UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

BERNARDO TREVIZAN

Homogeneous ensemble feature selection for breast cancer biomarker identification from microarray data

Advisor: Prof^{a}. Dra. Mariana Recamonde Mendoza

Porto Alegre
Maio 2021

To my beloved sister

Para a minha amada irmã

AGRADECIMENTOS

Eu agradeço aos meus pais por sempre incentivar os meus estudos e por fornecer todo o suporte, principalmente emocional, para que eu chegasse até aqui. Vocês são o meu maior tesouro. Meu agradecimento e admiração à minha orientadora, Profa. Dra. Mariana, por todos os ensinamentos, pelo apoio desde a concepção à escrita deste trabalho e pelos cinco anos que trabalhamos juntos. Ao Rodrigo Haas Bueno pelo suporte técnico no processamento dos dados utilizados neste trabalho, muito obrigado. Ao meu amigo de infância, doutorando em Computação, Gabriel Paludo Licks, pelas correções na escrita e por todo o apoio, muito obrigado. Eu agradeço, também, todos os discentes, docentes e servidores da UFRGS pelo apoio durante toda a minha graduação. A todas as pessoas envolvidas, a minha maior admiração e o meu muito obrigado.

ABSTRACT

In precision medicine, the identification of biomarkers could help speed the diagnosis and tailor the treatment to each patient increasing the quality of health care. Omics data, such as microarray, generates high-dimensional data that has enabled the analysis of genes expression profiles to extract candidate biomarkers. However, high-dimensional data requires advanced computational methods for data analysis. In this work, we proposed a homogeneous ensemble feature selection (EFS) strategy to identify candidate biomarkers for breast cancer from multiple microarray datasets. We applied the state-of-the-art random effect model from meta-analysis as a comparison method. We also compared five feature selection (FS) methods as base selectors and four classification algorithms. Our results showed that FS method *variance* is the most stable among other FS methods. We showed that stability is higher within datasets than across datasets, indicating high sample heterogeneity among studies. The top 20 genes selected by variance showed the best trade-off between the number of selected genes and performance. Our approach outperform meta-analysis in four out of six independent microarray studies evaluated. Support Vector Machine classifier presented, in general, the best mean F1-Scores and K-Nearest Neighbors classifier the best mean Recall values. We conclude that homogeneous EFS is a promising methodology for candidate biomarkers identification, demonstrating stability and predictive performance as good as the reference statistical method.

Keywords: Feature selection. microarray. biomarker. breast cancer.

Seleção de atributos com um ensemble homogêneo a partir de dados de microarranjo para identificação de biomarcadores de câncer de mama

RESUMO

Na medicina de precisão, a identificação de biomarcadores pode ajudar a agilizar o diagnóstico e adequar o tratamento a cada paciente, aumentando a qualidade da assistência à saúde. Dados ômicos, como os de microarranjo, geram dados de alta dimensionalidade que permitem a análise de perfis de expressão gênica para extrair cadidatos a biomarcadores. No entanto, dados de alta dimensionalidade requerem métodos computacionais avançados para análise de dados. Neste trabalho, propusemos uma estratégia de seleção de atributos com um *ensemble* (EFS) homogêneo para identificar candidatos a biomarcadores para câncer de mama a partir de múltiplos dados de microarranjo. Aplicamos o método de meta-análise *random effect model* como método de comparação. Também comparamos cinco métodos de seleção de atributos (FS) como seletores base e quatro algoritmos de classificação. Nossos resultados mostraram que o método de FS *variância* é o mais estável entre os outros métodos de FS. Mostramos que a estabilidade é maior dentro dos conjuntos de dados do que entre os conjuntos de dados, indicando alta heterogeneidade entre os estudos. Os 20 genes mais informativos selecionados por variância apresentaram a melhor troca entre o número de genes selecionados e o desempenho. Nossa abordagem superou a meta-análise em quatro dos seis estudos independentes de microarranjo avaliados. O classificador Support Vector Machine apresentou, em geral, os melhores valores médios de *F1-Score* e o classificador K-Nearest Neighbors os melhores valores médios de *recall*. Concluímos que o EFS homogêneo apresentado é uma metodologia promissora para a identificação de candidatos a biomarcadores, demonstrando estabilidade e desempenho preditivo tão bom quanto o método estatístico de referência.

Palavras-chave: seleção de atributos, microarranjo, biomarcadores, câncer de mama.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

DEA	Differential Expression Analysis
DNA	Deoxyribonucleic acid
EFS	Ensemble Feature Selection
FS	Feature Selection
FDA	United States Food and Drug Administration
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GSEA	Gene Set Enrichment Analysis
HGNC	HUGO Gene Nomenclature Committee
ID	Identification
J48	Decision Tree algorithm
KNN	K-Nearest Neighbors
LOOCV	Leave-one-out cross-validation
ML	Machine Learning
mRMR	Minimum Redundancy Maximum Relevance
NIH	National Institute of Health
NNET	Neural Network
PCA	Principal Component Analysis
REM	Random Effect Model
RMA	Robust Multi-array Average
SMOTE	Synthetic Minority Oversampling Technique
SVA	Surrogate Variable Analysis
SVM	Support Vector Machine
SVM-RFE	Support Vector Machine with Recursive Feature Elimination
WHO	World Health Organization

LIST OF SYMBOLS

χ^2	*Chi* squared statistic test
ϕ_c	Cramér's V measure of correlation
G_k	Set of expressions for the k^{th} gene
$logFC$	Log fold change
TP	True positive
FP	False positive
TN	True negative
FN	False negative
T_i	Training dataset i
T	Collection of training datasets
E_i	Test dataset i
E	Collection of test datasets
L_i	Local ranking i
L	Collection of local rankings

CONTENTS

1 INTRODUCTION

Cancer occurs when mutations in a cell replication process is not reverted, *i.e.,* the mutated cell does not die. The mutation is replicated and can form a benign or malignant tumor, which can negatively affect the organism essential functions. External factors such as the presence of radiation, ingestion of chemicals (smoke, alcohol, water or food contaminants), viruses, bacteria, or parasites could cause cancer. Smoking, for example, is the main reason for lung cancer cases, which has been diagnosed on over 2 million people and killed 1.8 million in 2020, according to the World Health Organization (WHO)[1]. Nearly 10 million deaths were caused by cancer in 2020, making the disease a leading cause of death. Breast cancer was the most common type of cancer with 2.3 million cases in 2020. There are many factors that could cause breast cancer. According to WHO, female gender is the strongest risk factor. Less then 1% of breast cancer cases are found in men. Family history and certain gene mutations could also increase the risk of breast cancer.

The disease arises from tumor cells in the breast glandular tissue (FENG et al., 2018). In the early stage, breast cancer causes no symptoms and does not harm the patient. However, the patient can have symptoms when tumor spreads to lymph nodes and to other organs in further stages. These further stages are more complicated to treat and could cause death. Therefore, one of the strategies to reduce breast cancer mortality is early detection. Finding a lump or thickening in the breast through self-examination is an evidence of breast cancer. In this case, breast cancer can be confirmed by a health practitioner through imaging and biopsy. The former is the most common way to detect tumors in the breast. Imaging could be followed by biopsy to confirm if the tumor is malignant or benign. A positive diagnosis - when cancer is present - is treated with radiation and, in more advanced cases, surgical removal of the breast.

WHO states that certain gene mutations indicate a greater risk of breast cancer - *BRCA1*, *BRCA2*, and *PALB-2*[2]. We go further in this work, aiming at identifying a small group of genes that can explain the presence of breast cancer, called biomarkers. A gene-level diagnosis could tailor the treatment to each patient reducing its side effects and increasing its effectiveness. This process is called precision medicine. Precision medicine has become feasible with the growth of digital medical records and high-throughput diagnosis devices (HODSON, 2016). High-throughput technologies have helped to un-

[1] <https://www.who.int/news-room/fact-sheets/detail/cancer>

[2] <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>

derstand diseases in a molecular level through the generation of genomic data, such as large-scale profiling of gene expression (*i.e.,* transcriptome) generated with microarray. Other types of genomic data exist and they refer to molecular profiling data comprising all genes, proteins, *etc.*, for several patients grouped by disease or condition. However, these recent technologies are expensive and genomic data analysis methods are not precise enough to confirm a diagnosis. Thus, a small group of biomarkers could also reduce the cost of collecting genes' information and increase performance in data analysis.

