Machine learning models to search relevant genetic signatures in clinical context

D. Urda
Universidad de Málaga,
Andalucía Tech
ETSI Informática (España)
Email: durda@lcc.uma.es

R.M. Luque-Baena, L. Franco,
and J.M. Jerez
Universidad de Málaga,
Departamento de Lenguajes y Ciencias de la Computación
ETSI Informática (España)
and Instituto de Investigación Biomédica de Málaga (IBIMA),
Inteligencia Computacional en Biomedicina
Email: {rmluque,lfranco,jja}@lcc.uma.es

N. Sanchez-Maroño
Universidad de La Coruña
Departamento de Computación
Email: nsanchez@udc.es

***Abstract*—Clinicians are interested in the estimation of robust and relevant genetic signatures from gene sequencing data. Many machine learning approaches have been proposed trying to address well-known issues of this complex task (feature or gene selection, classification or model selection, and prediction assessment). Addressing this problem often requires a deep knowledge of these methods and some of them demand high computational resources that may not be affordable. In this paper, an exhaustive study that includes different types of feature selection methods and classifiers is presented, providing clinicians an useful insight of the most suitable methods for this purpose. Predictions assessment is performed using a bootstrap cross-validation strategy as an honest validation scheme. The results of this study for six benchmark datasets show that filter or embedded methods are preferred, in general, to wrapper methods according to their better statistical significant results, in terms of accuracy, and lower demand for computational resources.**

I. Introduction

Machine learning (ML) and predictive modeling approaches are progressively being applied to data mining in personalized medicine, research field that relies on selecting optimal therapies based on the context of patient's clinical and genetic signature. Many studies are nowadays making use of ML procedures in prediction and prognosis of complex traits [1], [2]. In particular, there is a huge investment of resources in cancer research since the identification of genetic signatures correlated with clinical outcome remains as a challenging task in clinical assistance [3], [4], [5], [6], [7]

Nevertheless, the use of gene expression profiles in the estimation of prognosis models to find genetic signatures is a complex task in ML. It usually involves different steps that are not always easy to perform for clinicians, or even researchers, due mainly to the great variety of procedures and methods available in the literature, and the high-computing resources required by most of them. Feature (or gene) selection, classification and model selection, and prediction assessment are the three classical steps involved in the search of genetic signatures using ML approaches [8]. This family of procedures takes as input gene expression profiles from both Next Generation Sequencing (NGS) and DNA microarrays experiments [9], [10], [11].

Given the importance of these three steps involved in the estimation of genetic signatures, the impossibility of clinicians to test every ML method, and the high computational resources requirements, many authors have proposed different approaches trying to find molecular signatures with good prediction accuracy. In the feature selection step (genes to be included in the prognosis model), methods such as Partial Least Squares (PLS) regression [12], Information Gain (IG) [13], Minimum-Redundancy Maximum-Relevance (mRMR), and ReliefF [14] are among the statistical techniques proposed to address the problem. On the other hand, wrapper methods such as Stepwise Forward Selection (SFS) [15], [16], Ant Colony optimization [17], and evolutionary models [18], [19], [20], [21], [22] have been applied as heuristic methods from the computational intelligence perspective. With regard to classification model selection, different algorithms have been studied for the identification of differentially expressed genes in genomic data. Classification methods such as Multilayer Perceptron (NN) [23], [24], [15], Support Vector Machines (SVM) [25], Naive Bayes (NB) [26], k-Nearest Neighbour (kNN) [27], Decision Trees (DT) [28], and RF (Random Forest) [29] have been used in recent studies. Finally, prediction assessment refers to the performance of the predictive models. As few patient samples are typically available in genomic data, resampling techniques are a suitable methodology. In this sense, any dimensionality reduction technique should be performed within each resampling step in order to estimate prediction errors in a completely independent test set. This process is known as honest performance assessment [30], a necessary process if we are looking for generalizable results in independent cohorts, issue that has been overlooked in several works [15], [31], [32], [33]. On the other hand, honest validation strategies are presented in [8] and [34]. Specifically, in [8] the .632+ bootstrap method is highlighted for high-dimensional genomic studies and a number of existing bootstrap methods are compared (out-of-bag estimation and a bootstrap cross-validation (BCV) method [35]).

Despite all the extensive work that has been done in this research area, there is no conclusive results on which ML

TABLE I: Information about the six databases analysed.

Dataset	*#Genes*	*#Samples*	***"normal"***	***"cancer"***
West_ER	7129	49	25	24
Breast	24481	78	34	44
Leukaemia	7129	72	25	47
Lung	12533	181	150	31
Colon	2000	62	22	40
Prostate	12600	102	50	52

method performs better in order to estimate genetic signatures with relevance in the clinical practice. Up to date, these three steps have not yet been analyzed in a single work and tested over different datasets. Additionally, an extension of the classical evolutionary approach to carry out feature selection is presented. This procedure is considered a type of wrapper method. Therefore, this exhaustive analysis aims to help clinicians by providing a tool (or combination of FS procedures and classification algorithms) that may offer relevant results in terms of robustness, size and biological relevance of the genetic signature.

