Proceedings of the 53rd Hawaii International Conference on System Sciences | 2020

Applying Feature Selection to Improve Predictive Performance and Explainability in Lung Cancer Detection with Soft Computing

Nicolas Potie
Dept. of Computer Science and A.I.,
University of Granada
npotie@ugr.es

Stavros Giannoukakos
Dept. of Genetics,
University of Granada
sgiannoukakos@ugr.es

Michael Hackenberg
Dept. of Genetics,
University of Granada
hackenberg@go.ugr.es

Alberto Fernandez
Dept. of Computer Science and A.I.,
University of Granada
alberto@decsai.ugr.es

Abstract

The field of biomedicine is focused on the detection and subsequent treatment of various complex diseases. Among these, cancer stands out as one of the most studied, due to the high mortality it entails. The appearance of cancer depends directly on the correct functionality and balance of the genome. Therefore, it is mandatory to ensure which of the approximately 25,000 human genes are linked with this undesirable condition. In this work, we focus on a case study of a population affected by lung cancer. Patient information has been obtained using liquid biopsy technology, i.e. capturing cell information from the bloodstream and applying an RNA-seq procedure to get the frequency of representation for each gene. The ultimate goal of this study is to find a good trade-off between predictive capacity and interpretability for the discernment of this type of cancer. To this end, we will apply a large number of techniques for feature selection, using different thresholds for the number of selected discriminant genes. Our experimental results, using Soft Computing techniques, show that model-based feature selection via Random Forest is essential for both improving the predictive capacity of the models, and also their explainability over a small subset of genes.

1. Introduction

The importance of the Data Science area has been remarkably confirmed in recent years [1]. Indeed, stake-holders are realizing the importance of information, and especially the knowledge that can be extracted from data [2]. For this task, it is essential to use Machine Learning (ML) techniques that allow the generation of models that represent and explain the phenomena under study [3].

There exist plenty of areas of application in which the use of ML-based solutions is particularly beneficial. One clear example is biomedicine in general [4], and problems based on genomics in particular [5]. The human being is an inexhaustible source of data and taking advantage of them directly benefits society for a better quality of life.

Among different fields of study in biomedicine, cancer detection is beyond question one of the most important ones [6]. This is undoubtedly due to the high mortality of this disease, as well as the complexity of its correct diagnosis. Cancer is characterized by an abnormal increase of cell division in one or several parts of the body, leading to metastasis when it spreads uncontrollably. Regarding the different types of cancer that can affect the human body, lung cancer is possibly one of the most harmful [7, 8]. Specifically, in 2018, it comprised the 25% of the deaths caused by cancer in the USA, with more than 150,000 reported cases.

Currently, a number of diagnostic tools are under development. In particular, liquid biopsy [9] is one recent and promising new biotechnology that is emerging as an alternative to traditional imaging detection. The underlying mechanism is to capture cell information from the patient's bloodstream and then translate this information into gene expression data. This procedure is carried out via RNA-Seq [10], which is a precise measurement of the levels of transcripts involving thousands of genes under study.

In practical terms, clinicians cannot handle nor interpret such a large amount of genetic information. For the extracted models to be truly useful, it is essential to reduce the number of descriptive attributes to a volume that can explain the output of the diagnosis, and therefore present strong evidence when the system provides a given output label [11, 12]. Like lung cancer nodules can be seen from scanner images, cancer-driving mutations, genes expression or other features/sources must be explainable to be acceptable. It is more important that liquid biopsy pipeline is nearly fully computer-based after isolation of biological components of interests. Thanks to this, **personalized medicine** may become a reality [13].

Taking this into account, the model developed by the ML procedure must be as simple as possible, especially

URI: https://hdl.handle.net/10125/63952
978-0-9981331-3-3

concerning the number of involved genes. To this end, the use of feature selection techniques in this field is highly necessary [14]. However, there are different alternatives whose behavior needs to be analyzed in order to confirm the most promising methodology at the level of reduction, predictive capacity, and execution time.

Specifically, the importance of feature selection techniques allows for the identification of the so-called "differentially expressed genes", i.e. those that may include significant information for determining the corresponding health condition [15]. Among the most important techniques, we may stress solutions from statistical methods [16], fold change, rank products, or recursive feature elimination based on any given classification model [17, 18, 19].

However, most of the state-of-the-art work are focused on micro-array data [20], and therefore few research has been carried out in the scenario of liquid biopsy. With the objective of finding out the most appropriate feature selection approach in this context, we will carry out a thorough experimental study with several different methods. Specifically, we have selected univariate techniques with diverse scoring functions [21], recursive feature elimination [22], and a selection based on different learning models such as Random Forest [23]. Finally, as classification technique we will make use of two well-known Soft Computing approaches. On the one hand, the Gaussian Naïve Bayes classifier [24], as it has been shown to be a proper learning scheme in the event of sparse data, i.e. a greater number of variables than observations [25]. On the other hand, the Random Forest ensemble approach [23], a well-known and accurate solution in the field of Bioinformatics [5].

The rest of this contribution is arranged as follows. In Section 2 we will introduce the characteristics of the lung cancer diagnostics problem that will be used as case study. Then, Section 3 is devoted to describe the paradigms for feature selection that we have chosen for study. The experimental framework where we establish all the specifications for the analysis can be found in Section 4. Next, the experimental study where the results are analyzed is shown in Section 5. Finally, Section 6 will summarize and conclude this work.