In medical research, data analysis helps to derive conclusions from medical data creating evidence-based results. Meta-analysis is at the top of the hierarchy of clinical evidence approaches (HAIDICH, 2010). According to the Haidich (2010), "outcomes from a meta-analysis may include a more precise estimate of the effect of treatment or risk factor for disease, or other outcomes, than any individual study contributing to the pooled analysis". Therefore, meta-analysis is a state-of-the-art method for medical research, including extracting hypotheses through omics data analyses. In biomarker identification, meta-analysis groups the relevant genes from different studies according to statistical measures, such as *logFC* and *p-value*, which indicate the degree to which a gene has altered expression among two conditions and its statistical significance. However, a simple statistical method such as meta-analysis may not infer genes' relevancy in high-dimensional data, *i.e.,* genomic data. Hence, the need for more advanced computational methods to identify relevant candidate biomarkers and increase the quality of results.

Recently in literature, studies have applied machine learning feature selection methods to increase the quality of results in dimensionality reduction for low-sample high-dimensional datasets (KHAIRE; DHANALAKSHMI, 2019; BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2016; BOLÓN-CANEDO et al., 2014; HE; YU, 2010; YU; LIU, 2003). Mainly, ensemble feature selection (EFS) has been studied as an alternative approach due to its robustness compared to state-of-the-art feature selection methods (PES, 2019; BOLÓN-CANEDO; ALONSO-BETANZOS, 2019; ALI et al., 2018; Ben Brahim; LIMAM, 2017). Ensembles, in general, are a group of base models each of which independent from each other. Ensemble principles comes from the Wisdom of Crowds theory, which states that a group of diverse individuals are, on average, more correct than a single expert one (SUROWIECKI, 2005). EFS applies feature selection methods - base selectors - to identify relevant features and reduce redundancy in high-dimensional datasets.

Although many studies have reported increased stability through ensemble feature selection application on high-dimensional datasets ,specially for homogeneous EFS (ZHANG; JONASSEN, 2019; Ben Brahim; LIMAM, 2017; SEIJO-PARDO et al., 2017), to our knowledge few studies have compared EFS approaches against meta-analysis. Literature still lacks a comprehensive comparison between these approaches for cancer disease, including for breast cancer. Homogeneous EFS have presented satisfactory results (PES, 2019; BOLÓN-CANEDO; ALONSO-BETANZOS, 2019; SEIJO-PARDO et al., 2017), including other types of cancer, in studies focused on biomarker identification from microarray datasets. Among previously studied methods, homogenenous EFS have increased stability while maintaining performance in several distinct scenarios.

In this work, our goal is to apply a homogeneous EFS to increase stability while maintaining performance for breast cancer biomarker identification from a compendium of microarray datasets. We compare the results with the state-of-the-art method, meta-analysis. Our findings could help: (i) guide the design of a cheaper and faster diagnosis approach; (ii) and, thus reduce the number of deaths caused by breast cancer; (iii) tailor the treatment for each patient by using its genes' profiles; (iv) guide the application of ensemble feature selection as an alternative method to reduce dimensionality; and (v) guide the choice of methods to deal with low-sample high-dimensional data analyses in medical research. On the other hand, important questions still remain unanswered. As an important health issue, we must address these important questions in future studies to assure the quality and safety of our approach for practical usage.

The work is organized as follows. Chapter 2 explain biological (Section 2.1) and computational (Section 2.2) methods and metrics applied. Chapter 3 presents previous findings to guide our EFS design and experimental setup. Chapter 4 explains how data were collected (Section 4.1) and processed (Section 4.2), the experimental pipeline, which includes the EFS (Section 4.3) and meta-analysis (Section 4.4) design and its parameters (Section 4.5). Finally, Chapters 5 and 6 presents our work's findings, its discussion and still unanswered questions.

2 THEORETICAL BACKGROUND

There is a vast set of machine learning, feature selection and evaluation algorithms in the literature (BREIMAN et al., 1984; HAYKIN, 1998; KENT, 1983; KUNCHEVA; RODRÍGUEZ, 2018). In this work, the main object of study are the algorithms applied to solve high dimensional problems, such as classification using microarray datasets. The problem's biological background will also be investigated in order to understand and better interpret the final results. Therefore, the next sections are dedicated to review state-of-the-art algorithms and biological concepts.

2.1 Biological Background

Literature commonly refers to biomarkers and meta-analysis to identify diagnostic genes. However, these concepts must be reviewed to understand the comparison made between state-of-the-art methods for biomarker identification and more advanced computational methods, such as feature selection. The next sections present essential biological concepts that guides the theoretical assumptions in this work.

2.1.1 Precision Medicine and Biomarkers

Through advanced technological tools, precision medicine tailors health care testing and medication to each patient according to its characteristics. Precision medicine has become feasible with the growth of digital medical records and high-throughput diagnosis devices (HODSON, 2016). Since the first sequenced human genome, more recent technologies - such as microarray (Section 2.1.2) - have helped to understand diseases in a molecular level through the genomic data. As stated by Ginsburg and Phillips (2018), "the inclusion of genomic data in a knowledge-generating health care system infrastructure is a way to harness the full potential of that information to optimize patient care". Harnessing the knowledge of high-throughput medical data requires advanced computational methods such as machine learning algorithms to identify molecular patterns in different patients, and thus improve health care quality.

The importance of genomic data lies in the identification of biomarkers. According to a definition published by the U.S. Food and Drug Administration (FDA) and the

National Institute of Health (NIH), biomarkers indicate, in a broad sense, a characteristic in biological or pathogenic processes (FDA-NIH Biomarker Working Group and others, 2016). Among more specific definitions, however, in this work only diagnostic biomarker is taken into account. A diagnostic biomarker indicates the presence or condition of a disease. Therefore, these biomarkers can be used to identify the presence of cancer in patients, and even guide the understanding of different types of cancer. In this sense, biomarkers' stability must be considered. As discussed by He and Yu (2010), stable biomarkers refer to a set of markers representative of the study, *i.e.,* it can be reproduced in other sub-study as being diagnostic biomarkers. Biomarker stability on machine learning algorithms will be further discussed in the next sections.

2.1.2 Omics data

To understand omics data, some concepts must be explained. A DNA is a sequence of coding (genes) and non-coding segments. Genes are responsible for protein production. The activity of a gene in a cell is called expression. Recent high-throughput technologies - such as microarray - allows to extract genes' expressions from a cell. A human cell has around 20.000 genes, which means that microarray can extract 20.000 expressions simultaneously. Alternatively, expressions from different groups - pathological and healthy, for example - can be measured and compared through mRNA microarray. Figure 2.1 presents the process to collect genes' expressions through a microarray chip. As we can see, after processing the collected cells, the resulting biological matter is hybridized onto a microarray chip. A microarray chip is represented by the black square with colored circles, which are the probes used to identify specific genes. In this sense, omics data refers to data extracted by these technologies. Gene expression profiles collected in a large scale (*i.e.,* for all genes) is called transcriptome and it can be used to infer hypothesis about biomarkers. However, with high volume of data generated, large-scale data analysis requires robust computational and statistical methods.

High-throughput technologies generate massive amounts of data. To organize and make these data publicly available, the Gene Expression Omnibus (GEO) serves as a repository of gene expression data, specially for DNA microarray (EDGAR; DOMRACHEV; LASH, 2002). Up to this day, GEO archives more than 4 million samples of gene expression data divided by almost 150.000 series. A series defines the dataset from an experiment and helps to organize data into several biological processes such as toxicol-

Figure 2.1 – Gene expression extraction process from mRNA microarray.