The rest of the paper is structured as follows. Section II shows the databases used within this study. Section III-A describes the FS techniques tested in this work to obtain a subset of genes and estimate prediction errors. Section III-B presents the methodology of our approach. Section III-C describes several machine learning models used to predict the clinical outcomes and Section IV shows the experimental results over different databases. Finally, Section V provides the final conclusions of this paper.

II. Datasets

Six free public high-dimensional microarray datasets[1][2] have been used within this work. Although NGS clearly seems to be the predominant technology in the near future of biomedical research [36], expression arrays were selected because are still cheaper and easier when used in clinical research [37], [38], [39], [40], [41]. In fact, from a practical point of view, ML procedures need to be fed with expression profiles in matricial format independently of the sequencing technology (NGS or microarray). The information of each dataset is shown in Table I, and each is related to the study of a specific cancer: breast, leukaemia, lung, colon, and prostate cancer diseases.

The West_ER dataset analyses primary breast tumours in relation to estrogen receptor (ER) status; the Breast dataset was reported for patients' outcome prediction related to breast cancer disorder; the Leukaemia dataset contains measures that correspond to Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML) samples; the Lung dataset has samples for two subtypes of lung cancer such as malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA); the Colon dataset with healthy samples versus patients diagnosed with colorectal cancer, and, finally, the Prostate dataset contains different neoplastic samples as these tumours are among the most heterogeneous of cancers both histologically and with respect to highly divergent clinical outcomes.

TABLE II: Summary of the different feature selection methods considered in this work together with some references where they have been previously used. We propose one GA extension denoted with *** as another wrapper method.

Filter	**Wrapper**	**Embedded**
CFS [43]	SFS [44]	SVM-RFE [45]
Cons [46]	**GA** ***	-
mRMR [47]	-	-
IG [48]	-	-
ReliefF [49], [50]	-	-

III. Methods

A. Feature selection framework

Feature selection procedures aims to select the most significant subset of genes, in terms of prediction accuracy, that are correlated with a clinical outcome. Filter, wrapper, and embedded methods are the three main categories into which feature selection techniques can be divided [42]. In filter methods, some statistical procedures are applied to remove irrelevant features, as it is a method that is completely independent of the classifier. Wrapper methods evaluate different subsets of features within a classification algorithm comparing their accuracy, thus requiring more computational resources in contrast to filter methods. Finally, embedded methods could be seen as a mix of filter and wrapper methods where the search space is composed of the feature-selection procedure and the classification algorithm as a whole, thus also being a classifier-dependent method.

Table II shows the different feature selection methods implemented in this paper to compare the predictive performance of the different estimated genetic signatures. According to the three families of methods, we considered: *(i) filter methods:* Correlation-based (CFS), Consistency-based (Cons), Information Gain (IG), Minimum-Redundancy Maximum-Relevance (mRMR) and ReliefF; *(ii) Wrapper:* Stepwise Forward Selection (SFS) and Genetic Algorithms (GA); *(iii) Embedded:* Support Vector Machines with Recursive Feature Elimination (SVM-RFE).

In this work, we propose an extension of classical evolutionary strategies as another wrapper method to perform feature selection by considering together the predictive performance of the genetic signature and the correlation among the features selected and the target class. GAs are a class of optimization procedure, inspired by the biological mechanisms of reproduction, in which a fitness function $f(\mathbf{x})$ should be maximized or minimized over a given space X of arbitrary dimension. A simple encoding scheme representing as much of the available information as possible was employed in which the chromosome is a string of bits whose length is determined

[1] http://datam.i2r.a-star.edu.sg/datasets/krbd/

[2] http://cilab.ujn.edu.cn/datasets.htm

by the total number of genes. Each gene is associated with one bit in the string. If the i^{th} bit is active (value 1), then the i^{th} gene is selected in the chromosome. Otherwise, a value of 0 indicates that the corresponding gene is ignored. In this way, each chromosome represents a different genes subset. Both the active genes and their number are generated randomly. In all the experiments, a population size of 100 individuals was used.

A selection strategy based on a roulette wheel and uniform sampling was applied, while an elite count value of 10 (the number of chromosomes which are retained in the next generation) was selected. Scattered crossover, in which each bit of the offspring is chosen randomly, was the choice for combining parents of the previous generation. The crossover rate was set to 0.8. In addition to that, a traditional mutation operator which flips a specific bit with a probability rate of 0.2 was considered. A modification which involves mutating a random number of bits between 1 and the number of active genes of the individual is introduced. Since it was empirically verified that the best subsets include few features, this change avoids an increment in the number of active features in the last generations of the GA.

The fitness function assesses each chromosome in the population so that it can be ranked against all the other chromosomes. The main goal of gene subset selection is to use fewer genes to achieve the same or better performance. Additionally, it has been found that the combination of genes with low redundancy among them, that is, that provide different information about the target class, and with a certain resemblance to the target class can improve the performance rates [47]. Therefore, the fitness function should contain three terms: the misclassification error, the number of features selected, and a redundancy measure among them. Datasets are split into training and testing sets in order to evaluate the generalization ability of the proposed chromosome.

