed ensemble method in "Proposed method: bagging feature selection" section, (iv) we conduct an experiment on microarray datasets in "Experiment" section, and (v) we discuss the results and conclude the paper.

### Literature review

With the unprecedented growth of data dimensionality appears the complexity of understanding the relation between data attributes and the class label. Feature selection arises as a savior of this issue. In each machine learning problem, we aim to select the subset of the most relevant features to the target with minimum redundancy amongst them. Selecting a set of representative features found to be a very crucial not only to reduce computational complexity but also to improve learning performance [16–20]. However, due to the fact that introducing new samples to the data may introduce a noticeable amount of perturbation in the selected subset which make a confusing to the domain experts [14, 21–25].

Machine learning in the medical field is a significant tool nowadays for diagnosis and prognosis. Feature selection is extensively utilized to reduce the complexity of the data representation [5, 26–29]. However, due to the nature of medical data where a limited number of features are believed to be very related to the problem and a large number of redundant features exists, the reproducibility of the selected set is a challenge. Thus, robustness of selected features is gaining greater importance in the medical domain [10, 24, 30–33].

Stability of feature selection algorithms has drawn a significant attention lately due to the importance of robust selection with high performance as well [14, 15, 25, 34, 35]. Stability has been tackled differently. Several researchers have investigate stability from data perspective, such as: data noise, data characteristics, imbalanced classes, and feature redundancy [6, 31, 36, 37]. Others chose to tackle it from modeling perspective; including: sampling techniques, and multi-objective algorithm [24, 33, 35, 38–40]. Others investigated the evaluation metrics of stability [14, 25].

In [39], the authors proposed a multi objective feature selector based on evolutionary algorithms and wrapper model. The algorithm nominates different sets of features and the classifier evaluates the quality of the sets and optimizes the training and validation data accordingly to avoid overfitting. The authors claimed that the proposed solution was fairly stable in feature ranking using Spearman Correlation Coefficient. In another work, network-based algorithm was found to be stable with proteomics dataset [32]. They have found that statistical feature selection methods are not intrinsically stable with medical data, hence, it should be carefully tested and evaluated.

Combining different selectors is also was found to be stable [35]. An ensemble approach was investigated where they proposed to use different feature selection algorithms to aggregate the best subset. Different families of algorithms were utilized; including: filter, embedded, univariate, and multivariate. Also, ensemble approach using bagging and boosting was found accurate with respect to classification error metric in an online feature selection [41]. Similar findings were proposed in [11, 21]. The former proposed an ensemble ranker for feature selection and clustering that was found efficient. While the latter proposed feature grouping based on kernel density estimation.

Embedded feature selection algorithms employed on medical dataset to predict diabetes kidney disease consequences [33]. Their robustness analysis found that only Gradient Boosting Machine was fairly stable, while the rest of the results were not. Thus, the model clearly lacks generalization ability for other feature selectors. This is an issue due to the need to perform extensive experimentation and evaluation before choosing the proper approach that would satisfy the desired performance and robustness.

Others investigated selection stability from data perspective. Shanab et al. [31] studied the impact of data noise on selection stability. The noise was injected to the data; which was medical datasets. Since the original datasets were imbalanced, the training was conducted with different sampling techniques. They found that 50:50 class ratio, i.e. balanced data, and algorithms that handles data noise were better in terms of selection stability. In addition, in [6], the authors proposed a noise reduction framework using low rank matrix approximation technique to reduce noise and variance within the class. This approach was empirically proven powerful with regard to reduce data variance, hence, improving selection stability.

### **Feature selection algorithm**

Feature selection is one family of dimensionality reduction algorithms [17, 20, 42]. The other family is the feature extraction algorithms. The latter projects features into lower dimensional space; the newly generated features are not in the original space [43, 44]. Therefore, it cannot be interpreted or understood by domain experts. In the medical domain, if we apply Principal Component Analysis PCA, for instance, on a microarray, features in the new lower dimensional dataset are not genes [45]. Hence, biologists will not be able to determine which gene is relevant to the problem and which is not. On the other hand, feature selection keeps the original representation of features, and as a result, is interpreted, justifiable, and understood by domain experts.

A large number of feature selection algorithms have been proposed to improve learning performance while reducing computational cost. These algorithms can broadly be categorized by either according to the utilization of class information and *supervised or unsupervised algorithms*. [2, 3] reviews these two categories of algorithms. In this paper, we will be using supervised feature selection algorithms since we have datasets with class labels [46]. Another categorization is with respect to utilization of a classifier during the selection process. The *filter model* is the most widely used in this category; it is independent of any classifier. For example, Information Gain is a filter model that measures the amount of information of feature  $f_i$  given the class label  $y$ ,  $IG(f_i|y)$  [47, 48].

In contrast to the filter model, the *wrapper model* is dependent on a given classifier. Generally, a greedy approach is utilized to select different sets of features, and the classifier evaluates the quality of the selected sets with the best being Support Vector Machine Recursive Feature Elimination (SVM-RFE). It is one popular example of this model [27]. From its name, the algorithm starts by selecting the whole set of features and recursively eliminates features that are not effective in predicting the class label using the Support Vector Machine SVM classifier.

There are several benefits for each model. While the filter process is not as computationally costly as the wrapper model, the latter is more accurate in class prediction than the former since it uses the classifier in its selection process. Yet, the wrapper model usually tends to overfit especially with higher dimensionality, whereas filter generalizes easily. Therefore, a *hybrid model* was proposed to merge the two models and to benefit from the advantages of each while minimizing and removing the disadvantages. It selects different sets using the filter methodology and then compares the different sets using the wrapper approach. Thus, it is theoretically less complex than the wrapper model while more effective than the filter model.

### Feature selection stability

Biologists believe that genes related to a specific type of cancer to be a limited set of genes. In other words, a variant of cancer can be caused by permutations of genes, and these genes are almost always the same with different patients. However, feature selections algorithms show instability in selected feature sets with a small perturbation in the data set, such as: introducing new samples. Therefore, instability would lower the confidence of domain experts in the selection algorithm even if it performs well due to a contradiction with a biological fact [30, 34, 38]. Hence, the stability of feature selection algorithms is a highly desired characteristic.