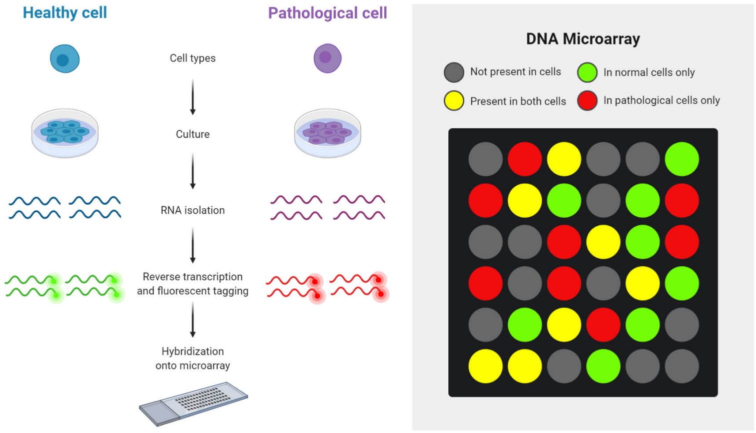


Image By Sagar Aryal, created using biorender.com

Adapted from <https://microbenotes.com/dna-microarray>

ogy and metabolic processes. In this work, DNA microarray extracted from breast cancer tumors and publicly available on GEO will be analyzed.

2.1.3 Transcriptome Meta-analysis

As the proposed methodology in this work, meta-analysis represents another viable approach for biomarker identification. In a broad sense, meta-analysis groups several studies to analyze jointly, adopting approaches to combine their statistics and improve findings. In terms of genomics data, meta-analysis groups results from different studies or series. Therefore, with this process, the same gene from different experiments can be analyzed.

Meta-analysis is a two-step process. The first step of meta-analysis measures the effect size - strength of a phenomenon - through differential expression analysis (DEA). The DEA process applies a statistical test, such as Student's t-test or LIMMA (RITCHIE et al., 2015) for genomic data, to extract a log fold change (*logFC*) and a *p-value* for each gene expression profile. The *logFC* is simply the difference between the average of expressions of pathological cases and the average of expressions of control cases (provided that expression values are represented in a log_2 scale). The *p-value* evaluates the significance of change in expressions between the two groups (*e.g.,* pathological and control). For different studies, there will be different *logFC* and *p-value* for the same gene. Therefore, the second step of meta-analysis is to measure the significance of such genes for all samples in the study, aggregating statistics derived from each study. In this work, Random Effect Model (REM) (DERSIMONIAN; KACKER, 2007) will be applied for such purpose.

Meta-analysis presents different methodologies according to the problem domain, such as rank combination, p-value combination, and effect size combination. In this work, the focus will be on effect size combination due to its performance in meta-analysis for microarray datasets. Methods based on effect size combination models the combined difference of differential expression between pathological and control groups (TORO-DOMÍNGUEZ et al., 2020). Specially REM which assumes different effect sizes between studies, *i.e.,* effect sizes follow a distribution. In this study, microarray data does follow a distribution (Table 2.1) becoming suitable for REM in opposed to fixed effects model that assumes common effect sizes. Although, as mentioned by Toro-Domínguez et al. (2020), the difference between effect sizes must be subtle.

Table 2.1 – Example of the distribution of effect size values (*logFC*) for different studies.

Study ID (GSE)	Gene 79608	Gene 22974	Gene 1308	Gene 9055
38959	-1.347	2.4753	-2.358	2.8467
42568	-0.039	1.6424	0.1014	2.5296
45827	0.1309	-0.061	0.1576	-0.295
53752	0.0046	-0.126	-0.068	-0.131
62944	-1.511	2.9949	-3.581	2.1556
70947	-0.330	1.5939	-0.282	1.5210
7904	-0.111	2.7816	-2.220	2.8474

2.2 Computational Background

Currently, the growing volume of data and the complexity of problems to be solved has made it essential to apply algorithms for classification, prediction, and knowledge discovery from data. This process is called Machine Learning (ML). A ML algorithm trains a model that represents a hypothesis - or a function approximation - from past recorded experiences (FACELI et al., 2011), *i.e.,* a set of examples and their features. An example x_i, such as a cancer patient, is a vector of features $x_i^1, x_i^2, ..., x_i^j$ (*e.g.,* birth year, average sleep time, body fat...) and a known target value y_i which the trained model must learn to predict. The application of a ML algorithm can vary according to the target value's class. Classification algorithms are applied when the target value - value to be predicted from data - is categorical. For example, $normal$ or $tumor$ for cancer classification tasks.

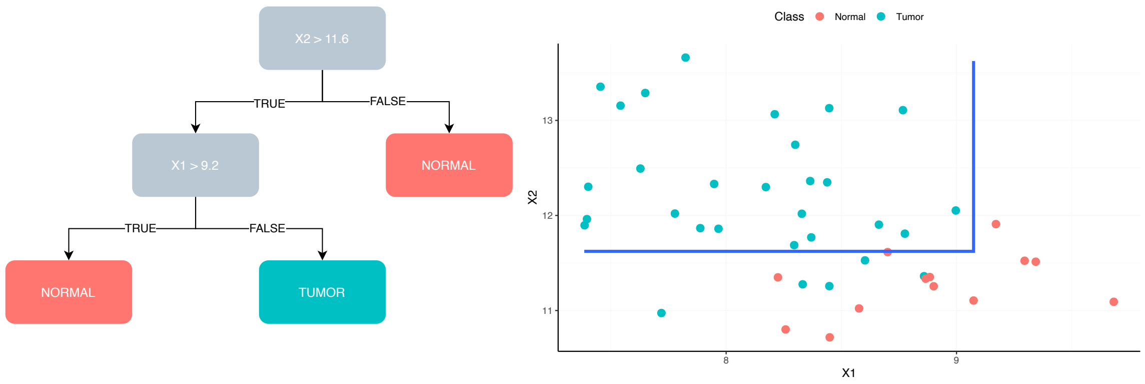
Regression algorithms are applied when the target value can be continuous, such as the number of years a cancer patient will live. Furthermore, a ML algorithm can be classified into supervised and unsupervised learning. The former knows the past experiences' target value while the latter does not have this information and must infer it.

2.2.1 Supervised Learning Algorithms

Regardless of algorithms' classification, each machine learning algorithm has its own inductive bias. In consequence, each algorithm learns from data differently, *i.e.,* different algorithms could infer different hypotheses for the same problem. Usually, the best algorithm to solve a specific problem is unknown. Considering the inductive bias and literature reviews, however, a small set of algorithms could be selected as viable candidates to solve the specified problem. In this sense, the most frequent algorithms found in the literature for cancer classification problems were selected and are defined as follows:

- *Decision Tree* (J48) classification algorithm learns decision rules inferred from training data to predict a target variable. J48 is a C4.5 (BREIMAN et al., 1984) implementation in Java (ARNOLD; GOSLING; HOLMES, 2005), which iteratively selects the feature that best splits the subset (training data for the first iteration) according to a homogeneity metric. An homogeneity metric measures the target value's homogeneity for a subset, such as information gain (see Section 2.2.2). For J48 trees, the algorithm has two parameters: (i) confidence interval C, the minimum gain from splitting a subset; and (ii) M, the minimum number of samples in the leafs. When C or M are not satisfied, the algorithm stops. As a result, a decision tree is returned in which each node represents a rule built with the most informative feature selected and the leafs are the final target value (Figure 2.2a). The decision boundary (Figure 2.2b) shows the algorithms limitations in terms of data representation: J48's decision boundaries are always parallel to features' space.
- K*-Nearest Neighbors* (KNN) is an instance-based algorithm (AHA; KIBLER; ALBERT, 1991) that computes classification through a majority vote among the K nearest neighbors of each new data point. In this sense, KNN does not train a model. However, a representation of its decision boundaries is presented in Figure 2.3b. The classifier is commonly based on the Euclidean distance between the new

Figure 2.2 – Decision Tree representation and its decision boundaries.