Statistical techniques such as mutual information [51] give us an idea of the correlation between a pair of features. The mutual information between two continuous random variables y and z is given by

$$I\left(y,z\right)=\int\int p\left(y,z\right)\log\left(\frac{p\left(y,z\right)}{p\left(y\right)p\left(z\right)}\right)dy\,dz \qquad (1)$$

where $p\left(y,z\right)$ is the joint probability density function of y and z, and $p\left(y\right)$ and $p\left(z\right)$ are the marginal probability density functions of y and z respectively. The mutual information is symmetric.

Moreover, it is non-negative, with a zero value indicating that the variables are independent. The more correlated two variables are, the greater their mutual information. Advantages of mutual information are that the dependency between variables is no longer restricted to being linear and it can handle nominal or discrete features. Although it is hard to compute for continuous data, the probability densities can be discretized using histograms, which are considered as good approximations [52]. A measure which incorporates the correlation of the features with the target class and penalizes the redundancy among the selected features is described as follows [47]:

$$corr(\mathbf{x})=\frac{1}{t}\sum_{i=1}^{k}\sum_{j=i+1}^{k}I(x_j,x_i)-\frac{1}{k}\sum_{j=1}^{k}I(x_j,C) \qquad (2)$$

where k is the number of features selected, C is the target class, and t is the number of combinations among the pairs of chromosome x analysed. Finally, the function to be minimized is represented as follows:

$$fitness(\mathbf{x})=(1-ACC(\mathbf{x}))+\lambda\frac{k}{\mathcal{N}}+\beta corr(\mathbf{x}) \qquad (3)$$

where $fitness(\mathbf{x})$ is the fitness value of the feature subset represented by $\mathbf{x}$; $ACC(\mathbf{x})$ is the accuracy rate obtained by the classifier using the test set; $\mathcal{N}$ is the total number of extracted features; finally, $corr(\mathbf{x})$ defines the correlation among the features and the target class, with the aim of avoiding the redundancy in the feature vector (Equation 2). The parameters λ and β can take values in the interval $(0,1)$ and were empirically chosen to be 0.4 and 0.25, respectively.

Therefore, if two subsets achieve the same performance while containing different numbers of features, the subset with fewer features is preferred. We also prefer the mixture of features that are less redundant among them, which is considered a good quality for classification tasks. Nevertheless, among the three terms - error, feature subset size, and correlation - the first one is our major concern.

B. Validation scheme

In this paper, an honest validation strategy is applied with the aim of obtaining a final subset of genes with high prediction capabilities. Stratified boostrap cross validation was chose as validation procedure since its good behaviour in estimating misclassification error with microarray datasets has been previously demonstrated in [35], [53]. A high-level description of our methodological approach is shown in Figure 1 as well as a brief pseudocode of the algorithm is described in Algorithm 1. In concrete, the developed procedure executes a 50-bootstrap resampling as external validation and 5-k-fold for internal validation techniques. Thus, this scheme will lead us to find subsets of features with high generalization rates in the prediction stage, as this is essential for determining the probability of suffering from a specific condition.

Moreover, in the case of wrapper methods, Welch's t-test [54] is applied assuming that the two classes (the patient does or does not have cancer) have unknown and unequal variances, because it is not advisable to use the basic form if we are unsure whether the requirements of the test are satisfied [42]. The top 200 of the total number of genes are retained according to the p-value descending sort, which will be the input of a wrapper method feature-selection procedure.

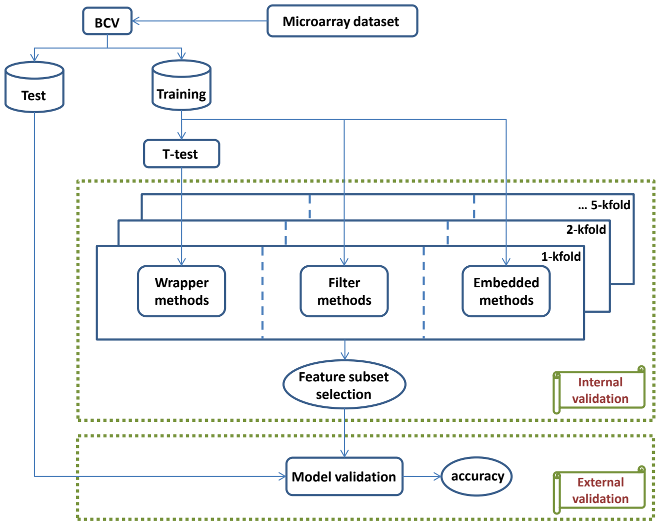


Fig. 1: Honest validation scheme used in the feature-selection procedure and prediction error estimation.

