An Ensemble Feature Selection Framework Integrating Stability

1st Xiaokang Zhang
Computational Biology Unit
Department of Informatics
University of Bergen
Bergen, Norway
Xiaokang.Zhang@uib.no

2nd Inge Jonassen
Computational Biology Unit
Department of Informatics
University of Bergen
Bergen, Norway
Inge.Jonassen@uib.no

***Abstract*—Ensemble feature selection has drawn more and more attention in recent years. There are mainly two strategies for ensemble feature selection, namely data perturbation and function perturbation. Data perturbation performs feature selection on data subsets sampled from the original dataset and then selects the features consistently ranked highly across those data subsets. Function perturbation frees the user from having to decide on the most appropriate selector for any given situation and works by aggregating multiple selectors. Our study showed that function perturbation resulted in a low stability. We therefore propose a framework, EFSIS (Ensemble Feature Selection Framework Integrating Stability), combining these two strategies and integrating stability during the aggregation of selectors. Empirical results indicate that EFSIS highly improves stability and meanwhile, maintains the prediction accuracy.**

***Index Terms*—feature selection, ensemble learning, stability**

I. Introduction

Feature selection is a crucial technique in machine learning. It is widely used in many fields to help to find the most important features. In classification tasks, feature selection can help to improve the prediction accuracy by removing the noisy features and avoiding overfitting. But feature selection can also be very challenging, especially when there is a large number of features (high-dimension) and very few training samples, which is quite often the case in biomedicine and genomics. In such cases a small change in the samples used as training set, can sometimes lead to a large change in the set of selected features. The ability of a feature selection method to give a consistent set of features when the training data changes, is called stability. So, a good feature selection method should enable the chosen classifier to obtain high prediction accuracy and also be stable to provide similar selected feature subsets.

In the field of prediction, ensemble learning has been shown to improve the stability and prediction accuracy of the individual learners [2]. The ensemble logic has been more and more applied to feature selection problem in recent years.

Ensemble feature selection methods can mainly be divided into two categories: data perturbation and function perturbation [3], [4].

In data perturbation (sometimes referred to as the homogeneous ensemble approach), feature selection is performed on several subsets of the samples, each analysis generating potentially different feature subsets. In this case the same feature selection method is used to analyze all subsets. The resulting feature subsets are then aggregated into one final feature subset [5]–[9]. Pes et al. showed that data perturbation can improve the stability of the original feature selection method [9].

Function perturbation (also referred to as the heterogeneous ensemble approach) combines the outputs from several feature selection methods - to free the user from having to choose one selection method and to benefit from the strengths of a set of methods [8], [10]–[12]. In this approach, a set of selected feature selection methods are all applied on the same training set. According to the literature, function perturbation can maintain or improve classification performance.

However, we have not been able to find in the literature any study of the stability of function-perturbation based methods for feature selection.

The concern is that each feature selection method makes different sets of assumptions and rationale for choosing the important features; combining selected features from across different selectors may give inferior performance including decreased stability. Especially in the field of biomedicine or genomics, where the feature dimension is very high but the sample number is comparably low, such as microarray data, a small change in the dataset may produce large change in the resulting features. Therefore, we find it highly relevant to investigate the issue of stability in ensemble feature selection and especially in context of function perturbation approaches.

Through our preliminary experiments, we found that function perturbation could indeed result in low stability. Since data perturbation has been shown to improve stability, we propose a framework to combine these two strategies to solve the stability issue of function perturbation.

The framework includes two phases. In the first phase, data perturbation is applied to generate a number of data subsets and each of these is given to a number of feature selectors (also referred to as rankers since they rank the features). For each ranker, the results across data subsets are aggregated to produce one ranked list of features. In addition, for each ranker, a statistic reflecting its stability is calculated. In the second phase which is function perturbation, the results from each ranker

are aggregated - using the estimated stability of each ranker to weigh their votes - to produce a final ranking and a final feature subset. The framework is named EFSIS (Ensemble Feature Selection Integrating Stability) and the source code is available on GitHub (https://github.com/zhxiaokang/EFSIS).

As benchmarks for our experiments, we tested our method on six cancer datasets coming from microarray experiments. To better understand its performance, we compared EFSIS with each of the methods aggregated in EFSIS and also with basic function perturbation. The result showed that the stability was highly improved by using EFSIS. Meanwhile, the prediction accuracy was also maintained well.

The rest of the article is organized as follows: Section II describes the proposed EFSIS framework, along with basic function perturbation, the individual feature selection methods, and the metrics applied to evaluate stability and prediction accuracy; Section III introduces the experimental study, including experimental settings and results; Section IV discusses the experiments and concludes the work.

II. Methods

A. Methodology of EFSIS

Our proposed ensemble feature selection framework includes two phases: data perturbation and function perturbation. The framework is illustrated in Figure 1.