2. A Case Study on Lung Cancer Prediction via Liquid Biopsy

Cancer is one of the most harming diseases nowadays. There are two main characteristics that represent this condition. On the one hand, the abnormal and uncontrolled cell proliferation and growth. On the other hand, the possibility of invading several parts of the body that were unrelated to the location of the primary tumor, which is known as metastasis. The main cause of cancer is the mutation in some of the genes involved in cell growth or cell cycle control. Therefore, it is of extreme importance to be able to detect changes in gene expression levels for determining the presence of malignant cells.

Throughout this section, we will introduce some of the main properties of lung cancer, as well as some key aspects of the novel revolutionary biotechnology, known as liquid biopsy, which allows detecting cancer from blood samples (Section 2.1). Afterwards, we will introduce the dataset employed as a case study, together with its main characteristics, and the bioinformatics methodology followed to prepare the dataset from the raw information to ease the applications of the ML-based models (Section 2.2).

2.1. Diagnosis of lung cancer using liquid biopsy

Lung cancer is a quite aggressive type of cancer that, as its name suggests, is mainly located in the respiratory system. Amongst all cancer types, it is by far the most harmful, being the third cause of death overall in the USA [7, 8].

One of the main drawbacks of lung cancer is that individuals do not exhibit any symptoms at the initial stages of the disease. This issue implies a major difficulty for an early detection via traditional imaging diagnosis, i.e. low-dose computed tomography (LDCT). In addition to the former, this methodology has several drawbacks, as pointed out below:

1. The first one is the high false positive rate that varied from 3-30% in Randomised Cohort Trials and from 5% to 51% in cohort studies[26]. This is due to the fact that LDCT can detect benign nodules that are not related to any cancer. Taking into account the invasive nature of the possible treatments, it implies a large cost for the health of these individuals.

2. The second drawback is the invasiveness due to the radioactive dose of LDCT per examination, namely 1.5 mSv[26].

3. The third disadvantage is the price. The annual screening costs between $126,000 - $169,000 per Quality-Adjusted Life-Year (QUALY) for lung cancer patient in the USA. This value is beyond the $100,000 threshold that is considered to be as cost effective.

Another issue related to lung cancer detection is that it distinguishes between two major molecular types. On the one hand, Small Cell Lung Cancer (SCLC) that comprises approximately 10-15% of the cases. On the other hand, and most common, Non-Small Cell Lung cancer (NSCLC) that represents the remaining 85-90%. Nevertheless, a closer look to NSCLC reveals three major subtypes, namely adenocarcinoma, squamous cell carcinoma and large cell cancer [27].

Focusing on the origins of cancer conditions, we must consider that genome alterations have the potential to serve as a powerful diagnostic tool. However, sequencing of the genome (or a part of it) is traditionally based on tissues biopsy. Thus, the cancer status of the individual as well as its location must be known. A new strong actor, liquid biopsy, could possibly solve all these problems thanks to today's advances of biotechnology [9]. The false positive rate can be studied threw statistical methods based on genomic information. Liquid biopsy is minimally invasive; nowadays a blood test is a well-known practice without any consequence for the patient. Although a blood test is cost effective, the isolation, sequencing of biological components of interest and downstream analysis might be seen as expensive. Nevertheless, the sequencing cost keeps decreasing since the first sequenced human genome in 2001. Whole-exome sequencing can, nowadays, be done for less than $1000.

The working procedure of liquid biopsy for cancer detection is based on the interaction between tumors and bloodstream [9]. Indeed, it is likely to discover Circulating Tumor Cells (CTCs), which in turn may exchange information with "tumor-Educated Platelets" (TEPs) [28, 29]. TEPs can provide tumor's genetic material, more precisely RNA, which can lead to the study of the tumor's gene expression. Furthermore, these blood components are "big" enough to be easily extracted and analyzed by means of biological procedures.

2.2. Bioinformatics pipeline to develop the case study

Despite the promising results of liquid biopsy in cancer detection, there are still only a few datasets available for lung cancer that come from this methodology. In our particular case, we refer to the study conducted by Best et al. in 2015 [28], and subsequently extended in 2017 [29]. Both works are based on a pipeline called *ThromboSeq*, which starts from differentially spliced RNA from (TEPs) to create the gene expression matrices.

Taking a closer look into the extraction process, the followed procedure was as follows. First, the total blood TEP RNA was isolated from the cancer patients and the total platelet RNA from the non-cancer. Next, the RNA was then subjected to complementary DNA (cDNA) synthesis and amplification according to the standard protocol. This cDNA is of high importance as it is the input material for Next Generation Sequencing (NGS) methods, i.e. to extract the whole information from the genes.

As soon as this "raw" initial data is collected, the first step is the library preparation. This part includes fragmentation of the cDNA, barcode labeling (important for the sequencing process), and product PCR amplification. Once this preparation is completed, the library was sequenced with the use of an Illumina HiSeq 2500 machine. The result of the sequencing, i.e sequence of each cDNA, is written to a file. The data analysis started by removing adaptors from each cDNA sequence and filtered out low quality ones. Then aligned each cDNA to the reference genome (hg19) to identify which sequence belongs to which location in the genome. The transcription levels of the genes are directly proportional to the number of sequences aligned to it (read counts).