Algorithm 1 Pseudocode of our methodological approach

```
{initialization}
$[Train, Test]\{1..50\} \Leftarrow BCV(dataset, 50)$
$FeatureSelectionMethods$ =[SFS,GA,CFS, ... ,SVM-RFE]

for all method $m$ in $FeatureSelectionMethods$ do
    {first-step: feature selection process}
    for $i = 1 \rightarrow 50$ do
        $TR_i \Leftarrow Train[i]$
        if (IsWrapperMethod($m$)) then
            $TR\_Reduced_i \Leftarrow Ttest(TR_i)$
            $[IntVal_i, Features_i] \Leftarrow ExecWrapper(m, TR\_Reduced_i)$
            //involves execution of a classification method
        else if (IsFilterMethod($m$)) then
            $[Features_i] \Leftarrow ExecFilter(m, TR_i)$
            $[IntVal_i] \Leftarrow ExecClassificationMethod(TR_i, Features_i)$
        else
            $[Features_i] \Leftarrow ExecEmbedded(m, TR_i)$
            $[IntVal_i] \Leftarrow ExecClassificationMethod(TR_i, Features_i)$
        end if
    end for
    $InternalValidation \Leftarrow mean(IntVal_i)$

    {second-step: model validation}
    for $i = 1 \rightarrow 50$ do
        $T_i \Leftarrow Test[i]$
        $ExtVal_i \Leftarrow Accuracy(T_i, Features_i)$
    end for
    $ExternalValidation \Leftarrow mean(ExtVal_i)$
end for
```

C. *Classification models*

Several standard and well-known classification models have also been tested in this paper: Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), k-Nearest Neighbours (kNN), Naive-Bayes (NB), C-MANTEC (CM) [55] as a constructive neural network model, and a standard Multilayer Perceptron (NN). Nested cross-validation is used to perform model selection after doing a grid-search over the parameters of each machine learning model. In this sense, this procedure implies a huge effort in terms of computational and time resources, since in the case of wrapper methods many different subsets of features are evaluated in each iteration. Therefore, the authors propose to adjust the parameters of each method by using the top 200 variables after sorting the p-values obtained by the application of Welch's t-test. As a result of this suggested parameter estimation method, every configuration is labelled by an accuracy measure in a reasonable time. Finally, the parameter configuration with the highest average result in the outter folds is kept to fit the model to the training samples. The accuracy measure is obtained through the .632+ bootstrap method [56], as it is highlighted for high-dimensional genomic studies.

IV. Results

Figure 2 presents a summary of the final results obtained over each dataset. For each one, the bar diagram represents the performance of eight feature-selection procedures analysed in this paper: two different wrapper methods (SFS and GA), five different filter methods (CFS, Cons, IG, mRMR, and ReliefF), and one embedded method (SVM-RFE). This performance is computed after averaging the accuracy obtained with six machine-learning classifiers (LDA, SVM, kNN, NB, CM, and NN). In general, filter and embedded methods are distinguished by the most accurate results in contrast to wrapper methods, independently of the cancer microarray dataset analysed. In concrete, mRMR emerges as the one with the best performance in three out of six analysed cancer datasets (Leukaemia, Lung, and Colon). Therefore, the results suggest the use of filter or embedded methods instead of wrapper methods, since the latter are more highly computationally demanding, leading to lower performance results on average.

Regarding the six cancer microarray datasets analysed, for three of them (the Leukaemia, Lung, and Prostate datasets), a very good classification result is obtained independently of the feature-selection procedure (over a 90%). On the other hand, the West_ER and Colon datasets present good classification results (over 80%) while the Breast cancer dataset appears to be the most difficult problem as a success rate of only 65% is achieved, which could lead us to think that more patient samples are needed to estimate gene-expression-based predictors.

A. *Classification models' performance*

Regarding the classifier models' performance, a deeper analysis has been carried out and is represented in Figures 3 and 4. In particular, Figure 3 shows a pairwise analysis of the number of times that one classifier is statistically significant better than another one according to the multiple comparison test. The thickness of the lines connecting the different classifiers indicates four categories in which the comparison was done: classifiers that are better than the other less than four times are discarded and not represented on the graph; those that are four to six times better than the other are represented by the thinnest line on the graph; those that are seven to nine times better than the other are indicated with a middle thickness line; finally, the last category represents classifiers that are better than the other more than nine times (shown in the graph with the thickest line). The results suggest that the kNN, NN, or NB classifier

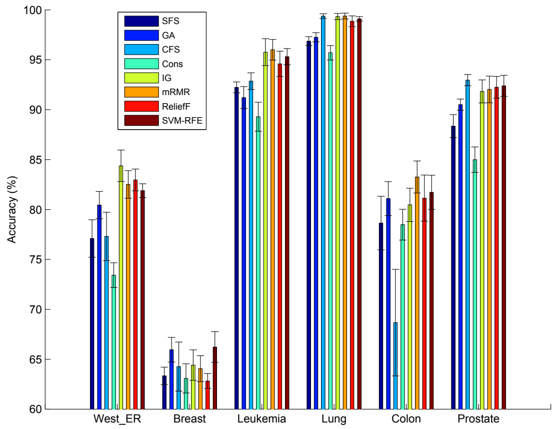


Fig. 2: Performance comparison (after averaging the accuracy of six machine learning classifiers) for eight feature selection procedures over six different cancer microarray datasets.