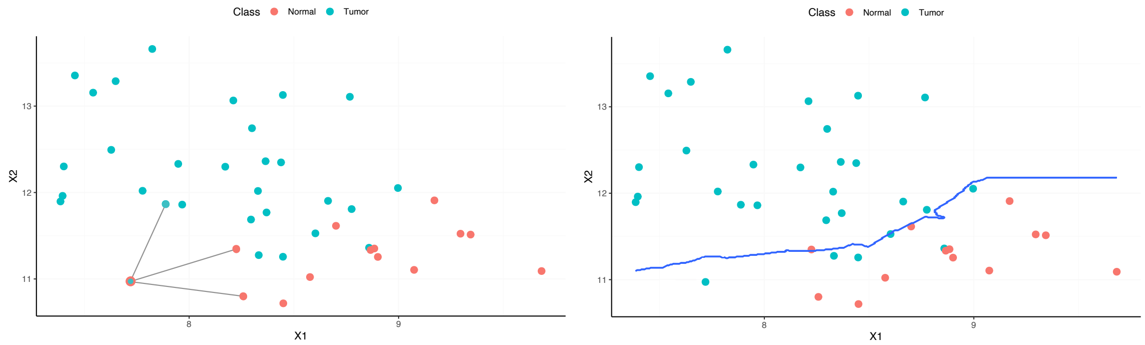


(a) Decision tree representation built with J48 for binary classification.

(b) Decision Tree's decision boundary for binary classification.

sample and the training samples. Therefore, for each new sample, KNN selects the K nearest neighbors. The majority class between the selected neighbors is assigned to the new sample. In the example presented in Figure 2.3a, a sample is wrongly classified with $K = 3$. However, it should not be the case if $K = 5$. Note that if $K = 4$, two neighbors could be from the $tumor$ class and two could be from the $normal$ class. In this scenario, KNN randomly assigns a class to the new sample.

Figure 2.3 – K-Nearest Neighbors' classification and its decision boundaries.

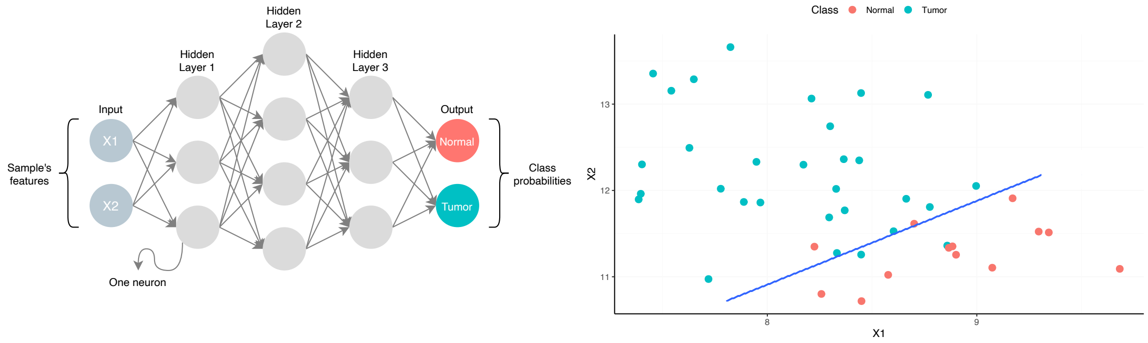


(a) Classification of a new sample with $K = 3$.

(b) K-Nearest Neighbors' decision boundary for binary classification.

- *Neural Network* (NNET) for classification uses a multi-layered perceptron algorithm that trains a model using batch gradient descent (HAYKIN, 1998). A simple neural network algorithm has four parameters: (i) n, the number of hidden layers; (ii) m, the number of neurons for each layer; (iii) α, the learning rate for gradient descent; and (iv) the activation function. Each input neuron represents a feature for a given sample - $X1$ and $X2$ in Figure 2.4a. 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 applied to an activation function. 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. This process is called *back-propagation*. In a sample classification, the model's output is the predicted probability for each class. The class with higher probability is selected. For binary classification, only the positive class probability is enough since negative class probability can be inferred from the other. In some cases, only the class label is returned. It is important to note that a high α tends to underfit the model, high values of n and m tend to overfit the model, a low α may take longer to converge to the optimal solution. Nevertheless, the activation function is also important to model the data correctly. In this sense, a fine tunning of parameters is required when training a neural network model. For example, a linear decision boundary inferred by a model trained with $n = 1, m = 3$ and $\alpha = 0.1$ (Figure 2.4b) has a satisfactory predictive performance.

Figure 2.4 – Neural Network's classification and its decision boundaries.



(a) A Neural Network model representation with 3 hidden layers with (from left to right) 3, 5 and 3 neurons respectively.

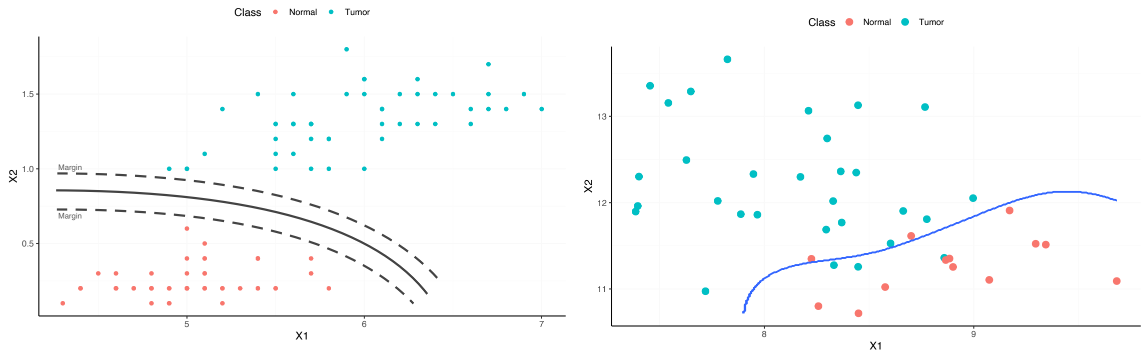
(b) Neural Network's decision boundary for binary classification.

- *Support Vector Machine* (SVM) creates a hyper-plane or a set of hyper-planes that is farthest from the nearest training samples. Thus, the hyper-plane is applied in classification problems by creating a functional margin based on the nearest samples of any class (CORTES; VAPNIK, 1995). The training of a SVM model is posed as an optimization problem in which it tries to maximize the hyper-plane distance from training samples (objective function) restricted to avoid training samples between the hyper-plane's margins. Due to the problem's restrictions, SVM is very sensitive to its parameters. The regularization parameter C inversely defines the hyper-plane's margin size, *i.e.,* a larger C defines a smaller margin size in favor of classification accuracy. 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. According to Cover (1965), a feature space is more likely to be linearly separable. Radial kernels takes a parameter σ which defines the kernel's radius. With a smaller radius, the model is restricted to a small feature space and may underfit. On the other hand, a kernel with a bigger radius may overfit the model. Note the represented margins in Figure 2.5a in contrast with a less separable data in Figure 2.5b. Both examples apply a radial kernel. The latter, however, has very strict margins due to samples proximity to the hyper-plane (decision boundary).

Figure 2.5 – Support Vector Machine's hyper-plane representation and its decision boundaries.

(a) Example of a hyper-plane learned from a SVM model with a radial kernel.

(b) Support Vector Machine's decision boundary with a radial kernel for binary classification.



2.2.2 Feature Selection

Machine Learning models' performance can be impacted by several factors. However, with the increase in the volume of data generated by recent high throughput technologies, such as microarray chips for gene expression measurement, studies have focused on solving the curse of dimensionality. High-dimensional problems, *i.e.,* when data have a high number of features, can impact negatively on models' performance. For example, irrelevant features may skew the classification of a new data point in KNN models. Redundant features may cause over-fitting, *i.e.,* when the algorithm perfectly infers the presented data points but fails to generalize the hypotheses. The curse of dimensionality states that the number of possible examples increases exponentially with a new feature.

