

Ensemble Feature Selection Compares to Meta-analysis for Breast Cancer Biomarker Identification from Microarray Data

Bernardo Trevizan[1,2] and Mariana Recamonde-Mendoza[1,2(✉)]

[1] Institute of Informatics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil

[2] Bioinformatics Core, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil

{btrevizan,mrmendoza}@inf.ufrgs.br

Abstract. Identifying stable and precise biomarkers is a key challenge in precision medicine. A promising approach in this direction is exploring omics data, such as transcriptome generated by microarrays, to discover candidate biomarkers. This, however, involves the fundamental issue of finding the most discriminative features in high-dimensional datasets. We proposed a homogeneous ensemble feature selection (EFS) method to extract candidate biomarkers of breast cancer from microarray datasets. Ensemble diversity is introduced by bootstraps and by the integration of seven microarray studies. As a baseline method, we used the random effect model meta-analysis, a state-of-the-art approach in the integrative analysis of microarrays for biomarkers discovery. We compared five feature selection (FS) methods as base selectors and four algorithms as base classifiers. Our results showed that the variance FS method is the most stable among the tested methods regardless of the classifier and that stability is higher within datasets than across datasets, indicating high sample heterogeneity among studies. The predictive performance of the top 20 genes selected with both approaches was evaluated with six independent microarray studies, and in four of these, we observed a superior performance of our EFS approach as compared to meta-analysis. EFS recall was as high as 85%, and the median F1-scores surpassed 80% for most of our experiments. We conclude that homogeneous EFS is a promising methodology for candidate biomarkers identification, demonstrating stability and predictive performance as satisfactory as the statistical reference method.

1 Introduction

Precision medicine aims at tailoring health care testing and medication to each patient according to their individual characteristics. It has become feasible with the growth of digital medical records and high-throughput diagnosis devices [16].

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

© Springer Nature Switzerland AG 2021

O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12949, pp. 162–178, 2021.
https://doi.org/10.1007/978-3-030-86653-2_12

Since the first sequenced human genome, more recent technologies, such as microarray, have helped to understand diseases at the molecular level through the generation of omics data. Microarray technology allows the characterization of the gene expression profile for all genes within an organism, also called transcriptome, which is especially useful when expressions are compared among different groups, such as a disease and a control group.

One of the key applications of transcriptome data for precision medicine is the identification of candidate biomarkers. Biomarkers are measurable parameters that act as indicators of the presence or severity of a disease. The exploration of biomarkers for detecting cancer or selecting the most appropriate therapy for a patient is one of the main branches of precision medicine research [17]. Nonetheless, harnessing the knowledge of high-throughput omics data requires advanced computational methods such as machine learning (ML) algorithms to identify molecular patterns in different patients that may be used to infer hypotheses about diagnostic biomarkers.

As discussed by [15], a fundamental property of biomarkers is stability. Stable biomarkers refer to a set of consistent markers identified from sampling variations. In other words, a biomarker is stable and thus representative of the population if the same can be reproduced in other similar subpopulations. Thus, the integration of multiple transcriptomic datasets has been advocated to improve biomarkers discovery [30]. In addition, biomarkers should have good discriminative power, meaning that based on their application, we can accurately distinguish tumor from non-tumor samples, for instance, if considering diagnostic biomarkers. The identification of stable and precise biomarkers is one of the critical challenges in transcriptome analysis in the scope of precision medicine. One of the state-of-the-art approaches for such investigation is statistical meta-analysis. Several meta-analysis methods exist, and they combine the primary statistics drawn from individual analysis of each experiment, such as p-values or effect sizes (*i.e.,* fold change) [30].

In the field of ML, biomarkers discovery has been frequently approached as a dimensionality reduction problem, for which feature selection (FS) methods may be naturally applied [4]. A plethora of FS methods exist; nonetheless, the high correlation among features, as frequently observed in omics data, leads to instability among results from multiple FS methods or from multiples runs over distinct training samples, reducing the confidence of selected features [19]. One of the proposed solutions to improve the stability of FS on high-dimensional data is ensemble feature selection (EFS). Ensemble learning principles come from the idea of "Wisdom of Crowds," which states that a group of diverse individuals are collectively smarter than any single individual, even expert ones [27]. In ML, ensembles have been explored to produce more robust and accurate learning solutions. In the context of FS, EFS methods use various base selectors and aggregate their results, aiming to obtain a more stable feature subset. Diversity among opinions, an essential property of ensemble learning, is achieved either by using distinct FS methods as base selectors (*i.e.,* function perturbation) or by using sampling variations across base selectors (*i.e.,* data perturbation) [7]. These approaches generate, respectively, heterogeneous and homogeneous ensembles.

In the present work, we aim at exploring homogeneous ensemble FS to identify candidate biomarkers for breast cancer from multiple microarray studies and compare its results to traditional meta-analysis. Our work is motivated by the large volume of transcriptomic data publicly available and the need to jointly analyze different gene expression datasets to derive more stable and precise hypotheses about candidate biomarkers. We believe that homogeneous EFS is as promising as the well-established meta-analysis approaches to extract the most relevant genes for breast cancer diagnosis. Moreover, to the best of our knowledge, there were limited efforts so far towards the comparison of stability and predictive performance among these different analytical approaches.