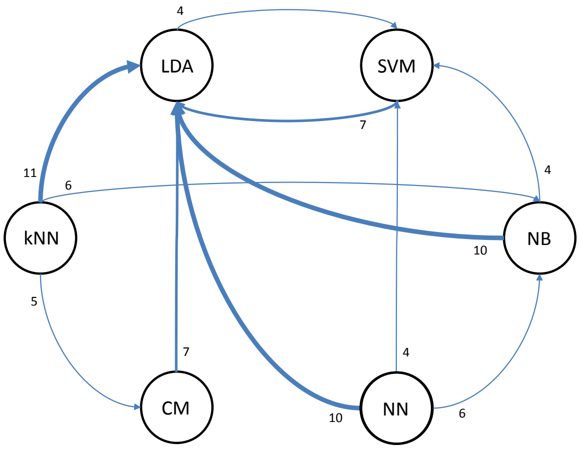
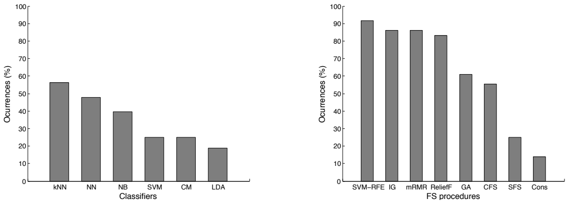


Fig. 3: Pairwise graph showing the number of times that the performance of a given classifier is statistically significant in comparison to other classifiers.

should be used rather than LDA and the rest. LDA is the only classifier that does not outperform any other classifier (does not reach the category that is seven to nine times better), so in principle it could be discarded as a classifier for the DNA microarray analysis.

Finally, Figure 4a) shows the percentage value of the number of times (occurrences) that a given classifier leads to statistically significant different results in comparison to a control group (the lowest in performance) computed among all analysed cases (different datasets and FS procedures), while Fig. 4b) shows a similar analysis but for different FS procedures among all datasets and classifiers. The histogram shown in 4a) indicates that kNN is the preferred classifier as it does outperform other classifiers almost 60% of the time, while LDA behaves quite poorly as it achieves the best results less than 20% of the time.



(a) (b)

Fig. 4: Summary of results. (a) Percentage value of the number of times (occurrences) that a given classifier leads to statistically significant different results in comparison to a control group computed for all analysed cases (different datasets and FS procedures). (b) Analysis similar to before as before but for different FS procedures performed over all datasets and classifiers.

B. Honest validation scheme

The use of an honest validation scheme is relevant, as performance results could be very optimistic otherwise. In this sense, and according to the results shown in Figure 2, we selected three a priori more difficult datasets (West_ER, Breast, and Colon) in order to perform a detailed analysis using SFS and GA as FS procedures. Table III shows the performance results of the LDA and SVM classifier models for each dataset with and without using an honest validation scheme. As expected, the behaviour of FS procedures is very optimistic if no honest validation is applied, independently of the classifier used. The final accuracy measures vary from honest validation schemes (H^+) to non-honest ones (H^-) approximately a 20% (West_ER), $16-20\%$ (Breast), and $5-18\%$ (Colon). Regarding the overfitting problem in the feature selection, an overfitting index (OI) was computed to analyse how much this effect affects the FS procedures, classifiers, and datasets. It was computed as $OI = 1 - (H^+/H^-)$, and averaged across datasets and classifiers, where OI considerably above zero will point out a clear overfitting scenario. The results of the overfitting index were $OI = 0.2297$ for the SFS procedure and $OI = 0.1845$ for the GA. Therefore, previous articles that did not use an honest validation scheme presented over-optimistic results as no test set was kept apart from the FS procedure.

C. Robustness of the FS procedures

An important aspect of the FS procedure is the variability observed in the set of selected subsets of genes in different executions of a given algorithm. In order to quantify this, we compute a robustness index for each FS procedure used, taking into consideration the subset of genes obtained for every resampling of the dataset. First, the absolute frequency for each gene is computed in order to retain those genes selected at least 5% of the time . Then the set of selected genes is sorted in descending order according to the relative frequency, discarding those genes for which the cumulative frequency is greater than 80%. Finally, the robustness measure is calculated as the average of the relative frequencies of the resulting genes.

TABLE III: Performance comparison among two different wrapper methods (SFS and GA) and two classifiers (LDA and SVM) using three datasets (West_ER, Breast and Colon). The results shown correspond to the accuracy of each classification method using the honest validation scheme proposed (H^+) in this work and without using it (H^-).

		SFS		GA	
	Classifier	H^+	H^-	H^+	H^-
West_ER	LDA	73.29	95.14	81.16	99.52
	SVM	77.78	96.63	79.58	99.32
Breast	LDA	63.35	79.45	66.23	95.10
	SVM	64.27	82.36	66.72	97.78
Colon	LDA	81.69	86.04	83.70	92.45
	SVM	78.39	88.27	78.61	95.28

Figure 5 shows the robustness value obtained for each dataset, depending on the FS procedure used. ReliefF could be considered the most robust FS procedure according to our analysis, since it leads to the highest robustness values for three datasets with competitive values for the other three. Moreover, IG and mRMR are a step backward in comparison to ReliefF but they also have competitive robustness values. On the other hand, the remainder of the FS procedures have values of less than 0.5 for almost all of the datasets, and thus it can be derived that on several executions of the algorithm, a different subset of genes will be obtained. It should be noted that there is no clear correlation between the robustness and the accuracy measure, since the most robust method (ReliefF) is not the same as the most accurate technique (SVM-RFE). Between the wrapper methods, GA overcomes SFS in both robustness and accuracy. To further confirm the results shown in the figure, permitting a more direct comparison of the robustness of the FS procedures, we compute a weighted average of the results shown by averaging the observed values re-scaled in relationship to the maximum value obtained within each dataset (an average value of 1 would indicate that the FS method obtained the best robustness index for all datasets), obtaining the following values: ReliefF: 0.87179; IG: 0.78078; mRMR: 0.64941; GA: 0.61396; SFS: 0.48686; CFS: 0.45904; SVM-RFE: 0.43212; Cons: 0.30266.