Several feature selection (FS) methods have been proposed to reduce dimensionality. Feature selection methods - base selectors in this work - are classified into two categories: rankers and subsetters. A feature selection ranker outputs all features ordered by a score. On the other hand, subsetters output a subset of the original features set. In this study, we use rankers due to its output completeness and scoring which will be used to aggregate all rankings. Among these two categories, feature selection methods are also classified between filters, wrappers, and embedded methods. Filters estimate - with statistical methods - the features importance on separating the different classes. Usually, filters output a score for each feature in order to create a feature ranking. In contrast, wrappers use classification algorithms to select the best set of features. Despite their capability to yield better results, wrappers are computationally more expensive than filters. On a similar approach, embedded methods use features' scores from classifiers, such as decision trees or support vector machines, to identify relevant features. In this case, feature selection is embedded in the classifier - information gain in decision trees, for example. Feature selection algorithms are also divided into univariate and multivariate approaches. The former only considers one feature at a time for selection. The latter is able to use values from other features to select just one.

In order to compare the final results with meta-analysis output, a full scored ranking of features is required. Thus, the choice of base selectors must consider the form of its outputs. Rankers feature selection methods assign each feature a score. The final result is an ordered rank of features. Among rankers - embedded, wrappers and filters - feature selection methods, filters present better trade-off between predictive and computational performance. Given more than 7.000 features, embedded and wrappers are not able to perform in a timely manner. Therefore, for the purpose of this work, only ranker-filter feature selection methods were selected. The most frequent ranker-filter methods presented in the literature and applied on this work are the following:

- *Information Gain* (KENT, 1983) measures the correlation between two random variables A and B given a prior entropy $H(B)$ (Equation 2.1) and a conditional entropy $H(A|B)$ (Equation 2.2). Defined by Shannon (1948), entropy quantifies the informative value of a variable, where $p(a)$ is the probability that $A = a$. The conditional entropy quantifies the informative value of a variable A given B, where $p(a, b)$ is the probability of $A = a$ and $B = b$. Therefore, information gain $I(A, B)$ (Equation 2.3) estimates the contribution of B in the overall informative value of A. In our case, $I(Y, G_k)$ defines the target values' information gain given a set of gene expressions.

$$H(A) = -\sum_{a \in A} p(a) \log_2 p(a) \tag{2.1}$$

$$H(A|B) = -\sum_{a \in A, b \in B} p(a,b) \log_2 \frac{p(a,b)}{p(b)} \tag{2.2}$$

$$I(A,B) = H(A) - H(A|B) \tag{2.3}$$

- ϕ_c (CRAMER, 1999) measures the correlation between two random variables A and B based on the χ^2 statistical test. Pearson (1900) defined the statistical hypothesis test, χ^2 (Equation 2.4), which assumes a normal distribution, where f_{ij} is the observed frequency of (A_i, B_j), f_i and f_j are the observed frequencies of A_i and B_j respectively and $e_{ij} = \frac{f_i f_j}{n}$ is the expected frequency (equally distributed). In this sense, Cramér's V coefficient (ϕ_c) represents the squared mean correlation between A and B (Equation 2.5).

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \tag{2.4}$$

$$\phi_c = \sqrt{\frac{\chi^2}{n \cdot min(r-1, s-1)}} \tag{2.5}$$

- *Symmetrical Uncertainty* (THEIL, 1957) is a weighted average between the entropy (Equation 2.1) of two random variables A and B and represents the prediction percentage of A given B, where $I(A,B)$ is the information gain (Equation 2.3), defined as follows:

$$U(A,B) = 2 \cdot \frac{I(A,B)}{H(A) + H(B)} \tag{2.6}$$

- *Minimum Redundancy Maximum Relevance (mRMR)*, defined by Hanchuan Peng, Fuhui Long and Ding (2005), best scores variables with higher correlation to the target value and lower correlation among other variables. For such, it takes the information gain between the response variable A given $B \in V$, where V is the set of variables, and subtract the pairwise information gain between other variables and B. $J(B)$ can be expressed as the relevance term minus the redundancy term (Equation 2.7).

$$J(B) = I(A,B) - \frac{1}{\mid V \mid} \sum_{C \in V - \{B\}} I(B,C) \tag{2.7}$$

- *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 (Equation 2.8).

$$S^2(A) = \frac{1}{\mid A \mid -1} \cdot \sum_{a \in A} (a - \mu_A)^2 \tag{2.8}$$

2.2.3 Ensemble Learning Feature Selection

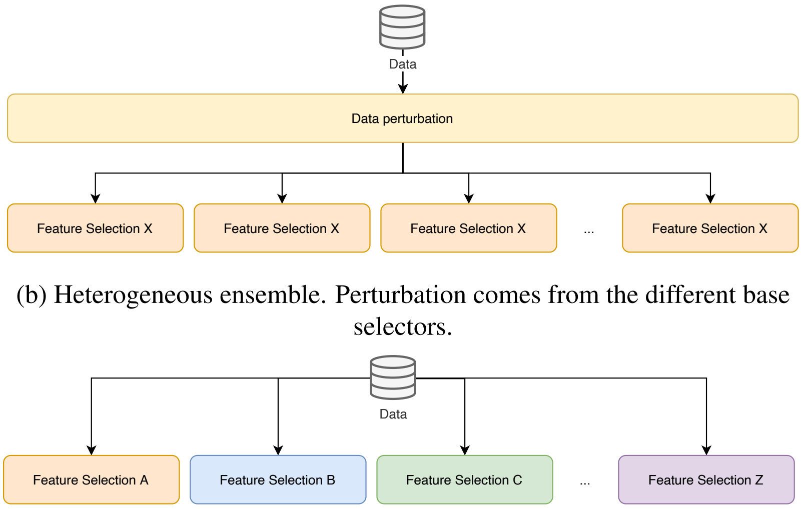
Each machine learning algorithm has its own limitations. For example, J48 decision boundaries are always parallel to features' space, NNET and SVM are very sensitive to different parameter values. Therefore, an ensemble attempts to reduce the individual algorithms' limitations by combining its models. Ensembles apply several algorithms so that one supplements the weakness of the other. They are a hybrid solution to find an optimal hypothesis.

In recent literature, ensemble feature selection (EFS) has been studied as a potential, more robust, solution to high-dimensional problems (PES, 2019; BOLÓN-CANEDO; ALONSO-BETANZOS, 2019; ALI et al., 2018; Ben Brahim; LIMAM, 2017). In this sense, ensemble feature selection on high-dimensional data can be used to increase stability on knowledge discovery. The idea is to use various base selectors and aggregate its results aiming to obtain a more stable feature subset. In order to design an efficient ensemble, Bolón-Canedo and Alonso-Betanzos (2019) mentioned five main decisions that must be taken: (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.

Ensembles can be classified into two main categories: homogeneous (Figure 2.6a) and heterogeneous (Figure 2.6b). The former uses different feature selection algorithms as base selectors. In this way, the ensemble perturbation comes from function-based approach. The latter is built with several instances of the same base selector. In contrast, the perturbation comes from data-based approaches such as bootstrap (EFRON; TIBSHIRANI, 1994). In machine learning, bootstrap is a method to resample data. Given a dataset with n samples, the method randomly selects n samples from the original dataset with replacement, which means the new dataset - also called *bag* - will probably have duplicated samples. The process can be repeated to generate more bags. In this way, bootstrap inserts heterogeneity between bags (data perturbation) which is essential for homogeneous ensembles.

Heterogeneous ensembles frees the user from choosing the base selector. According to Saha, Sarkar and Mitra (2009), Zhang and Jonassen (2019), heterogeneous ensembles can maintain or improve accuracy in comparison with state-of-the-art methods. On the other hand, homogeneous ensembles increase stability while maintaining accuracy when compared to other approaches (PES, 2019; PES; DESSÌ; ANGIONI, 2017; ZHANG; JONASSEN, 2019; SEIJO-PARDO et al., 2017).