2 Related Works

The literature regarding EFS has expanded recently, and several previous works have applied homogeneous EFS to high-dimensional data, including microarray. [22] evaluated stability on EFS methods across several domains, including biomedical data, demonstrating that homogeneous EFS approaches lead to a significant gain of stability. An earlier study [23] exploited the performance and stability of homogeneous ensembles applied to high-dimensional omics data. In every case, the EFS approach outperformed other methods. The authors achieved a high accuracy with 3% of selected features and a stability increase with homogeneous EFS compared to other methods. In [1], homogeneous EFS showed the best stability and improvement in performance on identifying biomarkers for cancer diagnosis from microarray datasets. Authors evaluate their approach for four types of cancer (Leukemia, colon, lymphoma, and prostate), using the traditional recursive feature elimination (RFE) as the baseline FS method.

In [5], authors designed an ensemble in which, for each base selector's output, a classifier was trained. The final predictions were combined by majority voting among classifiers. The heterogeneous ensemble approach yielded a stability index of 0.229 for breast cancer data and a predictive error of 28.11%. Using the same ensemble design, [21] achieved an error of 3.09%. However, in the latter, the authors grouped genes using information theory in which a gene is relevant if it is correlated to classes and not to other selected genes. The ensemble selects one gene for each group of highly correlated genes in order to reduce redundancy. The authors also reported an increase in stability compared to other methods. In [3], authors validated the results of genes selected from EFS algorithms with literature meta-analysis results. The authors were able to identify 100 microRNAs (*i.e.,* small non-coding RNAs) that could explain 29 types of cancer, most of which had previous evidence from the literature. Besides, reported accuracies are higher than 90% in all experiments.

In [25], the authors used seven datasets with different FS methods to improve training time and increase accuracy. The best results presented for microarray dataset used a homogeneous ensemble with SVM-RFE and mean aggregation of rankings. According to authors, "[...] *an ensemble approach would seem to be the most reliable approach to feature selection*" [25]. The authors in [6] reviewed the

performance of state-of-the-art FS methods across several domains of microarray - breast, prostate, brain, colon, and ovarian cancer. The authors concluded that, in general, performance depends on the FS method, on the classifier, and, mainly, on the problem domain. This emphasizes the importance of EFS methods, which partially alleviates the problem of lack of robustness for single FS methods. We note, however, that none of these works have compared EFS to meta-analysis approaches for dimensionality reduction in microarray data. In the approach presented by [26], authors combined ML with meta-analysis in a sequential manner, in which potential markers identified for mastitis disease in cattle using meta-analysis were further analyzed with decision trees to refine the hypotheses about most relevant features. Their approach was able to identify four biological markers with 83% accuracy in classification; nonetheless, their methodological and biological goals differ from ours. In fact, efforts towards comparing EFS and meta-analysis methodologies for identifying candidate biomarkers are still scarce in the literature.

3 Materials and Methods

3.1 Breast Cancer Microarray Datasets

Microarray datasets containing both tumor and non-tumor (*i.e.,* "normal") samples were collected from the Gene Expression Omnibus (GEO) database. A total of 11 studies were downloaded and pre-processed following the standard bioinformatics pipeline for each manufacturer (Affymetrix or Agilent). Due to limitations in the evaluation process, only datasets containing more than 10 samples in each class were selected as training datasets, defined as $T = \{T_1, T_2, T_3, ..., T_7\}$ (Table 1). The remaining datasets were assign for results' evaluation, defined as $E = \{E_1, E_2, E_3, ..., E_6\}$ (Table 2). To avoid redundancy on gene identification for later comparison of gene sets, all genes were mapped to their intrinsic *Entrez ID* using the R package *biomaRt* [11]. Furthermore, only genes common to all studies were kept for further analyses.

Table 1. Datasets used for training the EFS method.

	Dataset	Number of samples			Tumor/
		Tumor	Normal	Total	Total ratio
T_1	GSE38959	30	13	43	0.70
T_2	GSE42568	98	17	115	0.85
T_3	GSE45827	122	36	158	0.77
T_4	GSE53752	46	21	67	0.69
T_5	GSE62944	1119	113	1232	0.91
T_6	GSE70947	148	148	296	0.50
T_7	GSE7904	42	18	60	0.70

Table 2. Datasets used for testing the EFS method.

	Dataset	Number of samples			Tumor/
		Tumor	Normal	Total	Total ratio
E_1	GSE10797	27	5	32	0.84
E_2	GSE22820	74	10	84	0.88
E_3	GSE26304	109	6	115	0.95
E_4	GSE57297	25	7	32	0.78
E_5	GSE61304	57	4	61	0.93
E_6	GSE71053	6	12	18	0.33

3.2 Feature Selection Methods

Several FS methods have been proposed to reduce dimensionality. Here, we only considered ranker-based methods, *i.e.,* those that provide as output a ranking of all features ordered by a given score. The completeness in the output of these methods facilitates ranking aggregation among multiple selectors. We focused exclusively on filter methods among ranker-based methods, as they present the best trade-off between predictive and computational performance. This is important when working with high-dimensional data, as well as with homogeneous EFS since it involves multiple applications of the FS methods [22]. In this paper, the following FS methods were explored:

- *Information Gain* (IG) [18] explores the concept of entropy, which quantifies the uncertainty of each feature in the process of decision making. Entropy is used to measure the impurity of information, which in turn allows estimating the relevance of each feature by evaluating the extent to which impurity (and thus uncertainty) decreases when the value of that feature is known.
- Chi squared (ϕ_c) [10] measures the correlation between two random variables A and B based on the χ^2 statistical test. The larger the χ^2, the more important the feature is for the predictive task of interest.
- *Symmetrical Uncertainty* [28] is also based on the information-theoretical concept of entropy. It modifies the IG by introducing a proper normalization factor.
- *Minimum Redundancy Maximum Relevance (mRMR)*, defined by [13], best scores variables with higher correlation to the target value and lower correlation among other variables. For such, it takes the IG between the response variable A given $B \in V$, where V is the set of variables, and subtracts the pairwise IG between other variables and B.
- *Variance* measures the spread of a variable's values from its mean. As a statistical summary function, variance can be expressed as the average squared distances of A from its mean value μ_A.