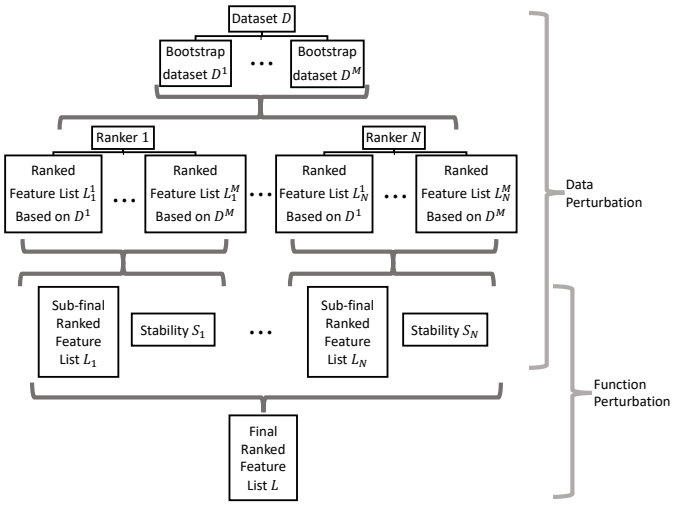

Fig. 1. The framework of EFSIS

Given the original dataset D, we use bootstrapping to get M perturbed variants of D ($\{D^1, ...D^m, ...D^M\}$) for the dataset D with p samples: we randomly draw p samples from D with replacement, allowing some samples to be picked multiple times while some samples may be absent in D^m. Each bootstrap dataset D^m is then passed to each of the included individual feature selection methods, each performing a ranking of all the features based on how well they distinguish samples from different groups. For simplicity, in the following, we call each feature selection method a *ranker*.

In the first phase which is data perturbation, let us take one ranker, ranker n ($n \in \{1, ...N\}$), as a general representative to explain the idea of data perturbation. Ranker n will rank the features based on the bootstrap datasets. Corresponding to each bootstrap dataset, one ranked list will be generated. Therefore, each ranker will end up with M ranked lists $\{L_n^1, ...L_n^m, ...L_n^M\}$. With an aggregation strategy (Equation (2) in Subsection 2.3), the M lists can then be combined into one list (L_n). In addition to L_n, a side product, the stability of ranker n, that we will denote as S_n, can be calculated using the stability definition described in Subsection 2.2: with a predefined threshold t, the top t features in L_n^m will be picked to constitute a feature subset, and then the M feature subsets will be used to calculate the stability of ranker n. The data perturbation procedure above will be applied to all N rankers to generate N sub-final ranked feature lists $\{L_1, ...L_N\}$.

In the phase of function perturbation, another aggregation strategy which integrates the stability of the rankers (Equation (3) in Subsection 2.3) combines those N sub-final ranked feature lists into one final list L. The top t features are kept as the selected important features by EFSIS.

B. Stability

A stable feature selection method should give similar feature subsets even given varying samples. We use the similarity between feature subsets derived from different sample sets to measure the stability of the corresponding feature selection method. We used the stability definition proposed by [5]:

$$S_n = \frac{\sum_{f \in F}(freq(f)/M)}{|F|} \tag{1}$$

Where S_n is the stability of a given feature selection method n; M is the number of feature subsets analyzed; F is the set of features that appear in at least one of the M subsets and $|F|$ indicates the cardinality of F; $freq(f)$ is the frequency of feature $f \in F$ that appears in those M subsets.

C. Aggregation strategies

There are two aggregations in the EFSIS paradigm shown in Figure 1. A very recent study [13] used intersection and union operations to aggregate the lists. But there is an extreme case where there is no intersection of the sub-lists. So we used another more robust strategy, rank product which is proposed by [14], to score each feature: the product of ranking positions of one feature in different ranked lists is used as its aggregated ranking score.

In data perturbation, the ranking score of a feature f from ranker n can be calculated as follows:

$$R_{f,n} = \prod_{m=1}^{M} R_{f,n}^m \tag{2}$$

where $R_{f,n}^m$ is the rank of feature f from ranker n on bootstrap set m. Based on this score, an aggregated ranked feature list L_n for ranker n can be obtained.

The function perturbation phase also applies the rank product aggregation strategy, but the stability of every ranker is

used as its weight. The ranking score of a feature f in the final ranked list can be calculated as follows:

$$R_f = \prod_{n=1}^{N} (R_{f,n})^{(1-S_n)} \quad (3)$$

where $1 - S_n$ is defined as the weight of ranker n, so that a more stable ranker is assigned a higher weight. Ranking the features based on this score, we get the final ranked list.

In fact, the basic function perturbation is a special case of the second phase in EFSIS: each ranker ranks the features based on the original dataset D, afterwards, it will apply the same rank product aggregation strategy, aggregating the rankings from different rankers in a similar way as EFSIS does in the second phase, except that there is no weight for each ranker ($S_n = 0$ in Equation (3)).