*Stability of feature selection algorithms* can be defined as how similar the selected feature sets with specific levels of perturbation on the training datasets in a controlled environment [21, 37, 49]. It is impacted mostly by data *variance*. The variance, in terms of feature selection, is the variability of the selection model for a given sample. Assume, we have a microarray dataset  $X$  with 100 samples. Then, we randomly generated two sets  $X_1$  and  $X_2$  out of  $X$ , where  $X_1 = \{x_1, x_2, \dots, x_{90}\}$  and  $X_2 = \{x_{10}, x_{11}, \dots, x_{100}\}$ . The overlap between  $X_1$  and  $X_2$  is exactly 80%. The following equations illustrate the stability notion:

$$f(X_1, y_1) = X'_1 \text{ and } f(X_2, y_2) = X'_2$$

Assume,  $\mathcal{F}(X'_1)$  is the selected set of features and we call it  $\mathcal{F}_1$  for short. Similarly,  $\mathcal{F}_2$  is the selected feature set of  $X_2$ . Without loss of generality, we are going to use  $\mathcal{F}$  to represent the selected features whenever we need to represent dataset  $X$ . The stability  $\varphi(f)$  of algorithm  $f()$  can be estimated by the evaluation of similarity of  $\mathcal{F}_1$  and  $\mathcal{F}_2$  as follows:

$$\varphi(f(X_1, y_1), f(X_2, y_2)) = \Psi(\mathcal{F}_1, \mathcal{F}_2); \quad (1)$$

where  $\Psi$  is a similarity measure between  $\mathcal{F}_1$  and  $\mathcal{F}_2$ . There are several measures to evaluate stability [15, 25, 50]. For example, the *Canberra Distance Measure* can be used to find the similarity between two ranked sets of features, while the *Euclidean Distance Measure* could be utilized to measure the distance between two sets of feature weights. Yet in this paper, we used the *Jaccard Index Measure* to measure the stability of the selected set, since  $\mathcal{F}$  contains a set of feature indexes with no specific order, rank, or weight.

### Jaccard Index

To evaluate the similarity between two selected subsets of features, Jaccard Index Eq. (2) calculates the proportion of intersection of the selected features to the total number of selected features in the two subsets combined.

$$\mathcal{J}(\mathcal{F}_1, \mathcal{F}_2) = \frac{|\mathcal{F}_1 \cap \mathcal{F}_2|}{|\mathcal{F}_1 \cup \mathcal{F}_2|} \quad (2)$$

Jaccard has several advantages as a stability measure. First, it is symmetric, thus,  $\mathcal{J}(\mathcal{F}_1, \mathcal{F}_2) = \mathcal{J}(\mathcal{F}_2, \mathcal{F}_1)$ . Also, it is monotonic. In other words, the larger the intersection between  $\mathcal{F}_1$  and  $\mathcal{F}_2$ , the larger  $\mathcal{J}$  will be. Besides that, Jaccard is always between 0 and 1.

On the other hand, Jaccard has some disadvantages in some cases. For instance,  $\mathcal{J}$  is larger when the size of selected feature subset is close to the original feature space size. In this case, intersection by chance is very likely. Yet, this is seldom in real cases since we usually aim to minimize the selected feature subset size. After all, Jaccard is widely used as a stability metric, and hence, we use it in this paper.

Generally, we evaluate the stability of an algorithm with respect to a data set by either (1) randomly sampling of different samples from the dataset or (2)  $l$ -fold cross-validation. Eq. (3) illustrates the mathematical form of  $l$  sampling time. The larger  $l$  is, the more accurate stability estimation will be.

$$S(\mathbf{F}') = \frac{2}{l(l-1)} \sum_{i=1}^{l-1} \sum_{j=i+1}^l \mathcal{J}(\mathcal{F}'_i, \mathcal{F}'_j) \quad (3)$$

where  $\mathbf{F}' = \{\mathcal{F}'_1, \mathcal{F}'_2, \dots, \mathcal{F}'_l\}$ .

### Proposed method: bagging feature selection

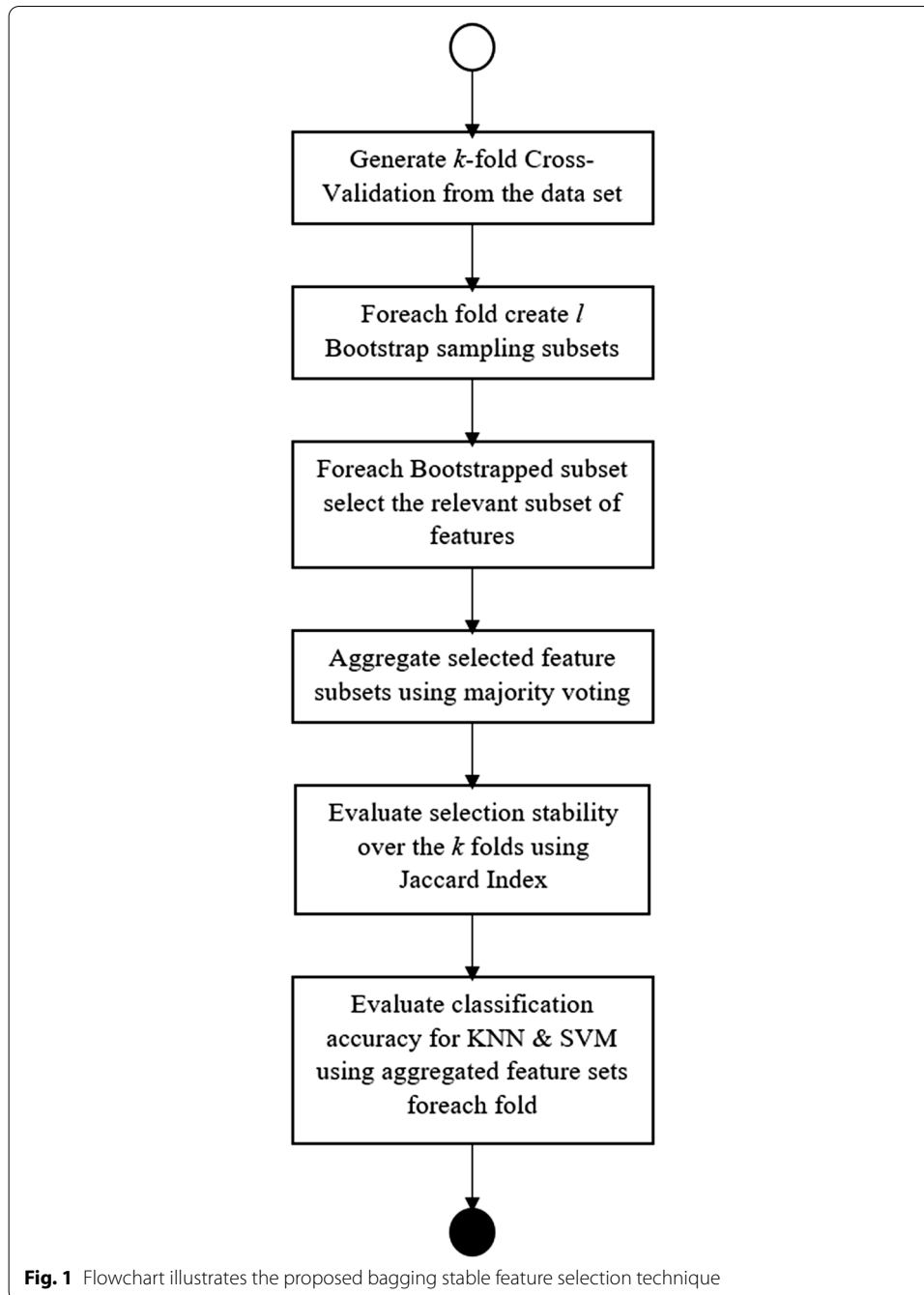
In "[Feature selection algorithm](#)" and "[Feature selection stability](#)", we have seen that feature selection algorithms are utilized to reduce dimensionality by removing irrelevant features while maintaining learning performance. In some domains; such as the medical field, they have been used to identify a subset of features, say genes, that are relevant to the problem, let us suppose colon cancer. Biologists believe that sets of genes related to a certain disease are almost always the same. However, when we apply a feature selection algorithm to a dataset for several times with introduced noise and/or perturbation we may end up selecting an entirely different set of features. Consequently, domain experts do not trust algorithms' results since this instability contradicts with a biological assumption.