3.3 Ensemble Feature Selection Design

The idea underlying EFS is to use various base selectors and aggregate their results aiming at more stable and accurate feature subsets. The design of an efficient EFS involves five main decisions: (i) type of base selectors; (ii) number of base selectors; (iii) number and size of different training sets; (iv) aggregation method, and (v) threshold methods [7]. In what follows, we discuss these details regarding the proposed solution.

We implemented a homogeneous EFS in this work, meaning that all base selectors employ the exact same FS method, and diversity among base selectors is introduced at the data level. The FS methods tested and compared in our experiments were listed in Sect. 3.2. In our framework, data diversity is generated in two ways. First, multiple microarray datasets are analyzed in parallel, each of which derives from a distinct study/population of patients with breast cancer (Table 1). Second,

we apply data perturbation upon sampling variations for each microarray dataset using the bootstrap method. Given a dataset with n instances, this method randomly selects n instances from the original dataset with replacement to generate a data sample, also called a *bag*. This sampling process is repeated B times, a parameter of our framework, and each bag is analyzed by the chosen FS method to generate a list of features ranked by relevance. We emphasize that only one type of FS method is applied over all bags for a single experiment. We tested a different number of bags (*i.e.,* number of base selectors) per microarray dataset, from 5 to 100, which will be further explained in Sect. 4.

After running all base selectors, multiple rankings are generated and must be aggregated into a consensus ranking. These consensus rankings are referred to as *local rankings* of our framework, as they reflect the consensus among B bags for a given microarray dataset. Many aggregation strategies have been proposed in the literature and were compared by previous works. According to [25], the best aggregation method for microarray datasets is the arithmetic mean, which takes the average of the relevance values yielded by the multiple rankings to reorder features into a consensus ranking. Thus, we adopted mean aggregation to generate a local ranking for each microarray study. Once each study has been individually analyzed, the multiple local rankings (*i.e.,* seven at total) are aggregated into a *global ranking*, which represents our framework's output. We note that both local and global rankings comprise the complete set of features in the datasets. An EFS-based subset of highly discriminative features is obtained by applying a threshold to keep only the top K most relevant features.

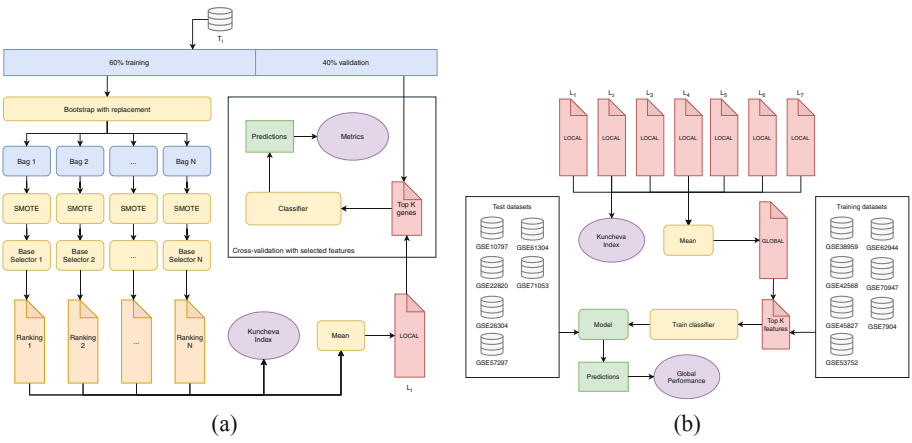


Fig. 1. Methodology applied for building and evaluating the proposed EFS approach. a) Homogeneous EFS is applied to each microarray dataset, generating a local ranking. b) The local rankings extracted from the multiple datasets are aggregated into a global ranking.

3.4 Ensemble Feature Selection Evaluation

The main goal in EFS is to improve the stability and predictive performance of the selected feature subset. According to [22], "*stable but not accurate solutions would not be meaningful; on the other hand, accurate but not stable results could have limited utility*". Thus, our methodology for evaluating FS results adopts both criteria, the stability of the selected features and the predictive performance of the classification models built exclusively with this subset of features. Our methodology to build and evaluate a homogeneous EFS is summarized in Fig. 1. For each training dataset $T_i \in T$, as summarized in Table 1, we created an ensemble following the aforementioned explanation.

To assess the predictive performance of the features subset, we applied a combination of stratified k-fold cross-validation (CV) and holdout to train and test classifiers using the selected top K features. For each iteration of k-fold CV, we split the training data (*i.e.,* training folds) into two parts, one for training (60%) and the other for testing (40%) the classifier, as shown in Fig. 1a. The training part is resampled to perturb data for B bags. B is also the number of base selectors, one for each bag, all running a previously chosen FS method. Since our datasets are characterized by class imbalance, we adopted SMOTE as an over-sampling method to deal with this issue. SMOTE is applied in each bag separately prior to FS analysis, and in the training partition to build the classifiers based on the top K features from the generated local ranking (*i.e.,* mean aggregation of the B rankings). Our EFS approach outputs three objects: (i) local stability for the set of rankings; (ii) final aggregated ranking, also called as local ranking L_i; and (iii) local performance metrics for the top K genes using 5-fold cross-validation.