D. Number of selected genes

The number of genes obtained by the different FS procedures studied varies depending on several factors. According to the raw data results shown in Table IV, it can be appreciated that the SFS, GA, and Cons procedures are more aggressive in the gene-selection procedure, as few genes are kept in the final solutions. In the case of the SFS procedure, this could be explained by the nature of the algorithm, as it begins from solutions with only one gene and then iteratively adds new genes, while the performance is statistically significantly better than in the previous iteration. In a similar way, the GA procedure includes in its fitness function the desired characteristics of the solutions, maximizing the accuracy result while at the same time keeping those configurations with a smaller number of genes.

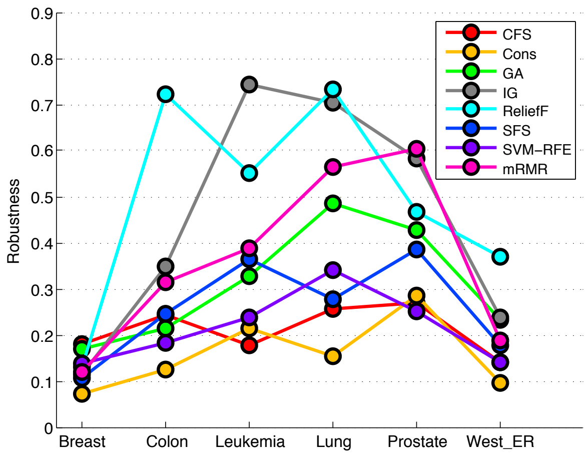


Fig. 5: Robustness measure for each FS procedure among the different resamplings of each dataset, computed taking into account the variability of the selected genes.

In comparison to wrapper methods (SFS and GA), filter and embedded methods are independent from the classification model. As these methods usually retain many more genes in their solutions, we established the following cut-off criteria: if the solution has more genes than the number of samples available in the dataset, then only the first $\#samples/8$ genes are kept (genes are sorted according to their suitability). Thus, there are some cases of FS procedures (i.e., SVM-RFE, mRMR, ReliefF, ...) where the number of selected genes is constant for all resamplings (the standard deviation is equal to zero in these cases). This criterion was set, firstly in order to reduce the number of selected genes per resampling and secondly in order to apply similar criteria over all filter and embedded methods so that a fair comparison could be made.

V. Conclusion

Based on the results discussed above, it is time to ask whether it is worth testing every ML model available to find relevant genetic signatures from gene expression data or not. In Figure 2 we presented a summary of the average accuracy results for all of the analysed datasets. The results of this study indicate the presence of three less complex datasets (Leukaemia, Lung, and Prostate) for which, independently of the FS and classification method used, the accuracy is always larger than 84%. Moreover, the use of honest validation schemes leads to less overfitting in the feature selection (an overfitting index was computed by dividing the accuracy with the proposed honest validation scheme and without using an honest validation scheme and then averaged across the three selected datasets and the two classification algorithms analysed).

TABLE IV: Number of selected genes obtained (*mean±standard deviation*) for the eight analysed feature selection procedures (SFS, GA, CFS, Cons, IG, mRMR, ReliefF and SVM-RFE) for three cancer microarray datasets (West_ER, Breast and Leukaemia).

	West_ER	Breast	Leukemia
SFS	2.51±0.74	3.10±1.20	2.25±0.63
GA	4.21±1.08	10.04±2.30	3.63±0.98
CFS	8.88±10.96	10.28±9.05	30.02±23.33
Cons	2.12±0.44	3.36±0.56	1.84±0.51
IG	6.00±0.00	9.00±0.00	9.00±0.00
mRMR	6.00±0.00	9.00±0.00	9.00±0.00
ReliefF	6.00±0.00	9.00±0.00	9.00±0.00
SVM-RFE	6.00±0.00	9.00±0.00	9.00±0.00

Regarding the classification models, the results shown in Figures 3 and 4 suggest that kNN and NN classifiers could be considered more robust methods independently of the FS method used and the dataset. Nevertheless, other classification techniques such as SVM or CM, which require the adjustment of several parameters, could lead to the achievement of similar results after a fine-tuning in the parameter estimation stage. Regading the FS methods, the embedded SVM-RFE and three other filter methods (IG, mRMR, and ReliefF) behave qualitatively better than the rest of the methods, indicating a superior performance in comparison to wrapper methods. Further, taking into account that wrapper methods tend to be more computationally intensive, the previous results clearly suggest an advantage of filtering (or embedded) FS schemes. In relation to the number of selected genes, SFS and Cons lead to more restricted sets, but with the disadvantage of worse performance, indicating that except when the size of the final set is a very important factor, these two FS procedures should not be the preferred option.