Boosting and Bagging are very popular ensemble techniques. Although they are similar in a sense that they both combine weak learners to aggregate a better model, they have several differences. First, Boosting performs data sampling without replacement. Then, it performs another deterministic sampling from the first round including a certain percentage of misclassified samples. This approach leads to lower classification error. In other words, learning bias is better with Boosting while data variance increases due to overfitting. On the other hand, *Bagging* (*Bootstrap Aggregating*) samples data with replacement which leads to randomly selecting data points. Unlike Boosting, Bagging is known for the ability to reduce data variance, i.e., the overfitting [51]. In "[Feature selection stability](#)" section, we mentioned errors due to variance which can be seen as a leading cause of selection instability. Therefore, we assume that reducing variance error may help improving stability. It was noted that we need to be careful here due to the fact that a lower variance leads to a higher bias, which is another error that is not preferred. Thus, our aim is to lower the variance error to an ordinarily acceptable level. If our assumption is valid, reducing variance will markedly improve selection stability. We propose to use Bagging technique to reduce data variance, hence, improving the stability. However, improving stability should not greatly impact learning performance. In the Results Section, Bagging selection proved to be stable and reasonably accurate.

Bagging [52] aggregates selection model by vote using the same algorithm on multiple bootstrap samples of the same dataset. We withdraw samples of dataset  $X$  for  $l$  times. Different sampling techniques are possible. Yet, leave-one-out, for instance, leads to a very simple model. Hence, a very high bias will occur. In contrast, model training using each data point ends up producing an overly complex model, and therefore, a very high variance and an overfitting model. As a result, a moderated selection is better. In "[Experiment](#)" section, we explain in details the selection model.

As shown in Algorithm(1), bagging algorithm starts by bootstrap sampling of the dataset, assume  $l$  samples. Then, we train the model on each bootstrap sample which produces a selected  $l$  sets of features. Where  $\mathbf{F}'$  is a two dimensional  $l - by - n$  matrix. The  $i^{\text{th}}$  row represents the selected feature set from the bootstrap sample  $(X', y')_i$ , while each  $i^{\text{th}}$  column is the value of the selected feature. This value could be the weight of that feature or simply a binary number representing whether the feature was selected. After





that, we aggregate the selected feature set. Each feature's value is summed and averaged over all bagging selection matrix,  $F'$ . We aggregate here by voting, where the most occurring feature will be assigned the highest weight. In the experiment, we will repeat this process several times in order to be able to evaluate the stability. We illustrate the proposed technique in a flowchart shown in Fig. 1.



---

```

Data: Input dataset  $(X, y)$ 
Result: 1-Selected Features  $\mathbf{F}' = \{\mathcal{F}'_1, \mathcal{F}'_2, \dots, \mathcal{F}'_l\}$ 
          2-Aggregated Features  $\mathbf{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n\}$ 
for  $i = 1 : l$  do
  |  $(X', y')_i = \text{Bootstrap Sampling of } (X, y);$ 
  |  $\mathcal{F}'_i = f(X', y')_i;$ 
end
 $\mathbf{F}' = \{\mathcal{F}'_1, \mathcal{F}'_2, \dots, \mathcal{F}'_l\}$ 
for  $j = 1 : n$  do
  |  $\mathcal{F}_j = \frac{1}{l} \sum_{i=1}^l \mathcal{F}'_{ij}$ 
end
 $\mathbf{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n\}$ 

```

---

### Algorithm 1: Bagging Algorithm

Without loss of generality, we can apply any feature selection algorithm  $f(\cdot)$  in Algorithm (1) due to our assumption that reducing data variance contributes positively to selection stability. In this work, we make use of five well-know feature selection algorithms. More details about the algorithms could be found in "Algorithms" section.

## Experiment

In this section, we perform several experiments using a variety of selection algorithms on different publicly available medical datasets, Table 1. To ensure generalization and to make it a comprehensive experiment, the datasets varied in sample size and dimensions. Similarly, the algorithms belong to different families of feature selection algorithms. Additionally, we conducted the experiment with different cardinalities of selected feature sets  $\mathcal{F}_i$ . In addition, to ensure that the improve of selection stability is not due to random behavior, we used two well-known classifiers, namely: SVM and KNN, as performance evaluation methods. Finally, we used the Jaccard Index to evaluate selection stability. The baseline to validate our proposed method is the same algorithms without bagging, since we claim that bagging improves selection stability while maintaining learning performance. In this following sub-sections, we are going to discuss our methodology and experiments in details.