Applying the basic homogeneous EFS approach to each microarray study only allows the assessment of the stability or predictive performance based on the specific dataset presented to the classifier, *i.e.,* T_i. However, our interest lies in the stability and performance across several microarray studies. Therefore, we also evaluate stability across the set of local rankings L_i for $i = \{1, 2, 3..., 7\}$ generated for the training datasets (Fig. 1b). Moreover, the local rankings are aggregated into a unique global ranking by the arithmetic mean, and the global ranking is finally used to select the EFS-based subset of most relevant features by applying a threshold K (*i.e.,* retaining the top K features). The predictive performance of the proposed homogeneous EFS approach is evaluated by training a classifier using this subset of EFS-based top K features and the union of datasets $T_i \in T$, and further evaluating its generalization power with the independent test datasets $E_i \in E$ (Table 2). Since these datasets contain instances never seen by the classifier, they allow an unbiased estimate of predictive performance. After studies aggregation, the output of our framework is the global stability, global ranking, and global performance metrics.

3.5 Evaluation Metrics

As aforementioned, two evaluation criteria are adopted in our framework: stability and predictive performance. The stability validation step calculates the

Kuncheva Index (Eq. 2) [20], which is the average of pairwise inconsistency indexes between a set of features subsets $\mathcal{S} = \{S_1, S_2, S_3, ..., S_N\}$. An inconsistency index, according to [20], increases proportionally to the intersection's cardinality $r = |A \cap B|$ between two subsets A and B with the same cardinality $|A| = |B| = k$. The maximum value 1 is achieved when $A = B$, and the minimum value is limited to -1. Equation 1 mathematically defines the inconsistency index where $A, B \subset X$ and $|X| = n$. The author also defines a threshold for high and low stability. The stability can be considered high when $\mathcal{I}(\mathcal{S}) \geq 0.5$, and low otherwise.

$$I(A, B) = \frac{rn - k^2}{k(n - k)} \tag{1}$$

$$\mathcal{I}(\mathcal{S}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=1}^{N} I(S_i, S_j) \tag{2}$$

The predictive performance is estimated by training a classifier solely with the most relevant features provided by the top K features. The top K features are extracted from the local ranking when our goal is to evaluate performance for a single microarray dataset or from the global ranking when we aim to carry out this analysis across multiple datasets. The classifier is evaluated using the independent testing sets to compute the precision, recall, and F1-Score in both cases. *Precision* defines the ratio of true positive (TP) predictions over all positive predictions made by the model, which is the sum of TP and false positives (FP). *Recall*, or sensitivity, defines the ratio of TP predictions over all samples labeled as positive, that is, TP and false negatives (FN). *F1-Score* ($F1$) defines the harmonic mean between precision and recall.

3.6 Classification Algorithms

To evaluate the effectiveness of potential feature sets for classification of breast cancer samples, we estimate the performance using the following supervised learning algorithms as basis for training the classifiers:

- *Decision Tree* (J48) classification algorithm learns decision rules inferred from training data to predict a target variable [8]. J48 iteratively selects the feature that best splits the subset (training data for the first iteration) according to a homogeneity metric, such as information gain.
- *K-Nearest Neighbors* (KNN) is an instance-based algorithm [2] that computes classification through a majority vote among the K nearest neighbors of each new data point. We adopted Euclidean distance as the distance measure.
- *Neural Network* (NNET) uses a multi-layered perceptron algorithm that trains a model using batch gradient descent [14]. Each input neuron represents a feature for a given sample. For training, a weight is assigned for each connection between two neurons. Therefore, a neuron's a_i value is the weighted sum of previous neurons directly connected to a_i. The values are

propagated toward the output layer where the prediction error is calculated. The error is applied on the weights update in the opposite direction using gradient descent and a learning rate α, *i.e.*,, by error *back-propagation*. The model classifies new samples by choosing the classes whose output neuron has the highest probability.

– *Support Vector Machine* (SVM) creates a hyper-plane or a set of hyper-planes that is farthest from the nearest training samples [9]. In this work, a non-linear SVM with a radial kernel is applied. Non-linear SVM applies a kernel function in order to expand the input space into a feature space to make the problem linearly separable.

Experiments were run in R environment, using the *caret* R package to train and evaluate models. Hyperparameter tuning was performed for each algorithm using the package's standard grid set of candidate models. By default, caret automatically chooses the hyperparameters associated with the best performing model.

3.7 Meta-analysis

Meta-analysis represents another viable approach for candidate biomarker identification, being a state-of-the-art method in this type of investigation. In a broad sense, meta-analysis aims at integrating the findings (*i.e.,* primary statistics) from multiple studies to extract more robust hypotheses. The first step of meta-analysis measures the effect size - the strength of a phenomenon - through differential expression analysis (DEA). Here we adopted the LIMMA [24], especially designed for transcriptomics, to extract genes' log fold change (*logFC*) and *p-value* for each study. The *logFC* is simply the difference between the average of expressions of tumor cases and the average of expressions of normal cases. The *p-value* evaluates the significance of a change in expressions between the two groups. In the second step, the meta-analysis method *per se* is applied, which integrates the primary statistics from individual studies to assess the significance of each gene across all populations. In this work, the focus will be on methodologies based on effect size combination due to their performance in meta-analysis for microarray datasets [29]. We used the Random Effect Model (REM) through the *MetaVolcano* R package, which outputs a ranking of genes from the most significant to the least significant to distinguish tumor from normal samples.