This type of dataset is called gene expression count matrix. Specifically, the "presence" of each gene is represented by the number of RNA sequences belonging to each gene (discrete values). Gene expression is a regulatory process by which information goes through the following stages: from DNA to RNA, then to proteins. Therefore, the expression of a given gene is frequently estimated by the abundance of the RNA transcribed from that gene by the cells. Thus, the more RNA of a gene A is detected, the more gene A is expressed. In this particular case study, a subset of 4,635 genes has been monitored in the TEPs (gene transcripts or features), from the number of up to 25,000 genes that the human genome approximately contains. Finally, this count matrix was normalized by trimmed mean of Mvalues (TMM) [30], which is the recommended procedure in the event of gene abundance data, setting the range of each gene/attribute between 0.0 and 2.0 approximately.

Once this procedure is completed, each one of the samples is linked with the clinical data in order to set the diagnosis label as lung cancer or non-cancerous individuals. We must stress that this dataset is focused on NSCLC without taking the subtypes into account. The final dataset contained 779 samples/individuals, 402 of which were diagnosed with lung cancer, and the remaining 377 as non-cancerous cases.

The dataset is freely available on Gene Expression

Omnibus (GEO) of the NCBI[1] under accession code GSE89843 [29].

3. Addressing High Dimensional Gene Expression Data via Feature Selection

When addressing any classification problem, the success of any ML algorithm is strongly dependent on the dataset's internal characteristics [31]. If we focus our attention on the area of bioinformatics, there are several issues that may hinder the classification ability such as the imbalanced problem (uneven class distribution), the curse of dimensionality (a high number of input attributes), or the class overlapping (same a priori class probability within a small cluster) [32].

Among these issues, in this work contribution, we focus on the so-called high dimensionality problem [33]. The most straightforward solution to overcome this problem is by means of preprocessing approaches, namely, feature selection methods. The objective of these approaches is to reduce the full feature set to a smaller subset with a similar (or even better) predictive capability.

There are different types of schemes to carry out this procedure. In the remainder of this Section, we will describe in detail the different approaches we have selected for the experimental study, namely univariate methods (Section 3.1), embedded model-based solutions (Section 3.2), and Recursive Feature Elimination (REF) (Section 3.3).

3.1. Univariate feature selection

The first group of approaches to be taken into account in this study are filtering techniques based on computing a ranking of the features depending on different scoring functions. Specifically, all attributes are considered independently based on univariate statistical tests. In this work, the selected scoring functions are the following ones:

1. *Chi-squared (Chi2)* [34]. It determines if the association between two categorical variables of the sample would reflect their real association in the population. In the case of feature selection, this test computes the dependence between any variable and the class, in order to detect whether it is relevant or not for the classification. The formula for Chi2 is shown next:

$$X^2 = \frac{(O_f - E_f)^2}{E_f} \tag{1}$$

where O_f refers to the observed frequency, i.e. the number of observations of class, and E_f is the expected frequency, that is, the number of expected observations of class if there was no relationship between the feature and the target. Since Chi2 is defined only for nominal variables, a *binning* discretization is applied [35].

2. *Correlation Score (CORR)* [36]. The filtering in this case is carried out using the correlation matrix via Pearson formula. Specifically, this value $\rho_{j,j'}$ is computed for each pair of attributes j and j', and subsequently computes the lowest absolute correlation, as follows:

$$CFS(j) = \min_{j'} |\rho_{j,j'}|. \tag{2}$$

3. *Mutual Information (MI)* [37]. It computes the amount of information about one attribute that can be gained by observing another one, as follows:

$$MI(j) = \sum_{y \in \mathbf{y}} \sum_{x \in \mathbf{x}_j} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{3}$$

where x and y are the various levels of attribute $\mathbf{x}_j$ and the target vector $\mathbf{y}$, respectively; whereas $p(x)$ and $p(y)$ are their marginal probability distributions, with $p(x,y)$ being their joint distribution. As it can be noted, this approach assumes that the covariates are nominal variables. As in the case of Chi2, MI can be used with numerical variables after binning them [35].

4. *F-Value Classification (FC)* [38]. The F-value scores examine if, when we group the numerical feature by the target vector, the means for each group are significantly different.

   In other words, it does a hypothesis testing model X and Y where X is a model created by just a constant and Y is the model created by a constant and a feature. The least square errors in both the models are compared and check if the difference in errors between model X and Y are significant or introduced by chance.

3.2. Embedded Model-based Selection

In practice, there are several classifiers from which the information on the contribution of each variable can be exported after the learning stage. Specifically, two different attributes can be considered depending on the

[1] https://www.ncbi.nlm.nih.gov/gds

used paradigm. On the one hand, the feature importance, measured for example the actual influence of each feature on the final inference, i.e. the depth in a decision tree. On the other hand, the estimated coefficients, which can be also a reliable measure regarding the absolute value.

Among different classification methods, we have selected three relevant alternatives, namely Linear Support Vector Machines (SVM) [39], Random Forest (RF) [40], and LASSO [41].

3.3. Recursive Feature Elimination

Standard univariate filtering methods have the drawback of not considering a global view of the system, but a local computation for each attribute in isolation. A similar problem is related to model-based feature selection since there can be a high correlation among several highly ranked variables.

Therefore, we must excel in the advantages of wrapper approaches. They consist of an iterative procedure that involves the classification model for determining the most appropriate features for the prediction.