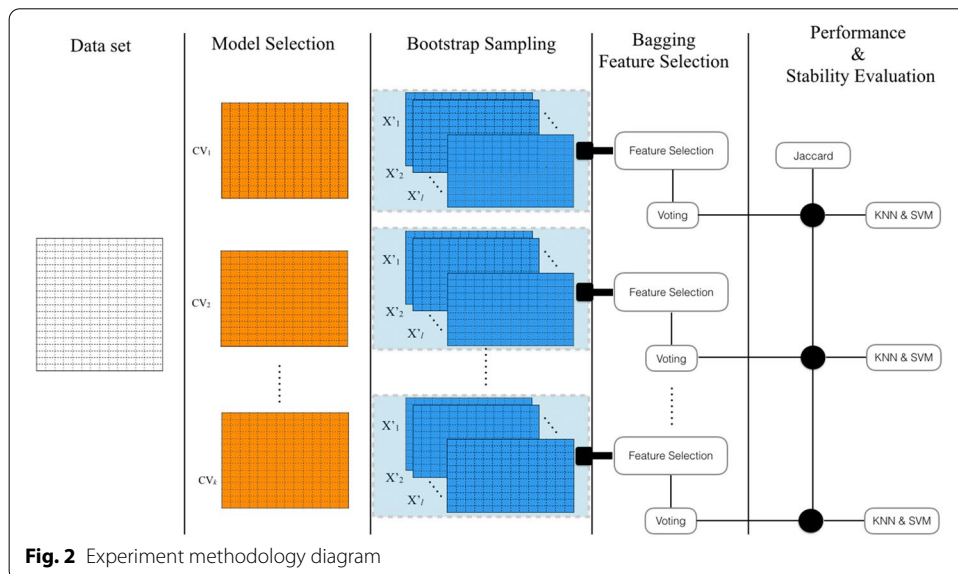
### Experiment methodology

As shown in Fig. 2, we start by taking each dataset and apply cross-validation  $l$  times. This step generates  $l$  training datasets and  $l$  test sets. After that, bootstrap sampling is applied to each training set. This generates another  $l$ -folds of each training sample. Then, we apply feature selection algorithms on each fold to select  $k$  number of features for each fold. Without loss of generality, the number of folds  $l$  may differ between the cross-validation stage and bagging stage.

Next, we evaluate the stability of the bagging feature selection algorithm. The baseline method that we compare our method with is the normal feature selection without bagging. The final stage is performance and stability evaluation. We evaluate stability using Jaccard Index and accuracy of the selected sets using SVM and KNN. We repeat the

**Table 1 Datasets statistics**

	Dataset name	#Samples $n$	Dimensionality $m$
1	BLOOD	89	2759
2	SMK-CAN	187	19993
3	Colon	62	2000
4	Leukemia	72	12582



mentioned steps for each data set on each algorithm and each cardinality of the selected feature set from 1 to 100 selected features.

**Medical datasets**

In this experiment, we used 4 microarrays listed in Table 1 which are publicly available.<sup>1</sup> All datasets are relatively high-dimensional with a very small sample size,  $n \gg m$  [53]. These characteristics are problematic in the machine learning task. Datasets suffer from what is known as the *curse-of-dimensionality*. This problem becomes even more challenging when we have a small number of samples. In such case, the learner will not be able to converge due to these two factors. Since, the more features we have, the more data samples we need.

Table 1 shows medical datasets, DNA microarrays or biomarkers. Microarrays are gene expressions collected from tissues and cells usually to diagnose tumors [29]. They are either continuous or discrete feature values. Feature selection in gene expression dataset usually helps removing irrelevant and redundant genes and to find relevant set of genes related to a certain kind of tumor. In this paper, we used different types of data sets with different characteristics to ensure generalization of proposed method.

<sup>1</sup> The datasets were obtained from: <http://featureselection.asu.edu/>.

### Algorithms

As we discussed in "Feature selection algorithm" section, there are different families of feature selection models. In this paper we used 5 algorithms to conduct the experiment. We consider each algorithm to be a baseline for the proposed bagging approach of the same algorithm since our assumption is that bagging should reduce variance, and hence, improve selection stability. The chosen algorithms are either filter or wrapper algorithms.

Fisher Score [54, 55] is a supervised filter model that utilizes fisher criteria to evaluate features independently. This leads to selecting a subset of features that maximizes the distance between data points in different classes while minimizing distance within a given class.

ReliefF [55, 56], similarly, is a supervised and filter feature selector. It is a heuristic algorithm that utilizes k-nearest neighbor to select features in multi-classes, noisy, dependent and incomplete data. Therefore, unlike the Fisher score, it can detect conditional dependencies between features [57].

Information Gain [47] measures the dependency between feature and class label. It represents the difference between the entropy of the feature and the entropy of the feature given the class label.

Chi-square score ( $\chi^2$ ) is similar in notion to Information Gain. It measures how independent the feature is from class label [42]. All the aforementioned algorithms are filter and supervised models. However,  $\ell_1$ SVM represents the wrapper approach [16, 58, 59]. It finds features that maximize separation between classes using SVM and 1-norm optimization since it utilizes SVM to evaluate the separability of each feature with respect to the classes.

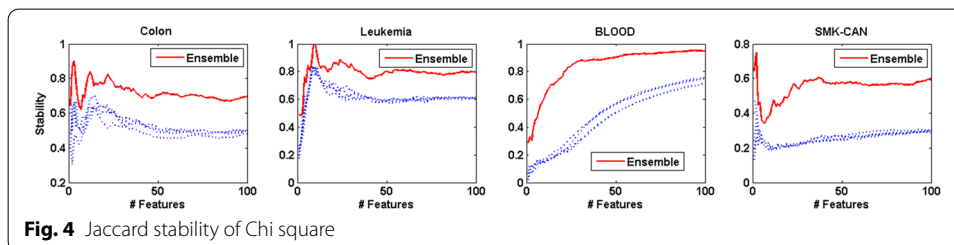
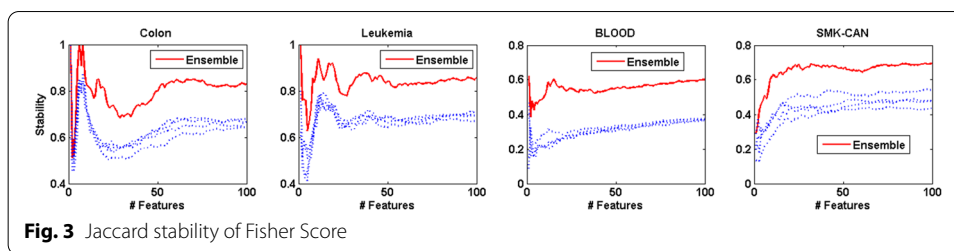
### KNN and SVM classifiers

Stability by itself is a highly desirable feature. However, it could be achieved by a characteristically ad-hoc algorithm that selects the same sets of features each time. It should be noted, that is not the purpose of feature selection. Indeed, we need stability, yet, stable selection should select very relevant features that are highly discriminative in their features. In other words, we are looking for an algorithm that performs well in terms of either accuracy, precision or any performance measure and performs stable selection as well.