Finally, the overall conclusion of the present study will depend on finding the right balance between the cost associated to carry out a wide analysis like this one and the performance that we may be interested in. In general, filter and embedded methods may be more suitable rather than wrapper ones, as they lead to more robust results in terms of both percentages of better statistical significant results and overfitting effects, and are also less computationally intensive. Some of these FS methods in combination with kNN or NN classifiers could lead to robust and relevant genetic signatures in the studied disorder, thus being a suggested set of methods to be tried first by clinicians.

Acknowledgments

The authors acknowledge support through grants TIN2014-58516-C2-1-R from MICINN-SPAIN which include FEDER funds, and from ICE Andalucía TECH (Spain) through a postdoctoral fellowship.

References

[1] M. W. Libbrecht and W. S. Noble, “Machine learning applications in genetics and genomics,” *Nature Review Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[2] S. Okser, T. Pahikkala, A. Airola, T. Salakoski, S. Ripatti, and T. Aittokallio, “Regularized machine learning in the genetic prediction of complex traits,” *PLoS Genetics*, vol. 10, no. 11, p. e1004754, 2014.

[3] S. Winslow, K. Leandersson, A. Edsjo, and C. Larsson, “Prognostic stromal gene signatures in breast cancer,” *Breast Cancer Research*, vol. 17, no. 1, p. 23, 2015.

[4] D. Bedognetti, W. Hendrickx, F. M. Marincola, and L. D. Miller, “Prognostic and predictive immune gene signatures in breast cancer,” *Current opinion in Oncology*, vol. 27, no. 6, pp. 433–444, 2015.

[5] X. Zhao, E. A. Rødland, T. Sørlie, H. K. M. Vollan, H. G. Russnes, V. N. Kristensen, O. C. Lingjærde, and A.-L. Børresen-Dale, “Systematic assessment of prognostic gene signatures for breast cancer shows distinct influence of time and er status,” *BMC Cancer*, vol. 14, no. 1, pp. 1–12, 2014.

[6] S. Irshad, M. Bansal, M. Castillo-Martin, T. Zheng, A. Aytes, S. Wenske, C. Le Magnen, P. Guarnieri, P. Sumazin, M. C. Benson, M. M. Shen, A. Califano, and C. Abate-Shen, “A molecular signature predictive of indolent prostate cancer,” *Science Translational Medicine*, vol. 5, no. 202, pp. 202ra122–202ra122, 2013.

[7] L. Chin, W. C. Hahn, G. Getz, and M. Meyerson, “Making sense of cancer genomic data,” *Genes & Development*, vol. 25, no. 6, pp. 534–555, 2011.

[8] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, “Prediction error estimation: a comparison of resampling methods,” *Bioinformatics*, vol. 21, pp. 3301–3307(7), 2005.

[9] X. Guo, S. Zhu, A. Brunner, M. van de Rijn, and R. West, “Next generation sequencing-based expression profiling identifies signatures from benign stromal proliferations that define stromal components of breast cancer,” *Breast Cancer Research*, vol. 15, no. 6, p. R117, 2013.

[10] J. Phan, A. Young, and M. Wang, “Robust microarray meta-analysis identifies differentially expressed genes for clinical prediction,” *The Scientific World Journal*, vol. 2012, 2012.

[11] X. Wang and R. Simon, “Microarray-based cancer prediction using single genes,” *BMC Bioinformatics*, vol. 12, 2011.

[12] S. Student and K. Fujarewicz, “Stable feature selection and classification algorithms for multiclass microarray data,” *Biol Direct*, vol. 7, no. 1, p. 33, 2012.

[13] D. Dittman, T. Khoshgoftaar, R. Wald, and A. Napolitano, “Similarity analysis of feature ranking techniques on imbalanced DNA microarray datasets,” 2012, pp. 398–402.

[14] Y. Zhang, C. Ding, and T. Li, “Gene selection algorithm by combining relieff and mrmr,” *BMC Genomics*, vol. 9, no. SUPPL. 2, 2008.

[15] L. J. Lancashire, R. C. Rees, and G. R. Ball, “Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach,” *Artif. Intell. Med.*, vol. 43, no. 2, pp. 99–111, 2008.

[16] H. Peng, Y. Fu, J. Liu, X. Fang, and C. Jiang, “Optimal gene subset selection using the modified SFFS algorithm for tumor classification,” *Neural Comput Appl*, pp. 1–8, 2012.

[17] H. Yu, J. Ni, and J. Zhao, “ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data,” *Neurocomputing*, vol. 101, pp. 309–318, 2013.

[18] J. A. Castellanos-Garzón and F. Díaz, “An evolutionary computational model applied to cluster analysis of DNA microarray data,” *ESWA*, vol. 40, no. 7, pp. 2575 – 2591, 2013.

[19] A. Sungheetha and J. Suganthi, “An efficient clustering-classification method in an information gain NRGA-KNN algorithm for feature election of micro array data,” *Life Science Journal*, vol. 10, no. SUPPL. 7, pp. 691–700, 2013.

[20] E. Keedwell and A. Narayanan, “Gene expression rule discovery and multi-objective ROC analysis using a neural-genetic hybrid,” *Int J Data Min Bioin*, vol. 7, no. 4, pp. 376–396, 2013.