Figure 2.6 – Types of ensemble exemplified by feature selection ensembles.
(a) Homogeneous ensemble in which all base selectors are the same and perturbation comes from data sampling or variation.



(b) Heterogeneous ensemble. Perturbation comes from the different base selectors.

Multiple results from ensembles can be aggregated through a vast range of algorithms. From statistical methods, such as $mean$, to social choice functions like *Borda Count* (RECAMONDE-MENDOZA; BAZZAN, 2016). The $mean$ method, for example, uses the feature selectors' scores as defined in Equation 2.9. On the other hand, methods such as *Borda Count* takes into account the position of a feature in the ranking. Therefore, with different approaches, the choice of the aggregation method can impact ensemble's final performance. In Seijo-Pardo et al. (2017), the authors experimented on 7 datasets with different statistical aggregation methods – $minimum$, $median$, $mean$, *Geometric Mean*, *Stuart* (AERTS et al., 2006) and *Robust Rank Aggregation* (KOLDE et al., 2012) - in order to study its behavior across several scenarios. The authors concluded that "[...] the choice of the aggregation method can impact the final results for microarray datasets". As presented by Seijo-Pardo et al. (2017), the best aggregation method for microarray datasets is $mean$ and it is defined as follows:

$$mean_i(s_i^1, s_i^2, ..., s_i^n) = \frac{1}{n}\sum_{j=1}^{n} s_i^j \tag{2.9}$$

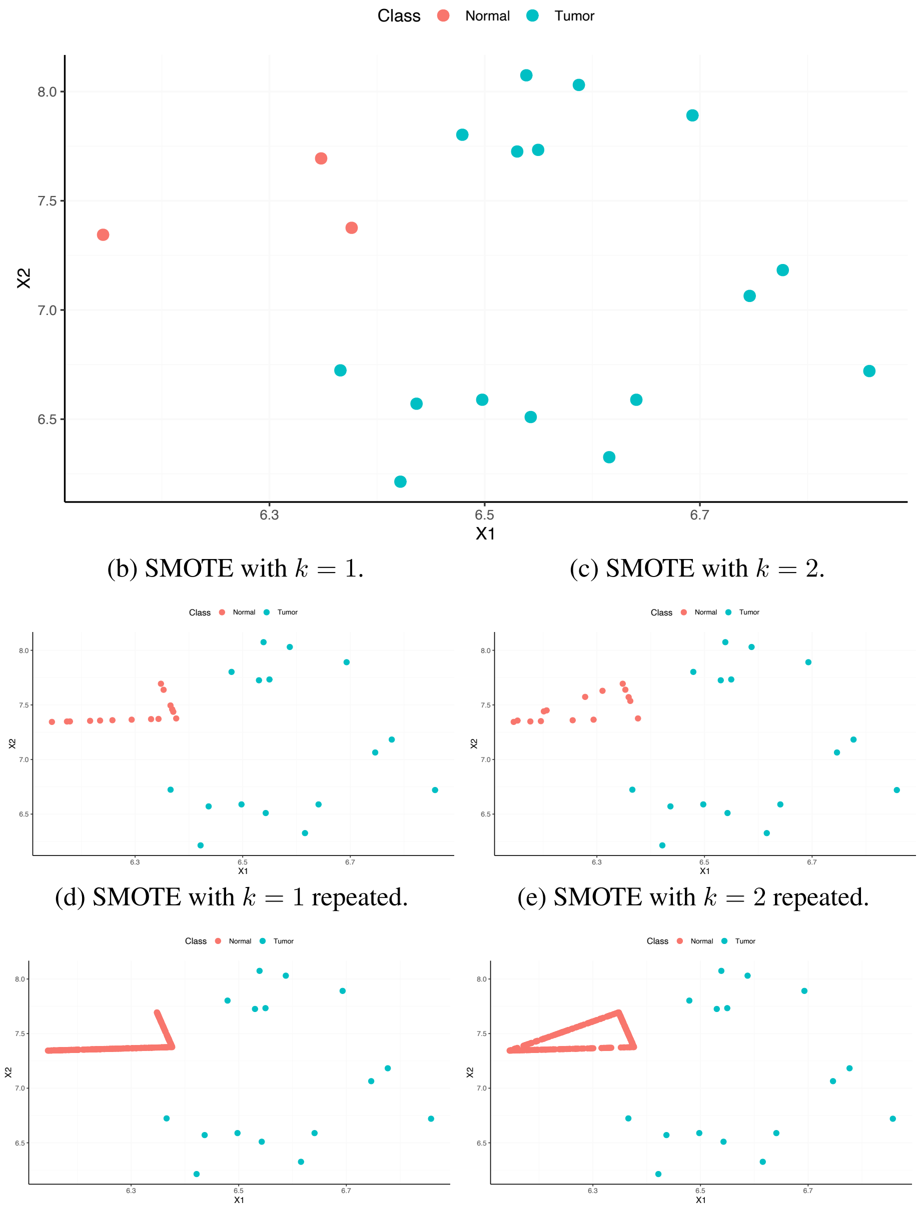
In Equation 2.9, the score of a feature i is calculated given n scores s_i^k from n base selectors rankings. The final ranking is computed by ordering features' mean score.

In order to select the best subset of features given an aggregated ranking, threshold methods use final scores to find the most relevant features. The study in Seijo-Pardo et al. (2017) also reported results for different threshold methods such as Fisher ratio, $log_2(n)$, 10%, 25% and 50%. For microarray datasets, 25% and 50% of features with higher scores presented better results in comparison with other methods. However, the authors concluded that "although satisfactory average test error results were obtained for the 50% threshold, it might not be a very suitable threshold for large dimensionality datasets". This is due to feature redundancy among relevant ones, according to Khaire and Dhanalakshmi (2019), He and Yu (2010). As mentioned by Liu, Liu and Zhang (2010), highly correlated features to its classes and uncorrelated to other features can reduce redundancy. However, according to Ali et al. (2018), in order to select the optimal subset of features, a domain expert is still required.

2.2.4 Sampling Methods

Unbalanced datasets pose a challenge in ensemble evaluation. For microarray datasets, where positive labels represent the majority of samples, few negative samples belong to stratified training and validation sets. Therefore, models' performance is wrongly interpreted. Over- and under-sampling methods try to solve the unbalance problem by leveling the number of positive and negative samples. Under-sampling randomly removes samples from the majority class. Another approach, which avoids sample removal and for that more suitable for low-sample datasets, can be found at Chawla et al. (2002). The synthetic minority over-sampling technique (SMOTE). Figure 2.7 visually represents how SMOTE works. For each sample s in the minority class - *normal*, SMOTE selects its k nearest neighbors through Euclidean distance and randomly generate synthetic samples between two existing and closest neighbors (Figures 2.7b and 2.7c). The over-sampling method tries to balance the dataset by generating new samples for the minority class. However, the minority class can become the majority one by repeating SMOTE over the

Figure 2.7 – Visual representation of how the synthetic minority over-sampling technique (SMOTE) works.



(a) Original dataset.

(b) SMOTE with $k = 1$.

(c) SMOTE with $k = 2$.

(d) SMOTE with $k = 1$ repeated.