In order to evaluate selection quality, we used two well-known accuracy measurements, namely: K-Nearest Neighbor KNN and Support Vector Machine SVM. KNN classifies data points by majority voting out of K nearest neighbors using a distance measure [60]. These distance measures could be Euclidean, Manhattan, Hamming Distance and so forth. On the other hand SVM creates separating hyperplanes between classes. It finds the optimal plane by maximizing the margin between classes [61].

### Model selection

Model selection in its simplest form is choosing the parameters that perform the best on test data. It is still an open problem [20]. In this experiment, Fig. 2, we have several models to be selected. To ensure the reliability of our proposed method, we needed to

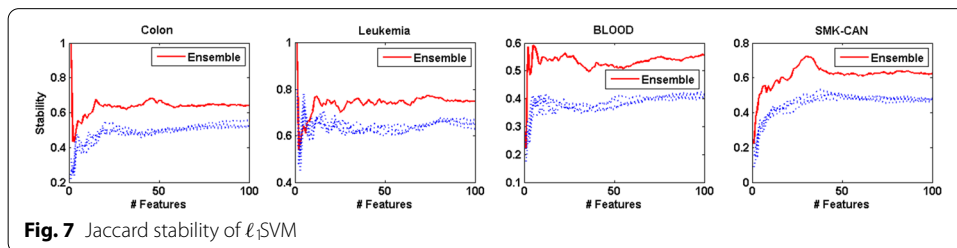
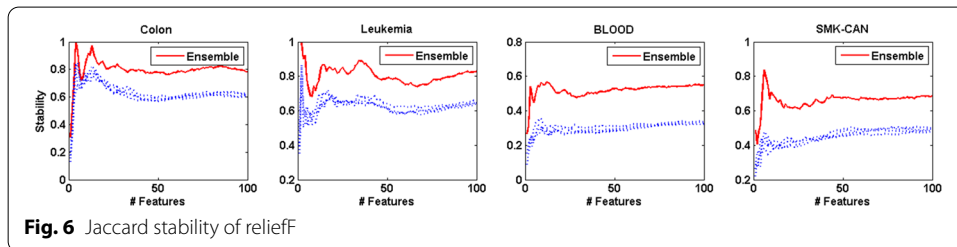
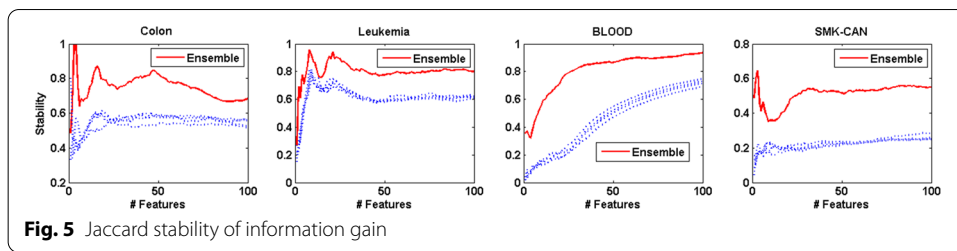


perform model learning several times. Also, we needed to randomly sample test sets of data points each time.  $l$ -fold Cross-Validation is proven to be effective in such tasks [20, 62–64]. Another selection to be chosen is the bootstrap sampling. We generally chose  $l = 10$  for both models since we had a small sample size in the original datasets  $X$ .

The last model to be selected is the number of selected features  $k$ . To the best of our knowledge, there is no rule of thumb of choosing an optimal number of features. Thus, practitioners consult domain experts in this open problem. In In [65], the authors tried to determine the proper  $k$  using false nearest neighbor from chaos theory for clustering. In this experiment, we chose to examine the behavior of the algorithms' stability and accuracy over different cardinalities where  $k$  ranges from 1 to 100. In other words, we performed the experiment 100 times for each algorithm on each dataset, and selecting one feature more each time. This number was reasonable in earlier studies, see [3, 36]. The reason behind this model selection is the believe that the number of genes related to a certain disease is small. Also, this range of  $k$  shows us the erratic behavior of the selection stability, if any. We will explain in details about the effect of that in “[Results and discussion](#)” section.