Among different alternatives, we have focused on RFE, which is an instance of backward feature elimination [42], thus starting with the full set of attributes and removing, in each step, the worst attribute remaining in the set. For computational reasons, it may be more efficient to remove several features at a time, at the expense of possible classification performance degradation. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

In order to be able to rank the features according to their importance, an external classification algorithm must be taken into account. In this case study, we will use RFE with two different estimators or classifiers, namely Logistic Regression (LR) [43] and Random Forest (RF) [23].

4. Experimental Framework

In this section we will analyze the behavior of the different approaches for feature selection, as introduced in Section 3. To do so, we will use as baseline classifier two different Soft Computing approaches, namely the Gaussian Naïve Bayes [24] and the Random Forest ensemble [23].

The former is probably the most well-known probabilistic classifier that is based, as its name suggests, in the Bayes' theorem of conditional probability with independence among the features given the class label. This issue implies the method to be a very appropriate solution in the event of high dimensional problems [25]. The decision rule for assigning the output class is given in Eq. 4.

$$\hat{y} = \underset{k \in \{1,\dots,K\}}{\mathbf{argmax}}\, p(C_k) \prod_{i=1}^{n} p(X_i|C_k) \tag{4}$$

In order to compute the parameter estimations, namely the class prior probability values $p(X_i|C_k)$, non-parametric models for the features must be obtained from the training set. For the particular case of continuous data, values associated with each class are frequently assumed to follow a normal (Gaussian) distribution, which is the alternative selected in our study. For any observation value v and given a class C_k, the following formula is applied (Eq. 5)

$$p(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\lambda_k)^2}{2\sigma_k^2}} \tag{5}$$

The Random Forest algorithm is a robust combination among the construction of uncorrelated trees using CART [44], bootstrap aggregation (bagging), and the random subspace method. As such, this learning scheme divides the original data into different subsets (with replacement) and learns a different decision tree for each one of them. In order to add more diversity between the classifiers, a number of random variables are also selected with each "bag".

In order to contrast the quality of the different feature selection alternatives, we have selected two complementary metrics. On the one hand, the standard predictive accuracy and, on the other hand the Area Under the ROC Curve (AUC) [45]. In the first case, i.e. accuracy, only correct hits are taking into account, disregard the output class. This way, we may evaluate whether the learned model is able to globally discriminate the concepts under study. In the second case, i.e. probabilistic AUC, the confidence degrees of each output are taken into account. In such way, a model in which true positives relay on high confidences, whereas false positives are related to low confidences, will present a higher performance value.

The estimates for both metrics will be obtained by means of a Stratified Fold Cross-Validation. Specifically, this procedure is carried out using 5 folds and, in accordance with the stochastic nature of the learning methods, each one of the 5-fold cross-validation is run 3 times. Therefore, experimental results for each method and dataset are computed with an average of 15 runs.

The rest of the experimental framework that we have set up for the thorough comparison of feature

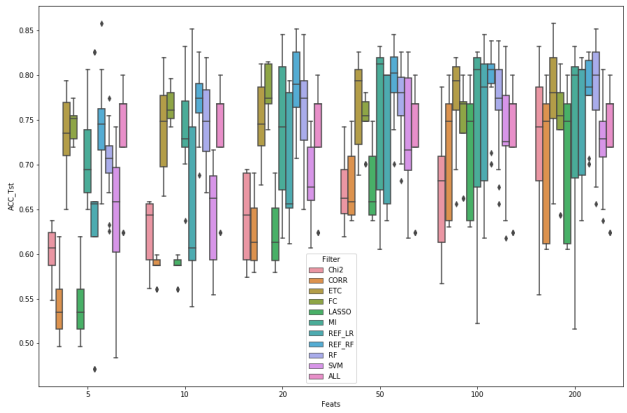
selection techniques is described as follows. First, we have carried out the learning procedure with the whole dataset, namely using all the input variables / genes, in order to establish the baseline performance values. Then, we have considered several thresholds for the number of variables selected, from just 5 genes per dataset, implying an interpretable final system, to a moderate value of 200 genes per problem. Specifically, the range of values is $\{5, 10, 20, 50, 100, 200\}$.

Finally, we must state that all the implementations for carrying out the experimentation, have been taken from the well-known Python ML scikit-learn library [46]. Parameters have been set by default implying, in the case of Random Forest, a number of 100 estimators, gini criterion for computing each split, and all trees to be expanded until all leaves are pure or contain less than 2 examples.

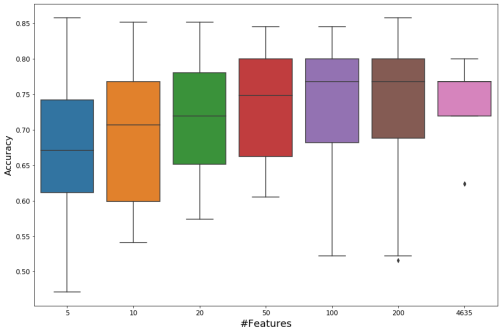
5. Analysis and Discussion of Results

Experimental results in the test partitions are shown by means of boxplots for the sake of comprising the most significant statistical information from the 15 runs per experiment. Figures 1 and 2 contains the results for the Naive Bayes classifier, and Figures 3 and 4 for the Random Forest. Each illustration is divided into three parts with aims at showing (a) a complete comparison for all feature selection methods regarding the different thresholds; (b) to contrast the best number of features for the learning task; and (c) to analyze the behavior of each feature selection method disregard the number of variables that is chosen. For the sake of including the baseline results with the complete set of features, these are noted as "ALL".