As one of ours goals is to compare EFS with meta-analysis methods, we evaluate the results from meta-analysis with a pipeline similar to the one adopted to EFS (Fig. 2). However, instead of the local ranking for a dataset T_i, meta-analysis calculates the effect size for each gene through DEA and the REM generates the global ranking. In this way, we assure a fair comparison between EFS and REM by applying the same data and executing the same evaluation procedure for both methods.

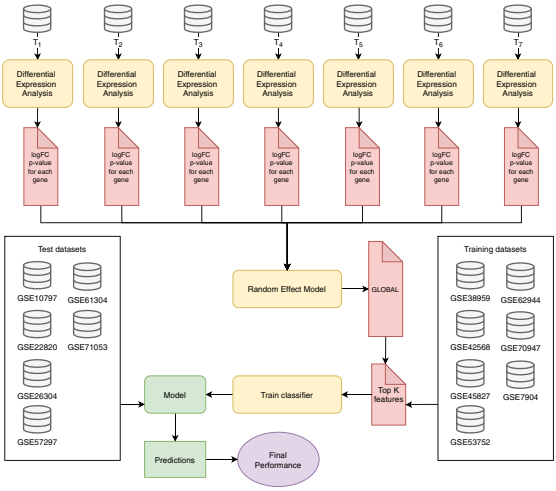


Fig. 2. Evaluation pipeline adopted for the REM meta-analysis method.

4 Results

In what follows we present the results for the homogeneous EFS applied to the selected breast cancer microarray studies in terms of stability and predictive performance. The section ends by discussing the results of the comparative analysis between EFS and the REM meta-analysis. All results are based on the average or values distribution for 5-fold cross-validation.

Our experiments were focused on evaluating every possible combination of the following parameters (and respective values): **Threshold**: 5, 10, 15, 20, 25, 30, 50 75, 100, 150, 200, 250, 500; Number of Bags: 5, 10, 25, 50, 100; **Classifier**: SVM, J48, KNN, NNET; **Base selector**: IG, ϕ_c, SU, mRMR, variance.

4.1 Stability Analysis

High stability for biomarker identification in EFS methods means that the most informative genes for one population are also the most informative for another, *i.e.,* it reflects the reproducibility of the selected genes across the populations. The ideal scenario is stability close to 1 when feature selection methods across different populations choose almost all the same genes. In general, the bigger the subset of selected genes, the higher the stability. Figure 3 shows the expected behavior for global stabilities, with values increasing for larger thresholds. Note that global stabilities, in general, are low, indicating a poor overlap of hypotheses driven from different populations. This may be due to heterogeneity across studies' samples. On the other hand, the stability for the top 5 genes does not follow the trend and is slightly higher than stability achieved until the top 20 genes. This suggests that local ensembles are most agreeing in rankings' top positions.

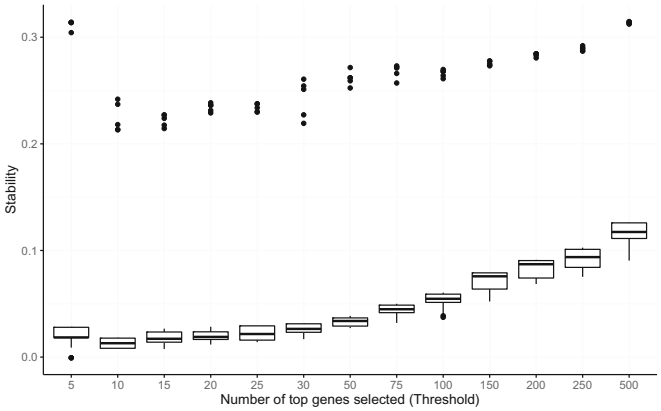


Fig. 3. Stability for different number of genes selected from the global ranking.

To investigate the outliers observed in Fig. 3, we analyzed results separately by base selectors and by the number of bags. Breaking by base selectors (Fig. 4a), it is clear that *variance* presents overall higher stability in comparison to other base selectors, which explains the mentioned outliers. In contrast, the number of bags (Fig. 4b) has an insignificant impact on the overall global stability. Therefore, stability is mainly impacted by the selected genes' subset size and the base selector bias. Since we want to select a minimum set of informative genes - the candidate biomarkers - high stability lies in the choice of the base selector.

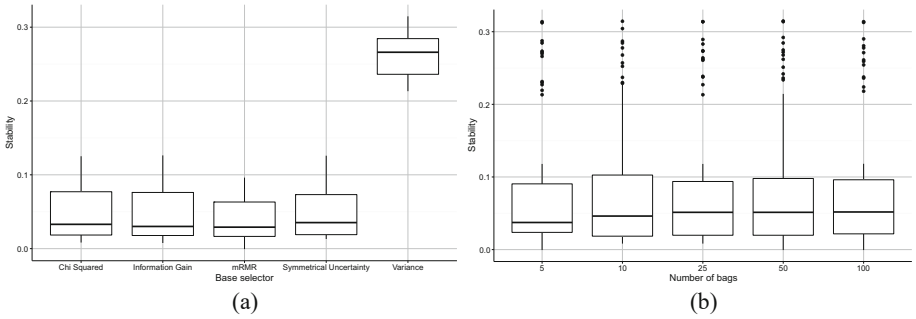


Fig. 4. Analysis of stability for a) each base selector and b) different number of bags.