## Results and discussion

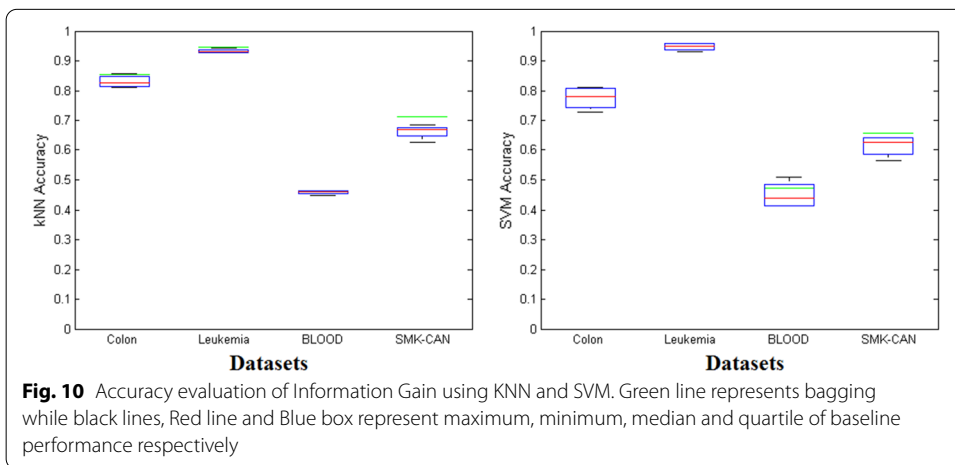
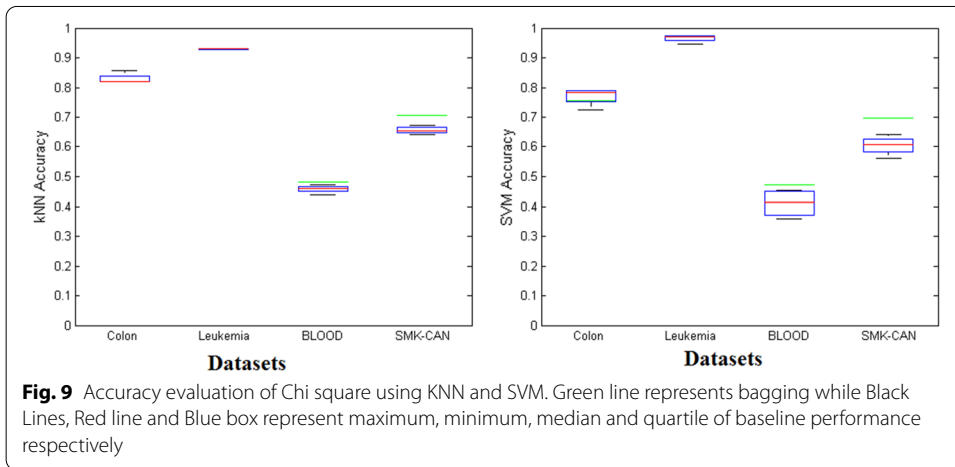
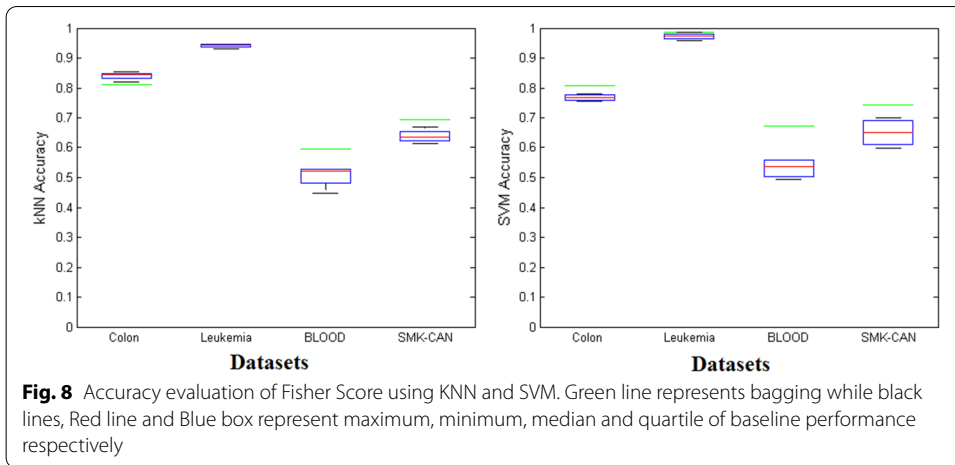
Our assumption was that Bagging technique is going to reduce data variance, hence, selection stability will improve. However, improving stability should not degrade classification accuracy. Otherwise, we could randomly select the same subset of features each time. Therefore, our aim is to improve selection stability while maintaining classification accuracy. The conducted experiments showed the effectiveness of our proposed method in both stability and classification accuracy. Figures 3, 4, 5, 6 and 7 illustrate the Jaccard stability of each algorithm on each microarray dataset using range of cardinalities of  $\mathcal{F}$ . As we can see in the figures, the proposed ensemble bagging technique improves the stability on each single case. The improvement mostly exceeded 20 percent. In some cases, the improvement exceeds 50 percent Fig. 4 on *Blood* and *SMK-CAN* datasets. We can observe some cases where selection stability reaches 100 percent with ensemble



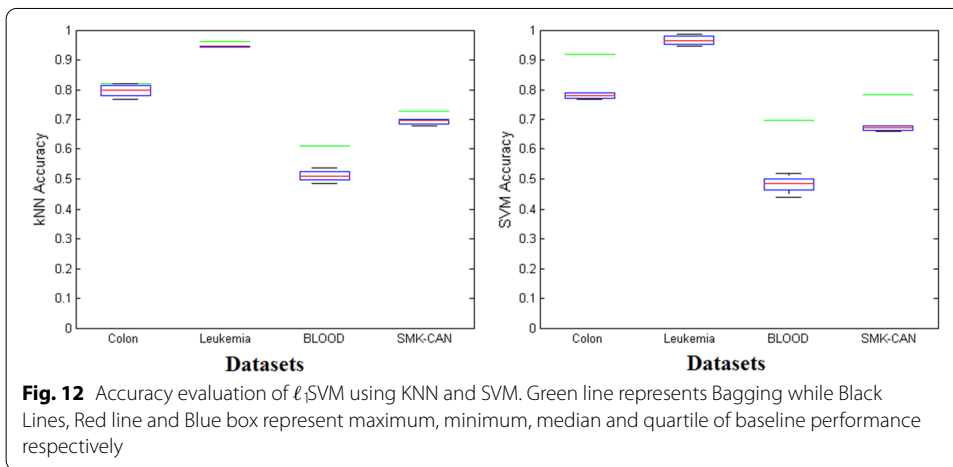
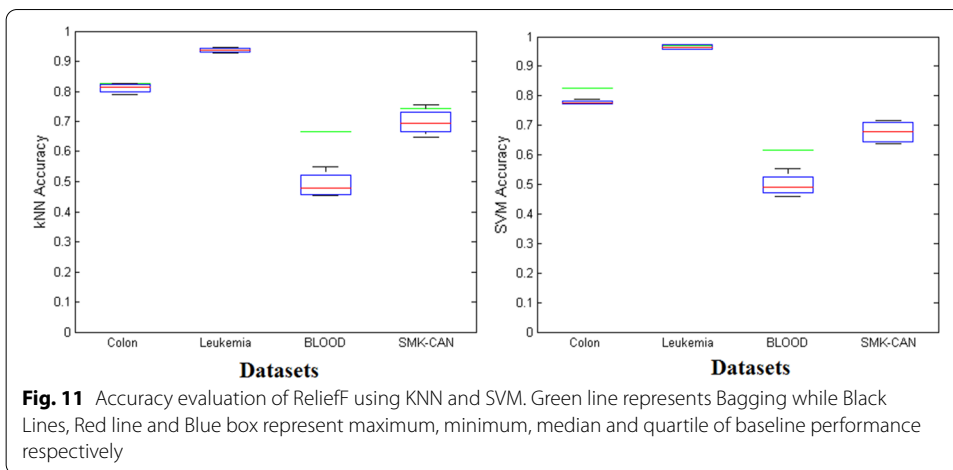
selection; see for example, Fig. 3 with Colon and Leukemia datasets and Fig. 4 with Leukemia. This pattern occurs when  $k$  is very small which indicates that the algorithms were able to identify the most relevant features in a very robust manner. These results means a lot in the medical domain since algorithms produce robust set of genes each time with respect to perturbation in data samples which is equivalent in this context to diagnosing.