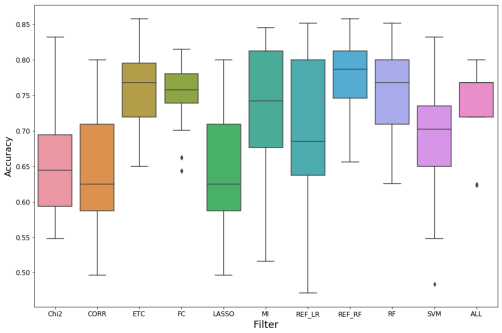
Focusing on this graphical representation of the results, we may observe a similar trend independently of the metric used, namely accuracy and AUC, and classifier. Focusing first on univariate methods, there are two different groups of approaches. On the one hand, those based on Chi2 and correlation (CORR) provide very poor solutions, even in the case of a moderate number of variables (200). On the other hand, those based on the F-test from ANOVA (FC) and Mutual Information (MI) are very competitive with the remaining options. In the particular case of FC, its behavior is especially excelled in the case of setting the threshold for a very low number of variables (5-20). Comparing the two methods/classifiers selected for the RFE, the best approach is by far the one based in RF, being the highest performing choice overall. Finally, contrasting the different alternatives for feature selection based on the models, i.e. ETC, LASSO, RF and SVM, those consisting on the ensemble schemes are the most



(a) Comparison of the 10 feature selection techniques



(b) Comparison grouped by number of features selected

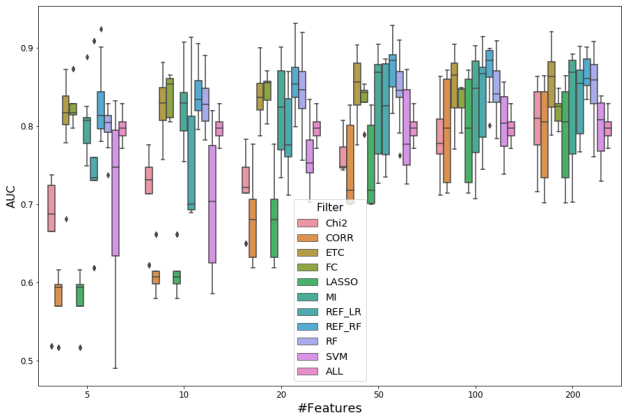


(c) Comparison grouped by feature selection technique

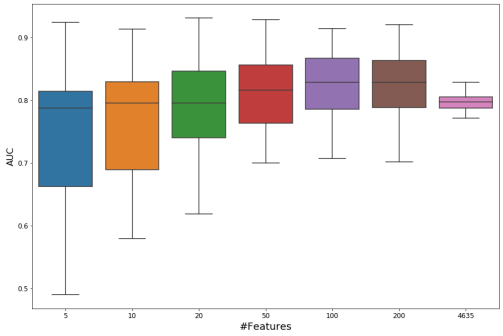
Figure 1: Boxplots with accuracy test results for Naive Bayes.

remarkable ones, that is, ETC and RF, being both of them equally competitive.

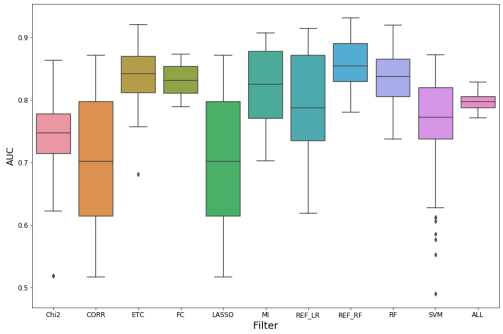
When we evaluate the goodness of the feature selection techniques versus the baseline approach, namely considering the whole 4,635 attributes, some small differences might be emphasized depending on whether accuracy or AUC is considered. In the former case, we may observe a quite similar performance for the lower number of variables, but then a significant boost in the case of the ensemble models, i.e. REF_RF, ETC and RF (filter approach) from 20 variables