D. Individual feature selection methods

In general, there are three categories of feature selection methods: filter methods which rank the features only based on their correlation with the targeted classes, wrapper methods which use an objective function (can be the prediction accuracy obtained by the classifier using the selected features) to evaluate features, and embedded methods where the classifier itself performs feature selection.

Since one motivation of the ensemble framework is to make it as general as possible, we want to make it classifier-independent. Therefore, we consider only filter methods in this context.

In our experiment, we used four very diverse feature selection methods which are based on different sets of assumptions, to demonstrate the generality of the proposed framework. In particular, we employed both univariate techniques which treat the features as independent from each other and multivariate techniques which take the interaction between features into consideration.

As representatives of univariate techniques, we used:

- Significance Analysis of Microarrays (SAM) that was originally designed to identify genes with significantly differential expression in microarray experiments [15]. It assigns a score to each gene based on the change in gene expression relative to the standard deviation of repeated experiments.
- Information gain which is one of the most popular univariate methods [16]. It evaluates each feature based on the entropy concept from information theory.

As representatives of multivariate techniques, we applied:

- The Characteristic Direction method (GeoDE) which is a geometrical multivariate approach [17]. It defines a separating hyperplane using linear discriminant analysis to characterize the differential expression of microarray or RNA-Seq data.
- ReliefF [18] is an extension of the original Relief algorithm [19], [20] that evaluates a feature according to how well it can distinguish among instances that are near to each other. Compared to Relief, ReliefF is more robust to noisy and incomplete datasets.

E. Classification algorithm

In evaluating the predictive performance of the selected feature subsets, we applied the classification algorithm Support Vector Machine (SVM) [21] to learn a classifier based on the selected feature subsets. Provided with a training dataset of samples marked with group labels (samples are characterized by the selected features), SVM will learn an optimal hyperplane separating the samples from different groups. And the optimal hyperplane will be used to predict the labels of the samples from test set. A prediction accuracy can be calculated comparing the predicted labels with the true labels. A better feature subset will enable the SVM to achieve a higher prediction accuracy. For simplicity, we chose a linear kernel for SVM and we used Area Under Curve (AUC) [22] to summarize the obtained prediction accuracy.

III. Experimental study

A. Datasets

EFSIS was tested on six gene expression datasets produced using microarrays to study different forms of cancer (datasets were collected by [23]). The main characteristics of the datasets, including numbers of features and samples, are given in Table I. Feature selection can provide valuable information in such applications. The selected features can be regarded as biomarkers and they reflect characteristics of the studied cancer forms and can help to classify the patients. Feature selection can allow the cancer researcher or clinician to focus on a small number of biomarkers instead of thousands of features, which can save lots of money and time for further studies. Biomarkers can also help to improve the understanding of the cancer forms on a molecular level.

TABLE I
DATASETS USED IN THE EXPERIMENTS

Name	Features	Samples	Refs
AML	12 625	54	[24]
CNS	7 129	60	[25]
DLBCL	7 129	77	[26]
Prostate	12 600	102	[27]
Leukemia	7 129	72	[28]
ColonBreast	22 283	52	[29]

B. Experimental procedure and settings

To evaluate the performance of EFSIS, it was compared with the aggregated individual rankers and the corresponding basic function perturbation aggregating the same four rankers. The performance was evaluated in two aspects: stability and prediction accuracy. Both stability and prediction accuracy depend on how many features are to be selected and used for classification (denoted t), hence we performed the assessment with a range of values for t.

In order to obtain an unbiased estimation of performance, we performed the experiments using a ten-fold cross-validation scheme [30], [31]. Thus, we obtained 10 selected feature subsets for each pre-defined threshold t, for each dataset and

for each ranker. By doing classification analysis with those 10 feature subsets, we obtained 10 prediction accuracy scores. At the same time, by calculating the similarity of those 10 feature subsets using Equation (1), we obtained an estimate of the stability of the corresponding ranker.

Considering the highly variable number of features in each dataset (as shown in Table I), instead of using an absolute number of features t, we used a percentage of the original number of features. We explored a range of values from 0.3% to 5%.

The main parameters for EFSIS are the number of bootstrap datasets M and number of rankers N. M was chosen based on the recommendation in [9] ($M = 50$). In our analysis, $N = 4$, the rankers are described in Subsection II.D. The competitors of EFSIS would therefore be the four individual rankers and the basic function perturbation of the same four rankers.