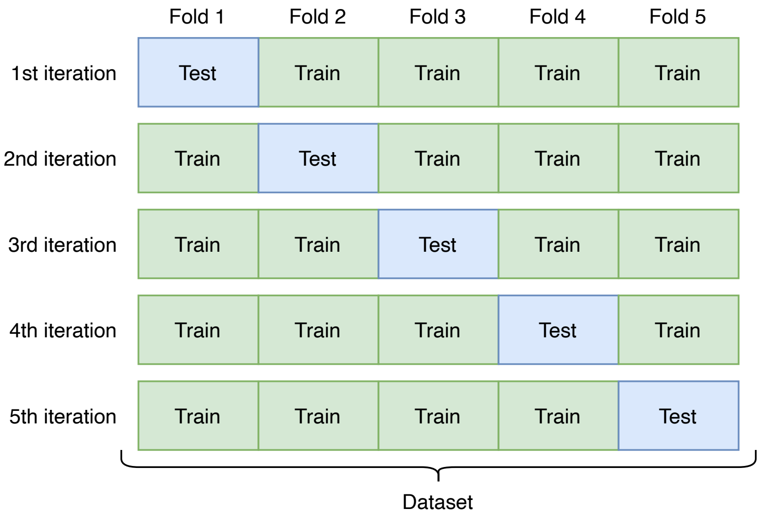
(e) SMOTE with $k = 2$ repeated.

original dataset, as shown in Figures 2.7d and 2.7e.

2.2.5 Predictive Power Evaluation

In Kuncheva and Rodríguez (2018), the authors presented an evaluation protocol that avoids data leakage on selecting features and evaluating the model on very low-sample-size data. The ideas of Kuncheva and Rodríguez (2018) are essential on ensemble design considering data partition on training and validation sets and the application of over-sampling methods to balance the datasets. The authors proposed a variation of cross-validation method which considers the feature selection step. As an evaluation method, cross-validation divides data into k *folds* each of which is an equally distributed - also known as *stratified* - and disjointed portion of the data. For every one of the k iterations, $k-1$ folds are used for training the classifier and the remaining one is used for model evaluation (Figure 2.8). A variation of k-fold cross-validation - called *leave-one-out cross-validation* (LOOCV) - leaves one sample for evaluation and the rest for training. Kuncheva and Rodríguez (2018) method introduces, prior to the training step, a feature selection using the same training portion of the k-fold cross-validation.

Figure 2.8 – 5-fold cross validation. For each iteration, a fold is selected for testing and the rest is selected for training.



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1st iteration	Test	Train	Train	Train	Train
2nd iteration	Train	Test	Train	Train	Train
3rd iteration	Train	Train	Test	Train	Train
4th iteration	Train	Train	Train	Test	Train
5th iteration	Train	Train	Train	Train	Test

In each iteration of k-fold cross-validation process, the model is evaluated by the testing set. The model's predictions generates a confusion matrix (Table 2.2). For binary classification, a confusion matrix $M_{2\times2}$ presents four main terms, defined as follows:

- *True positive* (TP) defines the number of samples labeled with and predicted as belonging to class the positive class.
- *False positive* (FP) defines the sum of instances whose prediction was positive but their true label is the negative class.
- *False negative* (FN) defines the sum of instances labeled as positive but predicted as negative.
- *True negative* (TN) defines the sum of instances labeled with and predicted as belonging to the negative class.

From the confusion matrix, we can derive the metrics for measuring a model's

Table 2.2 – Layout of a binary class confusion matrix.

Labeled Predicted	Positive	Negative
Positive	TP	FP
Negative	FN	TN

performance. Each metric will guide the choice of the best classifier according to the model's purpose. For microarray problems in which the model's predict the presence of cancer, it is important to reduce the false negative results, *i.e.,* when the model predicts negative for cancer but the sample is positive. In this way, we need to focus on maximizing recall, which is indirectly proportional to false negative errors. Recall and other metrics are defined and applied in this work as follows:

- *Precision* or positive predicted value defines the ratio of true positive predictions over all positive predictions. Thus:

$$Precision = \frac{TP}{TP + FP} \tag{2.10}$$

- *Recall*, sensitivity or true positive rate defines the ratio of true positive predictions over all samples labeled as positive. Recall value is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{2.11}$$

- *F1-Score* ($F1$) defines the harmonic mean between precision and recall. Thus:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.12}$$

2.2.6 Stability Evaluation

Besides predictive performance metrics, feature selection stability is equally important. In knowledge discovery applications, stability represents the selected features' quality and, therefore, the knowledge itself. Stability can also be defined as the robustness of a feature selection algorithm in face of perturbations in data. Usually, in high dimensional datasets, feature selection methods find more than one optimal subset - low stability - due to features' redundancy which leads to a decrease in the overall quality of the results. Recent research (PES, 2019; ZHANG; JONASSEN, 2019; ALI et al., 2018) have proposed the use of ensemble feature selection to deal with stability on high dimensional datasets. In particular, homogeneous EFS has been applied to improve stability due to its intrinsic characteristics. Homogeneous EFS explores the different visions - by data perturbation - from the same feature selection algorithm that has shown an increase stability, mainly on high dimensional datasets (ABEEL et al., 2009; PES; DESSÌ; ANGIONI, 2017).

He and Yu (2010) have shown three causes of instability. In the classic ML pipeline, a feature selection process directly precedes the model training. Thus, the lack of a validation step before training often leads to unstable results. The validation - between feature selection and model training - makes it possible to discover multiple optimal subsets of features that is possibly causing instability. This can happen when the dataset has several redundant features. Finally, specially for microarray datasets, the low-sample high-dimensional data can cause instability in feature selection due to the lack of more information for each feature.

In this work, the stability validation step calculates the Kuncheva Index (Equation 2.14) defined by Kuncheva (2007). The Kuncheva index is the average of pairwise inconsistency indexes between a set of subsets of features $\mathcal{S} = \{S_1, S_2, S_3, ..., S_N\}$. An inconsistency index, according to Kuncheva (2007), 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 2.13 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.

$$I(A,B) = \frac{rn - k^2}{k(n-k)} \tag{2.13}$$

$$\mathcal{I}(\mathcal{S}) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=1}^{N} I(S_i, S_j) \tag{2.14}$$

3 RELATED WORK

Analysis on microarray datasets can lead to erroneous conclusions. As explained by Ang et al. (2016), when dealing with microarray data we should consider several problematic factors. Technical errors on data collection are the start point. Factors such as type and quantity of reagents used, handling of data collection equipment, results discretization, data mislabeling can generate erroneous results. Furthermore, a large number of gene expression can cause over-fitting even after a feature selection, which connotes the difficulty of differentiating relevant and redundant gene expressions. According to He and Yu (2010), due to redundancy, there are many gene subsets that can explain the presence of cancer - biomarkers. Hence, the importance of feature selection design considering stability.

The authors in Khaire and Dhanalakshmi (2019) investigated the stability of feature selection methods. Stability measures indicates feature selection output's robustness given data perturbation across several runs. According to He and Yu (2010), Khaire and Dhanalakshmi (2019), highly correlated features usually lead to unstable outputs when applying state-of-the-art feature selection methods, which means the methods can identify more than one optimal feature set in high-dimensional data due to feature redundancy. In this sense, Pes (2019) evaluated stability on ensemble-based feature selection methods across several domains, including biomedical data. Pes (2019) demonstrated through extensive experimentation that homogeneous ensemble-based approaches lead to a significant gain in stability.

In an earlier study, Pes, Dessì and Angioni (2017) exploited the performance and stability of homogeneous ensembles applied to high-dimensional genomics data. In every case, the ensemble approach outperformed other methods. The authors were able to achieve a high accuracy with 3% of selected features and an stability increase with homogeneous ensembles compared to other methods. The study in Abeel et al. (2009) corroborates with Pes, Dessì and Angioni (2017) by using homogeneous ensembles on microarray datasets to increase stability. As in the former study, the authors reported better stability results for all datasets. In order to balance stability and predictive accuracy, however, Zhang and Jonassen (2019) proposed a hybrid ensemble approach which uses homogeneous and heterogeneous ensembles. The authors presented both high predictive accuracy and stability. On another study in which results corroborate with the latter, Saha, Sarkar and Mitra (2009) compared the robustness of feature selection methods either individually or as ensembles. Using Spearman correlation coefficient, ensembles yielded higher robustness than individual methods for all microarray datasets.