(a) Comparison of the 10 feature selection techniques

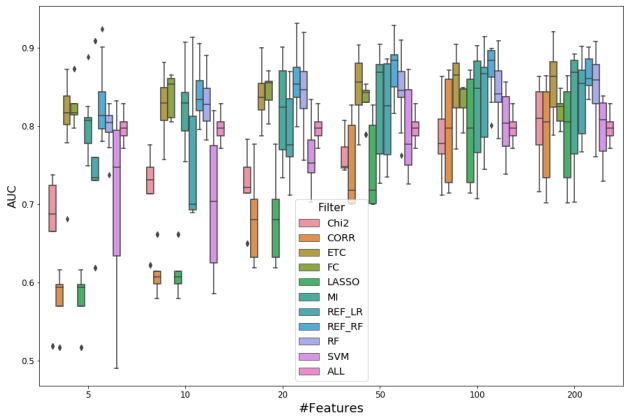


(b) Comparison grouped by number of features selected

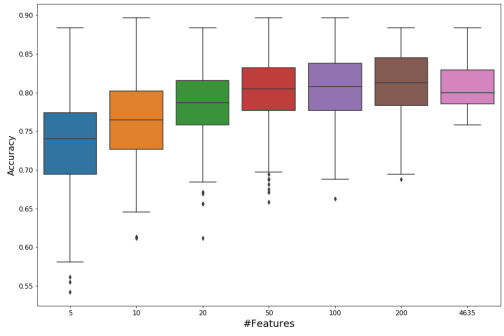


(c) Comparison grouped by feature selection technique

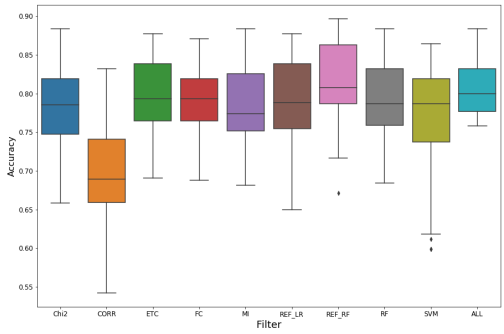
Figure 2: Boxplots with AUC test results for Naive Bayes.



(a) Comparison of the 10 feature selection techniques



(b) Comparison grouped by number of features selected

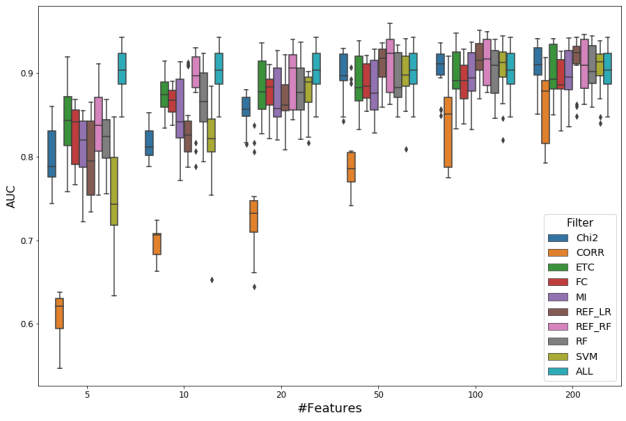


(c) Comparison grouped by feature selection technique

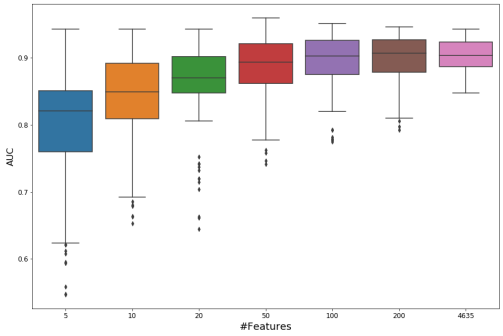
Figure 3: Boxplots with accuracy test results for Random Forest.

henceforth. Additionally, and as stressed in the general case study, FC univariate method is a very suitable approach in the case of a low number of variables. In the latter case, the baseline approach (all variables) is outperformed for all selected feature selection techniques. Specifically, in the case of 20 and 50 variables, REF_RF obtains up to ten points of higher performance, thus confirming the need for a proper preprocessing approach.

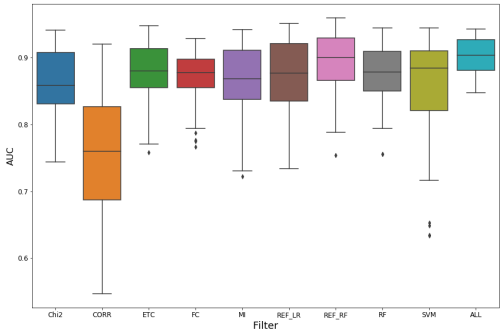
Continuing with the former analysis, very interesting conclusions can be extracted when focusing on the illustrations that represent the grouped comparisons, i.e. part (b) and (c) in all Figures. Specifically, observing the performance regarding the number of selected features (part (b)) very good properties can be highlighted from 20 to 200 variables, i.e. just considering from the 0,5% to the 5% of the original variables. Furthermore, when combining the results per feature selection technique disregard the threshold (part (c)), there is a clear superior behavior in the cases of ETC, FC, REF_RF and RF.



(a) Comparison of the 10 feature selection techniques



(b) Comparison grouped by number of features selected



(c) Comparison grouped by feature selection technique

Figure 4: Boxplots with AUC test results for Random Forest.

To conclude this experimental study, we compile the best results, i.e. baseline approach (ALL), REF_RF and RF for Naive Bayes and Random Forest classifiers in Table 1. The information shown include, for each the number of selected features, the average values of the accuracy and AUC, together with the standard deviation for the training and test partitions. We also show the elapsed learning time for the whole procedure, i.e. from the feature selection to the problem transformation and the generation of the classifier.

Table 1: Comparison of the accuracy and AUC results for the best performing feature selection techniques, REF_RF and RF. Baseline results (ALL) are also included.