C. Experimental results of stability performance

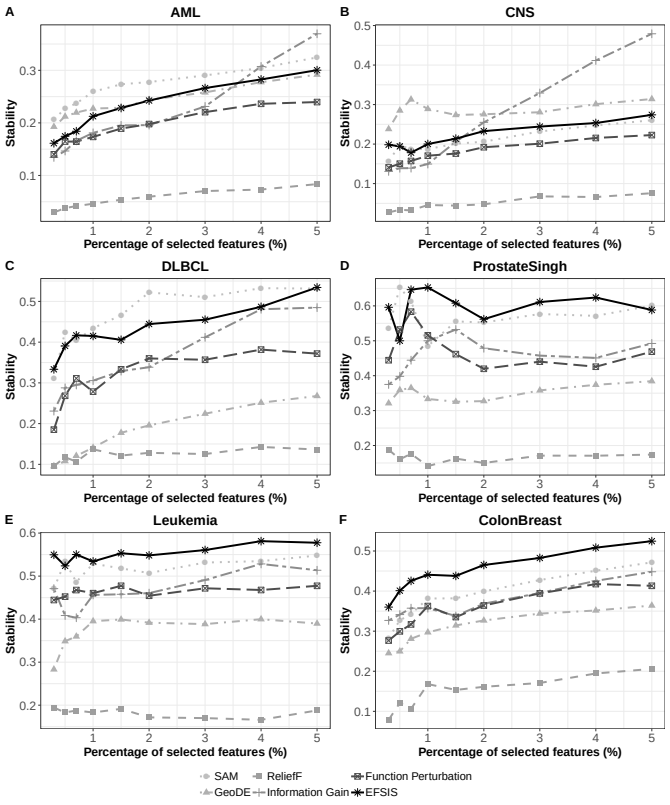

Fig. 2. Stability performance of six rankers on six datasets, tested with different percentages of selected features. For each dataset, four individual rankers (SAM, GeoDE, ReliefF, Information Gain), basic Function Perturbation, and EFSIS are considered.

In this section, we will study the stability of the rankers. The stability was tested on 6 datasets with 9 different percentages of selected features.

Figure 2 shows the performance of the four individual rankers and the two ensemble rankers. Let us firstly look at the individual ones. GeoDE has the same problem as in the previous section: it achieves a very high stability in the CNS dataset but a very low one in the DLBCL dataset. ReliefF seems to be a very unstable method with the lowest stability score across all datasets, even in the dataset DLBCL where it showed great predictive performance (as mentioned in the previous section).

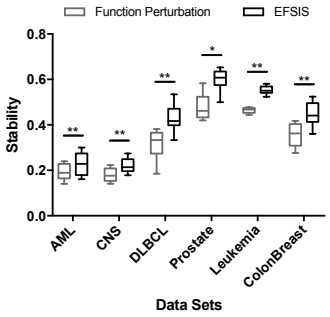

Fig. 3. Comparison of basic Function Perturbation and EFSIS in stability performance on six datasets. ** = P-value < 0.005, * = P-value < 0.01.

When we compare basic function perturbation with the four individual rankers across the 6 datasets as shown in Figure 2, we can find that basic function perturbation is either the second or the third worst one. In comparison, the performance of EFSIS is much more satisfactory: it is the second best one in the first 3 datasets (Figure 2 A-C), and it performs consistently better than all the individual rankers in the latter 3 datasets (Figure 2 D-F). If we compare between basic function perturbation and EFSIS, Figure 2 shows clearly that EFSIS performs always better than basic function perturbation. The box plot in Figure 3 shows the comparison between these two ensemble rankers on 6 datasets with the star ($*$) indicating the significance of difference (P-value was calculated using Wilcoxon Signed-Ranks Test [32]). We can see that the stability of EFSIS is significantly higher than basic function perturbation in all 6 datasets.

D. Experimental results of predictive performance

Even though stability is important, prediction accuracy cannot be ignored. The mean AUC (averaging the AUCs from ten-fold cross-validation) and associated standard deviation of four individual rankers and two ensemble ones (basic function perturbation and EFSIS) tested on 6 datasets with 9 different percentages of selected features are shown in Table II. For each combination of dataset and percentage of selected features, the best ranker (the one with the highest mean AUC and lowest standard deviation) is marked with dagger, and the ones that are significantly worse than the best one are marked with star and are in bold font (P-value < 0.05, Wilcoxon Signed-Ranks Test [32]). It shows a problem of the individual rankers: some individual rankers perform quite well in some datasets but poorly in some others. For example, GeoDE performs quite well in dataset CNS (it achieves the highest prediction accuracy among all rankers 7 times out of 9), but performs unsatisfactorily in dataset DLBCL (it achieves a significantly lower prediction accuracy than the best one 8 times out of 9, which makes it the worst for this dataset). But ReliefF

TABLE II

PREDICTIVE PERFORMANCE OF SIX RANKERS ON SIX DATASETS WITH DIFFERENT PERCENTAGES OF SELECTED FEATURES: MEAN AUC AND STANDARD DEVIATION.