[21] S. Gupta and S. Garg, “Multiobjective optimization using genetic algorithm,” *Advances in Chemical Engineering*, vol. 43, pp. 206–245, 2013.

[22] A. Kulkarni, B. Naveen Kumar, V. Ravi, and U. Murthy, "Colon cancer prediction with genetics profiles using evolutionary techniques," *ESWA*, vol. 38, no. 3, pp. 2752–2757, 2011.

[23] E. Hernández-Pereira, V. Bolón-Canedo, N. Sánchez-Maroño, D. Álvarez-Estévez, V. Moret-Bonillo, and A. Alonso-Betanzos, "A comparison of performance of k-complex classification methods using feature selection," *Information Sciences*, vol. 328, pp. 1–14, 2016.

[24] M. Pirooznia, J. Yang, M. Qu, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," *BMC Genomics*, vol. 9, no. SUPPL. 1, 2008.

[25] M. Zervakis, M. Blazadonakis, G. Tsiliki, V. Danilatou, M. Tsiknakis, and D. Kafetzopoulos, "Outcome prediction based on microarray analysis: A critical perspective on methods," *BMC Bioinformatics*, vol. 10, 2009.

[26] M. Wu, D. Dai, Y. Shi, H. Yan, and X. Zhang, "Biomarker identification and cancer classification based on microarray data using laplace naive bayes model with mean shrinkage." *IEEE/ACM T Comput Bi*, vol. 9, no. 6, pp. 1649–1662, 2012.

[27] S. Li, E. Harner, and D. Adjeroh, "Random knn feature selection - a fast and stable alternative to random forests," *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[28] Q. Han and G. Dong, "Using attribute behavior diversity to build accurate decision tree committees for microarray data," *JBCB*, vol. 10, no. 4, 2012.

[29] M. Burton, M. Thomassen, Q. Tan, and T. Kruse, "Gene expression profiles for predicting metastasis in breast cancer: A cross-study comparison of classification methods," *The Scientific World Journal*, vol. 2012, 2012.

[30] S. Dudoit and J. Fridlyand, *Statistical Analysis of Gene Expression Microarray Data*. Chapman and Hall, 2003.

[31] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002.

[32] A. R. Statnikov, C. F. Aliferis, I. Tsamardinos, D. P. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis." *Bioinformatics*, vol. 21, no. 5, pp. 631–643, 2005.

[33] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19(12), pp. 1484–1491, 2003.

[34] I. Walsh, G. Pollastri, and S. C. E. Tosatto, "Correct machine learning on protein sequences: a peer-reviewing perspective," *Briefings in Bioinformatics*, 2015.

[35] W. J. Fu, R. J. Carroll, and S. Wang, "Estimating misclassification error with small samples via bootstrap cross-validation," *Bioinformatics*, vol. 21, no. 9, pp. 1979–1986, May 2005.

[36] A. Srivastava, V. M. Philip, I. Greenstein, L. B. Rowe, M. Barter, C. Lutz, and L. G. Reinholdt, "Discovery of transgene insertion sites by high throughput sequencing of mate pair libraries," *BMC Genomics*, vol. 15, no. 1, pp. 1–9, 2014.

[37] S. Kar, K. D. Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing {PSO} and adaptive k-nearest neighborhood technique," *Expert Systems with Applications*, vol. 42, no. 1, pp. 612 – 627, 2015.

[38] S. Karimi and M. Farrokhnia, "Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 139, no. 0, pp. 6 – 14, 2014.

[39] E. Lotfi and A. Keshavarz, "Gene expression microarray classification using PCABEL," *Computers in Biology and Medicine*, vol. 54, no. 0, pp. 180 – 187, 2014.

[40] J. Nahar, T. Imam, K. S. Tickle, A. S. Ali, and Y.-P. P. Chen, "Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12 371 – 12 377, 2012.

[41] C.-K. Chen, "The classification of cancer stage microarray data," *Computer Methods and Programs in Biomedicine*, vol. 108, no. 3, pp. 1070 – 1077, 2012.

[42] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[43] M. A. Hall, "Correlation-based feature selection for machine learning," University of Waikato, Hamilton, New Zealand, Tech. Rep., 1998.

[44] A. R. Webb, *Statistical Pattern Recognition, 2nd Edition*, 3rd ed. John Wiley & Sons, 2011.

[45] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, Mar. 2002.

[46] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artif. Intell.*, vol. 151, no. 1-2, pp. 155–176, 2003.

[47] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[48] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[49] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," 1994, pp. 171–182.

[50] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*, ser. ML92, 1992, pp. 249–256.

[51] B. Guo and M. Nixon, "Gait feature subset selection by mutual information," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 39, no. 1, pp. 36 –46, 2009.

[52] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233 – 246, 1989.

[53] W. Pan, "Bootstrapping likelihood for model selection with small samples," 1998.

[54] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.

[55] J. L. Subirats, L. Franco, and J. M. Jerez, "C-mantec: A novel constructive neural network algorithm incorporating competition between neurons," *Neural Networks*, vol. 26, pp. 130 – 140, 2012.

[56] B. Efron and R. Tibshirani, "Improvements on cross-validation: The .632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. pp. 548–560, 1997.