Classif.	Filter	#Feats.	Acc_{Tr}	Acc_{Tst}	AUC_{Tr}	AUC_{Tst}	Time
NB	ALL	4635	.8123 ± .0073	.7359 ± .0636	.8671 ± .0158	.7981 ± .0196	.2873 ± .0591
	REF_RF	5	.7687 ± .0258	.7422 ± .0564	.8548 ± .0187	.8269 ± .0417	6,731.5 ± 218.08
		10	.7961 ± .0139	.7700 ± .0379	.8714 ± .0083	.8442 ± .0339	6,894.6 ± 179.66
		20	.8158 ± .0105	.7876 ± .0461	.8871 ± .0100	.8594 ± .0378	6,988.3 ± 71.65
		50	.8324 ± .0127	**.7926 ± .0399**	.9139 ± .0101	**.8738 ± .0337**	6,958.2 ± 61.94
		100	.8424 ± .0077	.7879 ± .0449	.9269 ± .0060	.8729 ± .0305	6,939.5 ± 61.74
		200	**.8449 ± .0089**	.7807 ± .0439	**.9289 ± .0095**	.8675 ± .0221	6,919.9 ± 485.23
	RF	5	.7452 ± .0432	.7047 ± .0413	.8449 ± .0218	.8014 ± .0241	2.2809 ± .3055
		10	.7777 ± .0245	.7477 ± .0462	.8631 ± .0151	.8276 ± .0301	2.2849 ± .0470
		20	.8022 ± .0184	.7602 ± .0554	.8787 ± .0119	.8427 ± .0460	2.2863 ± .0374
		50	.8153 ± .0139	.7687 ± .0419	.8902 ± .0066	.8483 ± .0396	2.2848 ± .0252
		100	.8218 ± .0109	.7683 ± .0544	.8967 ± .0123	.8466 ± .0357	2.1739 ± .0945
		200	.8225 ± .0063	.7799 ± .0629	.9048 ± .0093	.8474 ± .0427	**2.0740 ± .0645**
RF	ALL	4635	1.000 ± .0000	.8085 ± .0355	1.000 ± .0000	.9002 ± .0304	2.1180 ± .0193
	REF_RF	5	**1.000 ± .0000**	.7622 ± 0.0474	**1.000 ± .0000**	.8396 ± .0460	6,600.7 ± 271.60
		10	**1.000 ± .0000**	.8200 ± 0.0559	**1.000 ± .0000**	.8870 ± .0461	6,959.7 ± 60.09
		20	**1.000 ± .0000**	.8200 ± 0.0409	**1.000 ± .0000**	.8944 ± .0356	7,154.4 ± 91.22
		50	**1.0000 ± .0000**	**.8350 ± .0452**	**1.000 ± .0000**	.9119 ± .0350	7,416.9 ± 232.11
		100	**1.000 ± .0000**	.8333 ± .0409	**1.000 ± .0000**	**.9143 ± .0284**	7,250.5 ± 463.62
		200	**1.000 ± .0000**	.8294 ± .0416	**1.000 ± .0000**	.9108 ± .0303	6,662.4 ± 400.73
	RF	5	.9996 ± .0007	.7424 ± .0348	**1.000 ± .0000**	.8190 ± .0344	**2.5019 ± .0435**
		10	**1.000 ± .0000**	.7841 ± .0452	**1.000 ± .0000**	.8668 ± .0369	2.5088 ± .0430
		20	**1.000 ± .0000**	.7965 ± .0459	**1.000 ± .0000**	.8790 ± .0349	2.5581 ± .0311
		50	.9999 ± .0004	.8067 ± .0356	**1.000 ± .0000**	.8934 ± .0311	2.6438 ± .0374
		100	**1.000 ± .0000**	.8118 ± .0380	**1.000 ± .0000**	.8995 ± .0309	2.5740 ± .1055
		200	**1.000 ± .0000**	.8200 ± .0379	**1.000 ± .0000**	.9059 ± .0300	2.5843 ± .0834

From this complete Table of results, some interesting findings may be highlighted. First, the high overfitting in the case of the baseline case study (ALL) with respect to the use of feature selection techniques, in which the differences between the training and test results are almost minimal. Second, to confirm the excellent performance of both approaches, i.e. REF_RF and RF, especially from 20 variables, which can be seen as an interpretable number of genes for an in-depth study by the clinician. Finally, when we analyze the efficiency of the different feature selection techniques, there are clear differences between the REF and the model-based selection. The computational complexity of the recursive approach, although obtaining the best solution in terms of predictive ability, is probably not worth it as it is up to 3,500 times slower than to apply RF directly.

6. Concluding Remarks

In the field of biomedical applications, it is just as important to obtain high precision as to make the models generated to be explainable to clinical staff. For this reason, it is essential to apply intelligent techniques that are capable of learning effectively in these scenarios. In this work, we have focused our efforts on one of the most important problems in medical diagnosis, i.e., the detection of lung cancer. To this end, we have selected genetic data extracted using liquid biopsy technology, which allows us to have genetic information of almost 5,000 different characteristics of the genome of each sample or patient.

In order to obtain quality solutions, as well as identifying the most relevant genes that are linked with lung cancer condition, we have carried out a

thorough experimental study to contrast the behavior of different techniques for feature selection. Specifically, we have used univariate techniques, recursive feature elimination, and a selection based on different learning models such as Random Forest. Finally, both a Gaussian Naïve Bayes and Random Forest classifiers have been used for the prediction.

Different lessons have been learned from this case study. Firstly, it ratifies the need to apply any of the techniques for feature selection, since they have improved the predictive capacity of the classifier with the original data. Secondly, it has been shown that quality results can be achieved with a threshold of between 20 and 50 genes, which is a value manageable by the final expert. Finally, among different methods for selecting characteristics, those based on ensemble models (Random Forest and ETC) stand out, especially based on a simple ranking according to the importance of the characteristics due to their lower computational complexity.