Dataset	Ranker	Percentage of selected features (%)								
		0.3	0.5	0.7	1	1.5	2	3	4	5
AML	SAM	$\mathbf{0.69 \pm 0.17^*}$	0.73 ± 0.16	0.73 ± 0.20	0.76 ± 0.14	0.78 ± 0.17	0.75 ± 0.18	$\mathbf{0.74 \pm 0.20^*}$	0.76 ± 0.17	0.77 ± 0.16
	GeoDE	0.74 ± 0.16	0.69 ± 0.25	$0.76 \pm 0.20^\dagger$	$\mathbf{0.76 \pm 0.16^*}$	$0.80 \pm 0.16^\dagger$	$0.80 \pm 0.18^\dagger$	$0.84 \pm 0.15^\dagger$	0.79 ± 0.18	0.79 ± 0.22
	ReliefF	0.76 ± 0.21	0.73 ± 0.15	0.68 ± 0.21	$\mathbf{0.69 \pm 0.14^*}$	0.75 ± 0.15	0.76 ± 0.18	$\mathbf{0.72 \pm 0.14^*}$	0.74 ± 0.18	0.76 ± 0.15
	Info_Gain	$0.81 \pm 0.16^\dagger$	$0.75 \pm 0.18^\dagger$	0.74 ± 0.16	$\mathbf{0.73 \pm 0.17^*}$	0.79 ± 0.14	0.76 ± 0.17	0.77 ± 0.17	0.79 ± 0.16	0.80 ± 0.17
	Func_Pert	0.75 ± 0.16	0.73 ± 0.23	0.74 ± 0.15	$0.81 \pm 0.14^\dagger$	0.79 ± 0.13	0.79 ± 0.17	$\mathbf{0.75 \pm 0.21^*}$	$0.80 \pm 0.18^\dagger$	$0.81 \pm 0.14^\dagger$
	EFSIS	0.73 ± 0.20	0.74 ± 0.17	0.72 ± 0.22	0.75 ± 0.17	0.72 ± 0.18	0.73 ± 0.19	$\mathbf{0.77 \pm 0.16^*}$	0.75 ± 0.19	0.76 ± 0.17
CNS	SAM	0.72 ± 0.22	$\mathbf{0.69 \pm 0.22^*}$	0.71 ± 0.20	0.72 ± 0.17	$\mathbf{0.71 \pm 0.21^*}$	$\mathbf{0.72 \pm 0.17^*}$	$\mathbf{0.73 \pm 0.18^*}$	$\mathbf{0.69 \pm 0.19^*}$	$\mathbf{0.73 \pm 0.18^*}$
	GeoDE	0.63 ± 0.16	0.76 ± 0.08	$0.81 \pm 0.16^\dagger$	$0.82 \pm 0.13^\dagger$	$0.82 \pm 0.18^\dagger$	$0.88 \pm 0.17^\dagger$	$0.89 \pm 0.16^\dagger$	$0.88 \pm 0.14^\dagger$	$0.90 \pm 0.14^\dagger$
	ReliefF	0.69 ± 0.18	0.72 ± 0.14	0.75 ± 0.16	$\mathbf{0.68 \pm 0.15^*}$	0.74 ± 0.19	$\mathbf{0.70 \pm 0.19^*}$	$\mathbf{0.75 \pm 0.16^*}$	0.79 ± 0.21	$\mathbf{0.73 \pm 0.15^*}$
	Info_Gain	0.69 ± 0.17	$0.78 \pm 0.18^\dagger$	0.76 ± 0.19	0.71 ± 0.19	$\mathbf{0.65 \pm 0.17^*}$	$\mathbf{0.70 \pm 0.13^*}$	$\mathbf{0.66 \pm 0.18^*}$	$\mathbf{0.71 \pm 0.16^*}$	$\mathbf{0.78 \pm 0.15^*}$
	Func_Pert	0.72 ± 0.12	0.77 ± 0.18	0.68 ± 0.21	$\mathbf{0.68 \pm 0.21^*}$	0.77 ± 0.15	0.80 ± 0.21	$\mathbf{0.80 \pm 0.11^*}$	0.80 ± 0.13	$\mathbf{0.80 \pm 0.16^*}$
	EFSIS	$0.74 \pm 0.22^\dagger$	0.69 ± 0.16	0.68 ± 0.19	0.75 ± 0.15	0.79 ± 0.16	0.83 ± 0.14	0.82 ± 0.14	$\mathbf{0.78 \pm 0.11^*}$	$\mathbf{0.79 \pm 0.15^*}$
DLBCL	SAM	0.91 ± 0.13	0.90 ± 0.12	0.96 ± 0.08	0.96 ± 0.06	0.97 ± 0.07	0.97 ± 0.07	0.95 ± 0.11	0.94 ± 0.11	$0.97 \pm 0.07^\dagger$
	GeoDE	$\mathbf{0.86 \pm 0.10^*}$	$\mathbf{0.87 \pm 0.10^*}$	$\mathbf{0.86 \pm 0.12^*}$	$\mathbf{0.89 \pm 0.10^*}$	$\mathbf{0.88 \pm 0.16^*}$	$\mathbf{0.86 \pm 0.22^*}$	$\mathbf{0.85 \pm 0.22^*}$	$\mathbf{0.89 \pm 0.11^*}$	0.92 ± 0.10
	ReliefF	$0.96 \pm 0.08^\dagger$	0.94 ± 0.11	$0.99 \pm 0.03^\dagger$	0.96 ± 0.09	0.98 ± 0.06	0.94 ± 0.10	$0.99 \pm 0.03^\dagger$	$0.99 \pm 0.03^\dagger$	$0.97 \pm 0.07^\dagger$