In GEO, different datasets are deposited from studies within distinct populations. In this work, we will not address the biological implications and challenges of different data sources and sample heterogeneity. However, when stabilities are

compared locally within each training dataset (Fig. 5) based on bootstrap samples, the stabilities are much higher. We can see that, once again, variance clearly achieves higher stabilities than other base selectors. Furthermore, the GSE62944 dataset presents higher stability for every base selector in comparison with other datasets. We could hypothesize that a positive correlation exists between the number of samples and stability since this dataset has 1.232 samples while the other datasets have 123 ± 86 samples on average. However, this assumption needs to be validated in further experiments.

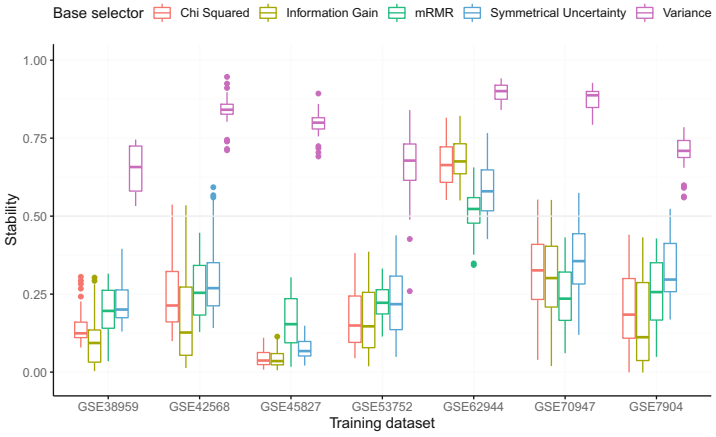
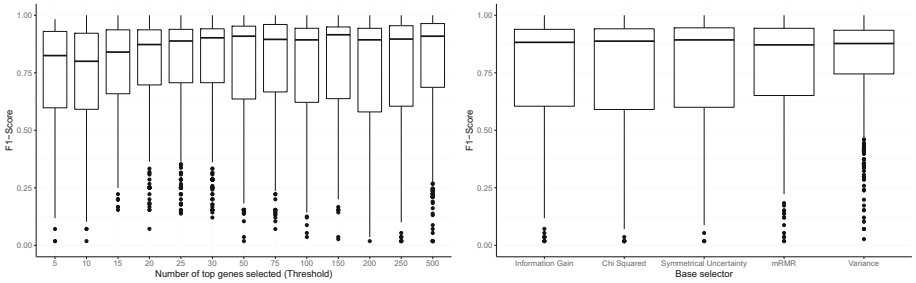


Fig. 5. Local stability for training datasets considering each base selector.

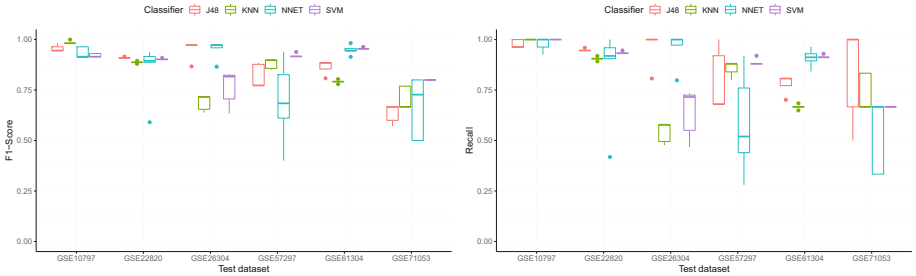
4.2 Predictive Performance Analysis

Classifiers predictive performance can also guide the identification of biomarkers. We observed that the threshold and the base selectors do not impact performance significantly (Fig. 6). We can see a slight increase on performance median until 50 genes are selected. However, in general, the top 5 genes already presented high F1-Scores (Fig. 6a). Among the base selectors, variance presented the most consistent results across multiple evaluations, which corroborates the stability results. Therefore, we will focus our next analyses on results presented by variance as base selector and the top 20 genes as threshold. Although other threshold values could be explored, the top 20 genes seem to show a good compromise between subset size, performance, and stability.



(a) (b)

Fig. 6. Analysis of the impact of a) threshold and b) base selector over performance.



(a) F1-Score. (b) Recall.

Fig. 7. Predictive performance on test datasets using variance and the top 20 genes.

In general, the EFS method presented satisfactory performance (Fig. 7). In terms of F1-Score (Fig. 7a), we note an insignificant difference between the highest median results within test datasets. We observed more stable results for precision (data not shown) than recall (Fig. 7b), which could be caused by a disproportionately smaller number of negative samples (*i.e.,* normal samples) compared to the number of positive samples, thus making recall more sensitive to prediction variations. Nonetheless, we note that predictive performance mostly varies between different test datasets rather than between different classifiers for the same dataset, due to data heterogeneity.

Table 3 shows the performance summary for F1-Score. There is no unique classifier able to perform satisfactorily for every test dataset. However, the mean absolute error (MAE), which is the mean difference between the highest performance and the other performances, in most cases is less than 0.1. MAE can be seen as a relative measure of how much better the best classifier is among the others. Note that when SVM outperforms other classifiers within test datasets, its MAE is higher if compared to the other cases. In this way, generally, SVM would be the best choice. On the other hand, in terms of recall, KNN seems to be a more robust choice (Table 4).