The results show that a moderately small  $k$  is very sensitive. Yet, it becomes more and more steady when  $k$  gets greater. Also, this pattern is consistent in all cases. Therefore, machine learning practitioners may need to consider this when they intend to build a learning model. This sensitivity is another nice property of this approach since it gives the practitioners an indicator of the best number of selected features. In other words, when the feature selection stability start to degrade and stabilize we can determine that point to be the proper number of selected features. This due to the fact that irrelevant and redundant features are now being selected randomly. This is an open problem that needed to be further investigated in the future.

The accuracy, on the other hand, was very satisfactory. Out of 40 experiment with four datasets, five feature selection algorithms, and two different accuracy metrics, the accuracy improves 21 times, was in the average interval 18 times, and less that the average in one case only. The improvement was significant where it exceeded 10 percents in some cases; such as: Blood with ReliefF; Fig. 11, and Colon dataset with  $l_1$ SVM; Fig. 12. In some cases, the proposed technique was able to maintain or improve the accuracy even when the average accuracy was already more than 95 percent for the baseline;



see Leukemia dataset in Figs. 8, 9, 10, 11 and 12. This is an evidence that the proposed method is able to maintain the accuracy and even improves upon a high performance



algorithms. The only degrade in accuracy occurs in Fig. 8 the case of Colon dataset with KNN metric. It is less than the minimum quartile with around 1 percent only. However, the same case improved with SVM as the second plot shows in the same figure.

We believe, it is good enough to improve stability while maintaining accuracy. However, the proposed method improved both stability and accuracy. In almost all cases the proposed method beats the maximum quartile among the baseline. Yet, there are few cases where the accuracy is maintained, i.e. in the range of the baseline accuracy. Figure 9 with Leukemia and Colon datasets are good examples of such cases.

Ultimately, the proposed bagging algorithm is experimentally proven to be very effective with medical data in selecting relevant genes and maintaining, or even improving, classification accuracy. We believe that bagging approach is a very stable in terms of feature selection due to the intrinsic power of reducing learning variance. As consequence, the influence of random sampling would be mitigated.



## Conclusion and future work

Medical datasets usually suffers from undesirable property; which is the high dimensionality with very small sample size. Also, the number of relevant features to the problem is usually small comparing the total number of features. With this ill-favored scenario, selection stability is a real challenge. Since selection stability is very related to data variance, our assumption was made so if we reduce data variance we can improve selection stability. Bootstrap Aggregating (Bagging) was proven to mitigate data variance.

We propose to incorporate Bagging technique in the process of feature selection to improve selection robustness. The experiment was conducted over four real-world medical datasets which suffer from the aforementioned undesirable property. Also, five different well-known feature selection algorithm were utilized. Besides, two classification algorithms were Incorporated. In almost all cases, the proposed method was found to overcome the baseline approach with considerable stability improvement ranging from 20 percent to 50 percent. Additionally, the classification accuracy was improved in more than half of the experiments and maintained in the remaining. This was an important finding since improving stability must not be at the expense of the accuracy.

The proposed method was a very successful in improving selection stability while maintaining classification accuracy. This proofs the assumption made in the paper that reducing the data variance significantly improves selection stability. We believe that employing bagging technique in feature selection is a compulsory step for robust models.

A further investigation and improvement is possible. choosing the proper number of selected features is an open research problem where stability might be an indicator of the optimal number. In the future, we will tackle this issue to optimize this hyperparameter. In addition, this method might be tested on unsupervised learning, e.g. clustering. This is an important problem to deal with since selection stability can be tackled, evaluated and interpreted differently.

### Abbreviations

CT-Scans: Computerized Tomography Scan; PCA: Principal Component Analysis; SVM: Support Vector Machine; SVM-RFE: Support Vector Machine Recursive Feature Elimination; Bagging: Bootstrap Aggregating.

### Acknowledgements

This work would not have been possible without the financial support King Khalid University. I would like to express my deepest gratitude to their generous support.

### Authors' contributions

SA contributed 100% to this paper. The author read and approved the final manuscript.

### Funding

This research was supported in part by Deanship of Scientific Research at King Khalid University research grant number (R.G.P2/100/41).

### Data availability

All datasets used in this paper are available in ASU feature selection dataset repository: <http://featureselection.asu.edu/>.

### Competing interests

I declare that I have no competing financial, professional, or personal interests that might have influenced the performance or presentation of the work described in this manuscript.