	Info_Gain	0.95 ± 0.11	$0.95 \pm 0.09^\dagger$	0.95 ± 0.11	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.96 ± 0.08	0.97 ± 0.08	0.96 ± 0.06
	Func_Pert	0.91 ± 0.12	0.92 ± 0.09	0.96 ± 0.08	0.96 ± 0.06	$0.98 \pm 0.05^\dagger$	$0.98 \pm 0.05^\dagger$	0.96 ± 0.07	0.94 ± 0.10	0.93 ± 0.11
	EFSIS	0.92 ± 0.10	0.94 ± 0.08	$\mathbf{0.93 \pm 0.10^*}$	$0.97 \pm 0.06^\dagger$	0.97 ± 0.06	0.97 ± 0.06	0.96 ± 0.07	0.97 ± 0.07	$0.97 \pm 0.07^\dagger$
Prostate	SAM	$0.95 \pm 0.08^\dagger$	0.95 ± 0.08	0.95 ± 0.08	$0.96 \pm 0.06^\dagger$	0.96 ± 0.07	0.95 ± 0.07	0.96 ± 0.07	0.96 ± 0.07	$0.96 \pm 0.07^\dagger$
	GeoDE	0.90 ± 0.15	0.93 ± 0.09	0.94 ± 0.09	0.95 ± 0.08	0.95 ± 0.09	$\mathbf{0.94 \pm 0.08^*}$	0.95 ± 0.06	0.95 ± 0.06	0.96 ± 0.06
	ReliefF	0.94 ± 0.08	$0.96 \pm 0.08^\dagger$	$0.96 \pm 0.06^\dagger$	0.94 ± 0.10	0.97 ± 0.04	0.96 ± 0.06	0.94 ± 0.08	0.96 ± 0.07	0.94 ± 0.09
	Info_Gain	0.94 ± 0.11	0.94 ± 0.10	0.94 ± 0.10	0.95 ± 0.09	$\mathbf{0.95 \pm 0.09^*}$	0.96 ± 0.07	$0.97 \pm 0.06^\dagger$	$0.97 \pm 0.06^\dagger$	0.96 ± 0.08
	Func_Pert	0.95 ± 0.09	0.94 ± 0.10	0.95 ± 0.10	0.95 ± 0.09	0.96 ± 0.06	0.96 ± 0.07	0.96 ± 0.06	0.95 ± 0.08	0.95 ± 0.09
	EFSIS	0.95 ± 0.09	0.94 ± 0.10	0.95 ± 0.09	0.95 ± 0.08	$0.97 \pm 0.07^\dagger$	$0.97 \pm 0.07^\dagger$	0.95 ± 0.08	0.94 ± 0.09	0.94 ± 0.09
Leukemia	SAM	0.99 ± 0.04	0.98 ± 0.05	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$
	GeoDE	0.98 ± 0.05	$0.99 \pm 0.02^\dagger$	0.99 ± 0.04	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$
	ReliefF	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	0.99 ± 0.04	0.99 ± 0.04	$0.99 \pm 0.02^\dagger$	0.99 ± 0.04	0.98 ± 0.04	0.98 ± 0.05	0.98 ± 0.04
	Info_Gain	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$
	Func_Pert	0.97 ± 0.08	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$
	EFSIS	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$	$0.99 \pm 0.02^\dagger$
ColonBreast	SAM	0.98 ± 0.08	0.98 ± 0.08	0.97 ± 0.06	$0.98 \pm 0.05^\dagger$	$0.99 \pm 0.03^\dagger$	0.97 ± 0.07	0.97 ± 0.06	0.97 ± 0.06	0.97 ± 0.06
	GeoDE	$0.99 \pm 0.04^\dagger$	$0.99 \pm 0.04^\dagger$	$0.99 \pm 0.04^\dagger$	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.05
	ReliefF	0.95 ± 0.08	0.95 ± 0.12	0.95 ± 0.11	0.94 ± 0.12	0.98 ± 0.05	$1.00 \pm 0.00^\dagger$	0.97 ± 0.08	0.96 ± 0.08	0.97 ± 0.05
	Info_Gain	0.98 ± 0.05	0.95 ± 0.08	0.95 ± 0.08	0.95 ± 0.08	0.98 ± 0.08	0.98 ± 0.08	0.99 ± 0.04	0.98 ± 0.08	0.95 ± 0.12
	Func_Pert	0.98 ± 0.08	$0.99 \pm 0.04^\dagger$	0.98 ± 0.05	0.98 ± 0.05	0.97 ± 0.06	0.99 ± 0.03	$0.99 \pm 0.03^\dagger$	0.98 ± 0.05	0.98 ± 0.05
	EFSIS	0.98 ± 0.08	0.98 ± 0.08	$0.99 \pm 0.04^\dagger$	0.98 ± 0.05	0.96 ± 0.07	0.98 ± 0.05	0.98 ± 0.05	$0.99 \pm 0.04^\dagger$	$0.99 \pm 0.04^\dagger$