Table 3. Classifiers mean F1-Score by test dataset for *variance* and the top 20 genes. Bold values indicate the highest performance achieved for a dataset.

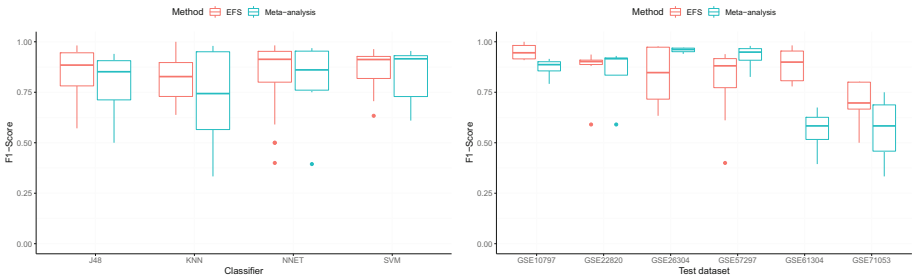
Dataset	Classifier				*MAE*
	SVM	J48	KNN	NNET	
GSE10797	0.92 ± 0.01	0.92 ± 0.03	$\mathbf{0.96 \pm 0.02}$	0.92 ± 0.02	0.04
GSE22820	0.72 ± 0.28	0.87 ± 0.05	$\mathbf{0.88 \pm 0.05}$	0.85 ± 0.16	0.06
GSE26304	0.86 ± 0.12	$\mathbf{0.88 \pm 0.14}$	0.77 ± 0.07	0.84 ± 0.22	0.05
GSE57297	$\mathbf{0.93 \pm 0.04}$	0.79 ± 0.10	0.89 ± 0.03	0.77 ± 0.18	0.11
GSE61304	$\mathbf{0.96 \pm 0.03}$	0.84 ± 0.09	0.84 ± 0.09	0.94 ± 0.04	0.08
GSE71053	$\mathbf{0.76 \pm 0.06}$	0.53 ± 0.14	0.68 ± 0.07	0.62 ± 0.17	0.15

Table 4. Average recall by classifier computed for the test datasets using the *variance* and the top 20 genes. Bold values indicate the highest performance for a dataset.

Dataset	Classifier				*MAE*
	SVM	J48	KNN	NNET	
GSE10797	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.02	0.03
GSE22820	0.71 ± 0.37	0.87 ± 0.10	$\mathbf{0.89 \pm 0.09}$	0.86 ± 0.23	0.08
GSE26304	0.54 ± 0.43	$\mathbf{0.85 \pm 0.22}$	0.65 ± 0.10	0.81 ± 0.28	0.18
GSE57297	$\mathbf{0.89 \pm 0.09}$	0.78 ± 0.20	$\mathbf{0.89 \pm 0.08}$	0.72 ± 0.25	0.14
GSE61304	$\mathbf{0.92 \pm 0.06}$	0.76 ± 0.16	0.74 ± 0.14	0.90 ± 0.06	0.12
GSE71053	0.62 ± 0.07	0.57 ± 0.29	$\mathbf{0.75 \pm 0.16}$	0.50 ± 0.16	0.19

4.3 EFS Versus Meta-analysis

To validate our EFS framework, we compared it with results generated by the REM meta-analysis approach, considering the top 20 genes in both cases and the variance FS method for EFS. Figure 8 shows the performance a) by classifier across all test datasets and b) by test dataset across all classifiers. Classifiers were able to perform satisfactorily across datasets, with median values surpassing the 0.75 mark in most cases. We note that performance for EFS is more consistent than meta-analysis and, in general, our approach has the interquartile range shifted towards higher values, which indicates better performance. When observing F1-Score per test dataset, our EFS approach outperforms meta-analysis in four out of the six datasets. In particular, for GSE10797, GSE61304, and GSE71053, the improvement brought by EFS over meta-analysis is remarkable. Table 5 summarizes the performance of both methods across all experiments, which corroborates the previous discussion. Note that for GSE613040 and GSE71053, EFS surpasses meta-analysis by 0.331 and 0.139, respectively.



(a) Results by classifiers. (b) Results by test dataset

Fig. 8. Comparison between the predictive performance of EFS and meta-analysis.

Table 5. Average F1-scores for EFS or meta-analysis based on test datasets, using *variance* and the top 20 genes. Bold values indicate the highest performance for a dataset.

Dataset	Method		*MAE*
	EFS	Meta-analysis	
GSE10797	**0.94 ± 0.02**	0.87 ± 0.05	0.079
GSE22820	**0.88 ± 0.07**	0.83 ± 0.16	0.048
GSE26304	0.83 ± 0.13	**0.95 ± 0.01**	0.120
GSE57297	0.82 ± 0.13	**0.92 ± 0.06**	0.098
GSE61304	**0.88 ± 0.07**	0.55 ± 0.11	0.331
GSE71053	**0.70 ± 0.10**	0.56 ± 0.18	0.139

When comparing the top 20 ranking generated by EFS and meta-analysis, an interesting finding is that the overlap among them is very low. Only one gene was common to both rankings, namely *S100A7*, which according to literature is highly expressed in breast cancer and may play a role in early tumor progression [12]. The high correlation among genes may justify to some extent the low overlap - a factor that should be further investigated. Among the top 5 genes found by our approach, we identified *TFF1*, *SCGB1D2*, *SCGB2A2*, and *PIP* with previous relation with breast cancer according to Genecards database. A more in-depth analysis of the genes selected by the EFS approach, including the investigation of their biological role, may be useful for better understanding their possible relation with breast cancer.