In accordance with the current experimental results, we plan to extend our research by designing a more sophisticated feature selection procedure aiming to minimize the number of biomarker genes. Additionally, we must study in-depth the effect of the confounding clinical variables in the prediction of the classifier. Finally, there is a need for including more datasets related to liquid biopsy to provide additional support to the findings extracted in this area of study.

7. Acknowledgments

This work has been partially supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Actions (grant agreement ELBA, 765492), and by the Spanish Ministry of Science and Technology under project TIN2017-89517-P, including European Regional Development Funds.

References

[1] L. Cao, "Data science: A comprehensive overview," *ACM Computing Surveys*, vol. 50, no. 3, 2017.

[2] Z. Ge, Z. Song, S. Ding, and B. Huang, "Data mining and analytics in the process industry: The role of machine learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017.

[3] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[4] P. Sajda, "Machine learning for detection and diagnosis of disease," *Annual Review of Biomedical Engineering*, vol. 8, pp. 537–565, 2006.

[5] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[6] S. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–86, 2002.

[7] R. Siegel, K. Miller, and A. Jemal, "Cancer statistics, 2018," *CA Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[8] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, pp. 394–424, nov 2018.

[9] I. Domínguez-Vigil, A. Moreno-Martínez, W. J.Y., M. Roehrl, and H. Barrera-Saldaña, "The dawn of the liquid biopsy in the fight against cancer.," *Oncotarget*, vol. 9, pp. 2912–2922, jan 2017.

[10] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[11] D. Castelvecchi, "Can we open the black box of AI?," *Nature*, vol. 538, no. 7623, pp. 20–23, 2016.

[12] A. Fernandez, M. J. del Jesus, O. Cordon, F. Marcelloni, and F. Herrera, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?," *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, 2019.

[13] M. Hamburg and F. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, no. 4, pp. 301–304, 2010.

[14] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys*, vol. 50, no. 6, 2017.

[15] I. Jeffery, D. Higgins, and A. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, 2006.

[16] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

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

[18] K.-B. Duan, J. Rajapakse, H. Wang, and F. Azuaje, "Multiple svm-rfe for gene selection in cancer classification with expression data," *IEEE Transactions on Nanobioscience*, vol. 4, no. 3, pp. 228–233, 2005.

[19] S. Liu, C. Xu, Y. Zhang, J. Liu, B. Yu, X. Liu, and M. Dehmer, "Feature selection of gene expression data for cancer classification using double rbf-kernels.," *BMC Bioinformatics*, vol. 19, no. 1, pp. 396:1–396:14, 2018.

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

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

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

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] H. Zhang, "The optimality of naive bayes," vol. 2, pp. 562–567, 2004.

[25] P. Bickel and E. Levina, "Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations," *Bernoulli*, vol. 10, no. 6, pp. 989–1010, 2004.

[26] P. B. Bach, J. N. Mirkin, T. K. Oliver, C. G. Azzoli, D. A. Berry, O. W. Brawley, T. Byers, G. A. Colditz, M. K. Gould, J. R. Jett, A. L. Sabichi, R. Smith-Bindman, D. E. Wood, A. Qaseem, and F. C. Detterbeck, "Benefits and harms of CT screening for lung cancer: a systematic review.," *JAMA*, vol. 307, pp. 2418–29, jun 2012.

[27] D. Sharma, T. G. Newman, and W. S. Aronow, "Lung cancer screening: history, current perspectives, and future directions.," *Archives of medical science : AMS*, vol. 11, pp. 1033–43, oct 2015.

[28] M. Best et al., "RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics," *Cancer Cell*, vol. 28, pp. 666–676, nov 2015.

[29] M. G. Best et al., "Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets.," *Cancer cell*, vol. 32, pp. 238–252.e9, aug 2017.

[30] M. Pereira, M. Wallroth, V. Jonsson, and E. Kristiansson, "Comparison of normalization methods for the analysis of metagenomic gene abundance data," *BMC Genomics*, vol. 19, no. 1, 2018.

[31] S. Das, S. Datta, and B. B. Chaudhuri, "Handling data irregularities in classification: Foundations, trends, and future challenges," *Pattern Recognition*, vol. 81, pp. 674–693, 2018.

[32] A. Fernández, C. J. Carmona, M. J. del Jesús, and F. Herrera, "A pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets," *Int. J. Neural Syst.*, vol. 27, no. 6, pp. 1–21, 2017.

[33] E. Keogh and A. Mueen, *Curse of dimensionality*, pp. 257–258. Springer, 2nd ed., 2017.

[34] H. Liu and R. Setiono, "Chi2: feature selection and discretization of numeric attributes," pp. 388–391, 1995.

[35] J. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*, vol. 24, pp. 175–186, 2014.

[36] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 359–366, 2000.

[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[38] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. De Schaetzen, R. Duque, H. Bersini, and A. Nowac, "A survey on filter techniques for feature selection in gene expression microarray analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 4, pp. 1106–1119, 2012.

[39] E. Byvatov and G. Schneider, "Support vector machine applications in bioinformatics.," *Applied Bioinformatics*, vol. 2, no. 2, pp. 67–77, 2003.

[40] Y. Saeys, T. Abeel, and Y. Van De Peer, "Robust feature selection using ensemble feature selection techniques," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5212 LNAI, no. PART 2, pp. 313–325, 2008.

[41] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[42] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[43] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[44] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Chapman and Hall (Wadsworth and Inc.), 1984.

[45] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.