†The best ranker in one experiment (of one specific dataset and percentage of selected features).
*The rankers that are significantly worse than the best individual one.

performs contrarily to GeoDE in these two datasets. Since the performance of feature selection methods varies from dataset to dataset, it is difficult for researchers to choose an adequate one for their dataset. That problem is actually a big motivation for function perturbation since it can free researchers from that difficult decision. Ensemble rankers (function perturbation and EFSIS) will combine the results from all candidate rankers.

The results in Table II show that the predictive performance of ensemble rankers is more stable across the different datasets analyzed. Function perturbation and EFSIS are slightly better than the individual rankers: they are significantly worse than the best ranker in 4 out of the 54 experiments (6 datasets $\times$ 9 percentages of selected features), while four individual rankers are worse in 8, 10, 6, 7 experiments, respectively.

IV. DISCUSSION AND CONCLUSIONS

We have described a new framework for ensemble feature selection, which combines data perturbation and function perturbation and utilizes the stability of the individual methods as weights. The new framework utilizes data perturbation's ability to improve stability to solve the low-stability issue of function perturbation. It possesses the advantages of both function perturbation and data perturbation: it combines the results from different individual feature selection methods and shows robust predictive performance, and it also provides more stable selected feature subsets. Therefore, it frees the researchers from choosing the most suitable feature selection method for their datasets. Also, compared to basic function perturbation, it provides higher stability. To be noted, EFSIS is a framework, meaning that researchers can put whatever they like in the framework. For example, they can replace the individual rankers with some specific ones that are commonly used in their research field or add new ones as more and more feature selection methods are being proposed.

A major shortcoming of EFSIS, however, is that it is more time-consuming and more computationally expensive compared to the other methods assessed here. However, it can be sped up by parallel computing. The parallelization can be done in multiple ways. What we have tried was to split the jobs by bootstrap datasets so that the job corresponding to one dataset was performed by one node. Parallelization can considerably shorten the computing time, but depends on available computing resources.

To our knowledge, our work is the first study exploring the stability of function perturbation and the combination of function and data perturbation. It can form the basis for further studies in this direction. In the EFSIS framework, we have chosen to perform data perturbation in the first phase so that each ranker (feature selection method) is performed on all bootstrap datasets to produce one ranking that is next combined with rankings from the other rankers. In this way we can obtain the stability of each individual ranker based on the same subsets of samples, enabling us to use the stability estimates when combining results across the rankers. However, it would be interesting to explore an alternative approach where function perturbation is applied to each bootstrap dataset, which will produce M ranked lists. In the next step, these M lists will

be combined (using for example rank product) to obtain the final ranked list. The idea behind this strategy is to make use of data perturbation's ability to improve the stability of function perturbation. Future studies will include this and other directions.

ACKNOWLEDGMENT

We would like to thank Computational Biology Unit at University of Bergen, where the work was carried out. We would like to thank the colleagues in Jonassen Group for helpful discussions, especially Fatemeh Zamanzad Ghavidel. The research has been done within the project dCod 1.0, funded by the Digital Life Norway initiative of the BIOTEK 2021 program of the Research Council of Norway (project no. 248840).

REFERENCES

[1] O. Sagi, L. Rokach, Ensemble learning: A survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8 (4) (2018) e1249. doi:10.1002/widm.1249.

[2] Z. He, W. Yu, Stable feature selection for biomarker discovery, Computational Biology and Chemistry 34 (4) (2010) 215–225. doi:10.1016/J.COMPBIOLCHEM.2010.07.002.

[3] V. Boln-Canedo, A. Alonso-Betanzos, Ensembles for feature selection: A review and future trends, Information Fusion 52 (2019) 1–12. doi:10.1016/J.INFFUS.2018.11.008.

[4] C. A. Davis, F. Gerick, V. Hintermair, C. C. Friedel, K. Fundel, R. Kuffner, R. Zimmer, Reliable gene signatures for microarray classification: assessment of stability and performance, Bioinformatics 22 (19) (2006) 2356–2363. doi:10.1093/bioinformatics/btl400.

[5] F. R. Bach, F. R., Bolasso: model consistent Lasso estimation through the bootstrap, in: Proceedings of the 25th international conference on Machine learning - ICML '08, ACM Press, New York, New York, USA, 2008, pp. 33–40. doi:10.1145/1390156.1390161.