5 Conclusions

We explored a homogeneous EFS approach for candidate biomarkers identification from microarray data related to breast cancer in this work. A wide range

of parameters and configurations were explored. We could note the impact of the FS method chosen as base selectors over performance and the effect of data heterogeneity across distinct, independent studies. By combining several feature rankings extracted from data bootstraps and multiple microarray studies, the classifiers trained with the top EFS-based ranked features distinguished tumor from non-tumor samples with very good and robust performance. A good stability was found for a study-centered analysis, although stability across multiple studies was lower, partially due to high heterogeneity among samples. Interestingly, our approach was shown to outperform in most experiments one of the state-of-the-art approaches for integrative microarray analysis, the REM meta-analysis method. Further analyses of our top-ranked features could introduce new insights into the molecular mechanisms underlying breast cancer.

Provided that feature selection from high dimensional data is a challenging task, mainly when applied in the domain of omics data, we believe that our results are auspicious and encourage further studies in this direction. In particular, for future works, we aim to expand our analysis for other types of cancer, investigate the role of gene expression correlation in the stability of the rankings, perform a further investigation of top-ranked genes, and explore new strategies to improve stability under disease heterogeneity among samples.

Acknowledgments. The authors thank Rodrigo Haas Bueno for his help with microarray data preprocessing.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3), 392–398 (2009)
2. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. **6**(1), 37–66 (1991)
3. Alejandro, L.R., Marlet, M.A., Gustavo Ulises, M.R., Alberto, T.: Ensemble feature selection and meta-analysis of cancer miRNA biomarkers. bioRxiv, p. 353201 (2018)
4. Ang, J.C., Mirzal, A., Haron, H., Hamed, H.N.A.: Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans. Comput. Biol. Bioinform. **13**(5), 971–989 (2016)
5. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A.: An ensemble of filters and classifiers for microarray data classification. Pattern Recogn. **45**(1), 531–539 (2012)
6. Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J., Herrera, F.: A review of microarray datasets and applied feature selection methods. Inf. Sci. **282**, 111–135 (2014)
7. Bolón-Canedo, V., Alonso-Betanzos, A.: Ensembles for feature selection: a review and future trends. Inf. Fusion **52**, 1–12 (2019)
8. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth and Brooks, Belmont (1984)
9. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)

10. Cramer, H.: Mathematical Methods of Statistics (PMS-9), vol. 9. Princeton University Press, Princeton (1999)
11. Durinck, S., et al.: Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21**, 3439–3440 (2005)
12. Emberley, E.D., Murphy, L.C., Watson, P.H.: S100A7 and the progression of breast cancer. Breast Cancer Res. **6**(4), 1–7 (2004)
13. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
14. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall PTR, Upper Saddle River (1998)
15. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. **34**(4), 215–225 (2010)
16. Krieken, J.H.: Precision medicine. J. Hematopathol. **6**(1), 1–1 (2013). https://doi.org/10.1007/s12308-013-0176-x
17. Karley, D., Gupta, D., Tiwari, A.: Biomarker for cancer: a great promise for future. World J. Oncol. **2**(4), 151 (2011)
18. Kent, J.T.: Information gain and a general measure of correlation. Biometrika **70**(1), 163–173 (1983)
19. Khaire, U.M., Dhanalakshmi, R.: Stability of feature selection algorithm: a review. J. King Saud Univ. Comput. Inf. Sci. (2019)
20. Kuncheva, L.: A stability index for feature selection. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA, vol. 2007, pp. 421–427 (2007)
21. Liu, H., Liu, L., Zhang, H.: Ensemble gene selection by grouping for microarray data classification. J. Biomed. Inform. **43**(1), 81–87 (2010)
22. Pes, B.: Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. Neural Comput. Appl. **32**(10), 5951–5973 (2019). https://doi.org/10.1007/s00521-019-04082-3
23. Pes, B., Dessì, N., Angioni, M.: Exploiting the ensemble paradigm for stable feature selection: a case study on high-dimensional genomic data. Inf. Fusion **35**, 132–147 (2017)
24. Ritchie, M.E., et al.: LIMMA powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**(7), e47 (2015)
25. Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V., Alonso-Betanzos, A.: Ensemble feature selection: homogeneous and heterogeneous approaches. Knowl.-Based Syst. **118**, 124–139 (2017)
26. Sharifi, S., Pakdel, A., Ebrahimi, M., Reecy, J.M., Fazeli Farsani, S., Ebrahimie, E.: Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. PLoS ONE **13**(2), 1–18 (2018)
27. Surowiecki, J.: The Wisdom of Crowds. Knopf Doubleday Publishing Group, New York (2005)
28. Theil, H.: A note on certainty equivalence in dynamic planning. Econometrica **25**(2), 346–349 (1957)
29. Toro-Domínguez, D., Villatoro-García, J.A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M.E., Carmona-Sáez, P.: A survey of gene expression meta-analysis: methods and applications. Brief. Bioinform. **22**(2), 1694–1705 (2021)
30. Walsh, C.J., Hu, P., Batt, J., Santos, C.C.D.: Microarray meta-analysis and cross-platform normalization: integrative genomics for robust biomarker discovery. Microarrays **4**(3), 389–406 (2015)