Received: 2 September 2020 Accepted: 26 November 2020

Published online: 07 January 2021

## References

1. Dy JG, Brodley CE. Feature selection for unsupervised learning. *J Mach Learn Res.* 2004;5:845–89.
2. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. *Data Classification: Algorithms and Applications.* 2014;37.
3. Alelyani S, Tang J, Liu H. Feature selection for clustering: a review. *Data Clust.* 2013;29:110–21.
4. Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. 2003;856–863.
5. Leung YY, Chang CQ, Hung YS, Fung PCW. Gene selection for brain cancer classification. *Conf Proc IEEE Eng Med Biol Soc.* 2006;1:5846–9.
6. Alelyani S, Liu H. Supervised low rank matrix approximation for stable feature selection 2012;1:324–329. *IEEE*
7. Zhang M, Zhang L, Zou J, Yao C, Xiao H, Liu Q, Wang J, Wang D, Wang C, Guo Z. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics.* 2009;25(13):1662–8.
8. Saeys Y, Inza I, Larraaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* 2007;23(19):2507–17.
9. Han C, Tao X, Duan Y, Liu X, Lu J. A cnn based framework for stable image feature selection, 2017;1402–1406. *IEEE.*
10. Boulesteix A-L, Slawski M. Stability and aggregation of ranked gene lists. *Brief Bioinform.* 2009;10(5):556–568. <http://bib.oxfordjournals.org/cgi/reprint/10/5/556.pdf>.
11. Drotár P, Gazda M, Vokorokos L. Ensemble feature selection using election methods and ranker clustering. *Inf Sci.* 2019;480:365–80.
12. Kuncheva LI. A stability index for feature selection. 2007;390–395.
13. Jurman G, Merler S, Barla A, Paoli S, Galea A, Furlanello C. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics.* 2008;24(2):258–64.
14. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowl Inf Syst.* 2007;12(1):95–116.
15. Alelyani S. On feature selection stability: A data perspective. PhD thesis, Arizona State University, 2013.
16. Bradley PS, Mangasarian OL. Feature selection via concave minimization and support vector machines. *Machine Learning Proceedings of the Fifteenth International Conference.* 1998;82–90.
17. Das S. Filters, wrappers and a boosting-based hybrid for feature selection, 2001;74–81.
18. Dash M, Choi K, Scheuermann P, Liu H. Feature selection for clustering - a filter solution. 2002;115–122.
19. Forman G. An extensive empirical study of feature selection metrics for text classification. *J Mach Learn Res.* 2003;3:1289–305.
20. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3:1157–82.
21. Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. 2008;803–811.
22. Loscalzo S, Yu L, Ding C. Consensus group stable feature selection. 2009;567–576.
23. Somol P, Novovicov J. Evaluating the stability of feature selectors that optimize feature subset cardinality. *Structural, Syntactic, and Statistical Pattern Recognition,* 2010;956–966.
24. Yu L, Han Y, Berens ME. Stable gene selection from microarray data via sample weighting. *IEEE/ACM Trans Comput Biol Bioinform.* 2011;9(1):262–72.
25. Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. *J Mach Learn Res.* 2017;18(1):6345–98.
26. Model F, Adorjn P, Olek A, Piepenbrock C. Feature selection for DNA methylation based cancer classification. *Bioinformatics.* 2001;17(Suppl 1):157–64.
27. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* 2002;46(1):389–422.
28. Cawley GC, Talbot NLC. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics.* 2006;22:2348–55.
29. Bolon-Canedo V, Sanchez-Marono N, Alonso-Betanzos A, Benitez JM, Herrera F. A review of microarray datasets and applied feature selection methods. *Inf Sci.* 2014;282:111–35.
30. Abeel T, Helleputte T, de Peer YV, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics.* 2010;26(3):392–8.
31. Shanab AA, Khoshgoftaar TM, Wald R, Napolitano A. Impact of noise and data sampling on stability of feature ranking techniques for biological datasets, 2012;415–422. *IEEE.*
32. Goh WWB, Wong L. Evaluating feature-selection stability in next-generation proteomics. *J Bioinform Comput Biol.* 2016;14(05):1650029.
33. Song X, Waitman LR, Hu Y, Yu AS, Robins D, Liu M. Robust clinical marker identification for diabetic kidney disease with ensemble feature selection. *J Am Med Inf Assoc.* 2019;26(3):242–53.
34. He Z, Yu W. Stable Feature Selection for Biomarker Discovery (2010). <http://www.citebase.org/abstract?id=oai:arXiv.org:1001.0887>.
35. Pes B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. *Neural Comput Appl.* 2019;1–23.
36. Alelyani S, Liu H, Wang L. The effect of the characteristics of the dataset on the selection stability, 2011;970–977. *IEEE.*
37. Gulgezen G, Cataltepe Z, Yu L. Stable and accurate feature selection. Berlin: Springer; 2009. p. 455–468.
38. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. *Machine Learning and Knowledge Discovery in Databases.* Berlin: Springer; 2008. p. 313–325.
39. González J, Ortega J, Damas M, Martín-Smith P, Gan JQ. A new multi-objective wrapper method for feature selection-accuracy and stability analysis for bci. *Neurocomputing.* 2019;333:407–18.
40. Baldassarre L, Pontil M, Mourão-Miranda J. Sparsity is better with stability: combining accuracy and stability for model selection in brain decoding. *Front Neurosci.* 2017;11:62.
41. Ditzler G, LaBarck J, Ritchie J, Rosen G, Polikar R. Extensions to online feature selection using bagging and boosting. *IEEE Trans Neural Netw Learn Syst.* 2017;29(9):4504–9.

42. Liu H, Setiono R. Chi2: Feature selection and discretization of numeric attributes, 1995;388–391.
43. Guyon I, Elisseeff A. An introduction to feature extraction. *Feature extraction*. 2006;1–25.
44. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. *Science*. 2000;290(5500):2323–6.
45. Abdi H, Williams LJ. Principal component analysis. *Wiley Interdiscipl Rev*. 2010;2(4):433–59.
46. Song L, Smola A, Gretton A, Borgwardt K, Bedo J. Supervised feature selection via dependence estimation, 2007.
47. Cover TM, Thomas JA. *Elements of information theory*. Hoboken: Wiley; 1991.
48. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. *J Royal Stat Soc*. 2008;70(1):53–71.
49. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms. *Data Mining, Fifth IEEE International Conference on*, 2005;8.
50. Chelvan PM, Perumal K. A comparative analysis of feature selection stability measures, 2017;124–128. IEEE.
51. Breiman L. Bias, variance, and arcing classifiers, 1996.
52. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123–40.
53. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, Liu H. Feature selection: A data perspective. 2017; arXiv preprint [arXiv:1601.07996](https://arxiv.org/abs/1601.07996).
54. Gu Q, Li Z, Han J. Generalized fisher score for feature selection. arXiv preprint [arXiv:1202.3725](https://arxiv.org/abs/1202.3725), 2012.
55. Zhao Z, Morstatter F, Sharma S, Alelyani S, Anand A, Liu H. Advancing feature selection research. ASU feature selection repository. 2010;1–28.
56. Kononenko I. *Estimating attributes: analysis and extensions of relief*. Berlin: Springer; 1994. p. 171–182.
57. Sikonja MR, Kononenko I. Theoretical and empirical analysis of relief and relief. *Mach Learn*. 2003;53:23–69.
58. Bi J, Bennett K, Embrechts M, Breneman C, Song M. Dimensionality reduction via sparse support vector machines. *J Mach Learn Res*. 2003;3(Mar):1229–433.
59. Joachims T, Informatik F, Informatik F, Informatik F, Informatik F, Viii L. Text Categorization with Support Vector Machines: Learning with Many Relevant Features, 1997. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.11.6124rep=rep1type=pdf>.
60. Witten IH, Frank E. *Data mining: Practical machine learning tools and techniques*. Portland: ACM SIGMOD Book; 2005.
61. Suykens J, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9(3):293–300.
62. Kohavi R et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. 1995;14(2):1137–1145. Stanford.
63. John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference*. 1994;121–129.
64. Ng AY. On feature selection: Learning with exponentially many irrelevant features as training examples. *Proceedings of the Fifteenth International Conference on Machine Learning*. 1998;404–412.
65. Andrade Filho JA, Carvalho AC, Mello RF, Alelyani S, Liu H. Quantifying features using false nearest neighbors: An unsupervised approach. 2011;994–997.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---