[6] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (3) (2010) 392–398. doi:10.1093/bioinformatics/btp630.

[7] B. Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, A. Alonso-Betanzos, Ensemble feature selection: Homogeneous and heterogeneous approaches, Knowledge-Based Systems 118 (2017) 124–139. doi:10.1016/J.KNOSYS.2016.11.017.

[8] B. Pes, N. Dessì, M. Angioni, Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data, Information Fusion 35 (2017) 132–147. doi:10.1016/j.inffus.2016.10.001.

[9] N. C. Tan, W. G. Fisher, K. P. Rosenblatt, H. R. Garner, Application of multiple statistical tests to enhance mass spectrometry-based biomarker discovery, BMC Bioinformatics 10 (1) (2009) 144. doi:10.1186/1471-2105-10-144.

[10] A. Ben Brahim, M. Limam, Robust ensemble feature selection for high dimensional data sets, 2013 International Conference on High Performance Computing & Simulation (HPCS) (2013) 151–157doi:10.1109/HPCSim.2013.6641406.

[11] S. Ahmed, M. Zhang, L. Peng, Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming Improving feature ranking for biomarker discovery in proteomics mass spectrometry data using genetic programming, Connection Science 26 (3) (2014) 215–243. doi:10.1080/09540091.2014.906388org/10.1080/09540091.2014.906388.

[12] K.L. Chiew, C.L. Tan, K. Wong, K.S.C. Yong, W.K. Tiong, A new hybrid ensemble feature selection framework for machine learning-based phishing detection system, Information Sciences 484 (2019) 15–166. doi:10.1016/J.INS.2019.01.064.

[13] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Letters 573 (1-3) (2004) 83–92. doi:10.1016/j.febslet.2004.07.055.

[14] V. Goss Tusher, R. Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proceedings of the National Academy of Sciences 98 (9) (2001) 5116—-5121.

[15] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2016.

[16] N. R. Clark, K. S. Hu, A. S. Feldmann, Y. Kou, E. Y. Chen, Q. Duan, A. Ma'ayan, The characteristic direction: a geometrical approach to identify differentially expressed genes, BMC Bioinformatics 15 (1) (2014) 79. doi:10.1186/1471-2105-15-79.

[17] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF, in: European conference on machine learning, Springer, Berlin, Heidelberg, 1994, pp. 171–182.

[18] K. Kira, L. A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of the tenth national conference on Artificial intelligence, AAAI Press, 1992, pp. 129–134.

[19] K. Kira, L. A. Rendell, A Practical Approach to Feature Selection, Machine Learning Proceedings 1992 (1992) 249–256doi:10.1016/B978-1-55860-247-2.50037-1.

[20] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297. doi:10.1007/BF00994018.

[21] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874. doi:10.1016/J.PATREC.2005.10.010.

[22] N. Lazzarini, J. Bacardit, RGIFE: a ranked guided iterative feature elimination heuristic for the identification of biomarkers, Lazzarini and Bacardit BMC Bioinformatics 18. doi:10.1186/s12859-017-1729-2.

[23] T. Yagi, A. Morimoto, M. Eguchi, S. Hibi, M. Sako, E. Ishii, S. Mizutani, S. Imashuku, M. Ohki, H. Ichikawa, Identification of a gene expression signature associated with pediatric AML prognosis, Blood 102 (2003) 1849–1856. doi:10.1182/blood-2003-02-0578.

[24] 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, T. R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (6870) (2002) 436–442. doi:10.1038/415436a.

[25] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. 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, T. R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nature Medicine 8 (1) (2002) 68–74. doi:10.1038/nm0102-68.

[26] 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, W. R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209. doi:10.1016/S1535-6108(02)00030-2.

[27] 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, E. S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring., Science (New York, N.Y.) 286 (5439) (1999) 531–7. doi:10.1126/SCIENCE.286.5439.531.

[28] D. Chowdary, J. Lathrop, J. Skelton, K. Curtin, T. Briggs, Y. Zhang, J. Yu, Y. Wang, A. Mazumder, Prognostic Gene Expression Signatures Can Be Measured in Tissues Collected in RNAlater Preservative, The Journal of Molecular Diagnostics 8 (1) (2006) 31–39. doi:10.2353/JMOLDX.2006.050056.

[29] D. D. Jensen, P. R. Cohen, Multiple Comparisons in Induction Algorithms, Machine Learning 38 (3) (2000) 309–338. doi:10.1023/A:1007631014630.

[30] I. Tsamardinos, E. Greasidou, G. Borboudakis, Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation, Machine Learning 107 (12) (2018) 1895–1922. doi:10.1007/s10994-018-5714-4.

[31] J. Demšar, Statistical comparisons of classifiers over multiple data sets, Journal of Machine learning research 7 (Jan) (2006) 1–30.