In Yu and Liu (2003), the authors proposed a fast correlation-based filter which can identify relevant and redundant features efficiently. The method achieved an average accuracy of 95.06% with 14 features selected among 650 total using C4.5. Similarly, Ali et al. (2018) presented an univariate ensemble-based feature selection method which can identify relevant features among redundant ones. The author reported higher predictive accuracy than state-of-the-art methods. The authors in Ben Brahim and Limam (2013) explored the reliability assessment based on an aggregation technique in which classification performance of features subsets determines features' confidence to assess reliability. According to the study, the ensemble-based method using KNN achieved 85.5% of F-Measure and outperformed other methods on breast cancer dataset.

The study in Das, Das and Ghosh (2017) proposed an ensemble of bi-objective genetic algorithm in which a stochastic search (through the genetic algorithm) is performed on a feature selection algorithm for subsets of data. The proposed method outperformed other methods for high-dimensional datasets. In Seijo-Pardo et al. (2017), the authors used 7 datasets with different feature selection methods in order to improve training time and increase accuracy. The best results presented for microarray datasets used homogeneous ensemble with SVM-RFE with mean as aggregation method. Among conclusions, the authors state that "[...] an ensemble approach would seem to be the most reliable approach to feature selection".

As consequence, the studies using microarray datasets evaluate performance either using single feature selection methods or ensembles. The authors in Bolón-Canedo et al. (2014) reviewed the performance of state-of-the-art feature selection methods across several domains of microarray - breast cancer, prostate, brain, colon, ovarian. With mMRM method and the top 10 genes selected, the authors achieved 100% recall for breast cancer dataset. However, the authors concluded that, in general, performance depends on the feature selection method, on the classifier and, mainly, on the problem domain. On another approach, Bolón-Canedo, Sánchez-Maroño and Alonso-Betanzos (2012) designed an ensemble in which for each base selector's output a classifier was trained. The final predictions were combined by majority voting. 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, Liu, Liu and Zhang (2010) 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 increase in stability compared to other methods.

On another approach, presented by Sharifi et al. (2018), a decision tree combined with meta-analysis were able to identify four biological markers with 83% accuracy. On the other hand, Alejandro et al. (2018) cross-validated the results of genes selected from ensemble feature selection algorithms with meta-analysis results and data from literature. The authors were able to identify 100 genes that could explain 29 types of cancer. In every case, Alejandro et al. (2018) reported accuracy higher than 90%.

Table 3.1 summarizes the main results found in the literature. Since the reported performance metric varies among studies, we only included works that have reported either AUC score or accuracy for the performance of the EFS approach. We note that works that explore homogeneous ensemble feature selection use sampling methods from the same dataset to generate data perturbation. In contrast, as we will explain in Chapter 4, in our approach data perturbation comes from different datasets and the sampling method bootstrap.

Table 3.1 – Main studies considering AUC score, accuracy (ACC) and the Kuncheva index (KI).

Reference	Method	AUC	ACC	KI
Zhang and Jonassen (2019)	Hybrid EFS	99%	-	-
Alejandro et al. (2018)	Hybrid EFS	-	92%	-
Ali et al. (2018)	Heterogeneous EFS	-	73%	-
Das, Das and Ghosh (2017)	Homogeneous EFS	-	92%	-
Pes, Dessì and Angioni (2017)	Homogeneous EFS	90%	-	0.96
Ben Brahim and Limam (2013)	Heterogeneous EFS	84%	-	-
Liu, Liu and Zhang (2010)	Hybrid EFS	-	97%	-
Abeel et al. (2009)	Homogeneous EFS	96%	-	0.72
Saha, Sarkar and Mitra (2009)	Homogeneous EFS	-	96%	-
Yu and Liu (2003)	Heterogeneous EFS	-	89%	-

4 METHODOLOGY

Based on results found in the literature, we have designed a pipeline to evaluate the performance of both ensemble feature selection and meta-analysis methods. Data collection (Section 4.1) was executed manually to safely insure the data integrity and source. Data pre-processing (Section 4.2) was a simple process to gather and to standardize all common genes between the datasets. The EFS and meta-analysis design and evaluation steps (Sections 4.3 and 4.4) were implemented and executed in R Language (R Core Team, 2020) using several R packages[1]. The next sections will detail each step of the evaluation process.

4.1 Data collection

There is a large volume of microarray datasets publicly available in databases such as Gene Expression Omnibus (GEO) (EDGAR; DOMRACHEV; LASH, 2002; BARRETT et al., 2012). In this work, we searched the GEO repository for datasets related to breast cancer, selecting those that presented both control and tumor samples and a good number of samples per group. We also restricted our search for Affymetrix or Agilent microarray platforms, for which pre-processing protocols are better established. All breast cancer microarrays were included, except those that included cancer patients receiving some kind of medical treatment.

In our search, 13 datasets were considered eligible for our study. 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 4.1). The remaining datasets were assign for evaluation defined as $E = \{E_1, E_2, E_3, ..., E_6\}$ (Table 4.2). In this case, let $D \in T \cup E$ be a set of samples, each one as a tuple $(x, y)_i = (x^i_1, x^i_2, x^i_3, ..., x^i_m, y_i)$ where x^i_k is the k^{th} gene expression for sample i and $y_i \in \{normal, tumor\}$ is the target value. Also, let $G_k = [x^1_k, x^2_k, x^3_k, ..., x^n_k]$ be the k^{th} gene's expression values and $Y = [y_1, y_2, y_3, ..., y_n]$ the target values for each sample.

We note that there is a large heterogeneity among different datasets in terms of patients' tumor characteristics. For instance, some datasets (*e.g.,* GSE38959, GSE53752) include patients with a triple negative tumor, which are those tested negative for the three most common type of receptors in cancer, *i.e.,* hormone epidermal growth factor receptor

[1] List of packages: <https://github.com/btrevizan/biomarker_id/blob/master/requirements.R>

Table 4.1 – Datasets used for training.

			Number of samples			
	Dataset	Features	Tumor	Normal	Total	Tumor/Total Ratio
T_1	GSE38959	19.750	30	13	43	0.70
T_2	GSE42568	20.471	98	17	115	0.85
T_3	GSE45827	20.917	122	36	158	0.77
T_4	GSE53752	18.318	46	21	67	0.69
T_5	GSE62944	23.368	1119	113	1232	0.91
T_6	GSE70947	32.577	148	148	296	0.50
T_7	GSE7904	20.896	42	18	60	0.70

Table 4.2 – Datasets used for evaluating the trained model.

			Number of samples			
	Dataset	Features	Tumor	Normal	Total	Tumor/Total Ratio
E_1	GSE10797	12.182	27	5	32	0.84
E_2	GSE22820	30.484	74	10	84	0.88
E_3	GSE26304	31.013	109	6	115	0.95
E_4	GSE57297	36.337	25	7	32	0.78
E_5	GSE61304	18.663	57	4	61	0.93
E_6	GSE71053	20.896	6	12	18	0.33

2 (HER-2), estrogen receptors (ER), and progesterone receptors (PR), and can be more aggressive. Others include patients with both ER positive and negative (*e.g.,* GSE42568) or with various tumor subtypes (*e.g.,* GSE45827). Also, dataset GSE62944 contains combined data for several types of cancer screened by The Cancer Genome Atlas (TCGA)[2], from which we analyzed only samples related to breast cancer. The datasets heterogeneity certainly poses challenges to data analysis, nonetheless, as we are interested in identifying general breast cancer candidate biomarkers rather than subtype-specific ones, we deem suitable to include datasets with different tumor subtypes.

4.2 Data Pre-processing

Microarray data pre-processing is important to adjust the effects on data from technical and experimental variations and allow samples from the same study to be comparable to each other. There are well established bioinformatics protocols to perform gene expression data pre-processing and prepare them for statistical or computational analyses. Here, we adopted standard pre-processing pipelines for Affymetrix and Agilent platforms, as reported in Bueno and Recamonde-Mendoza (2020)[3]. In summary, for Affymetrix

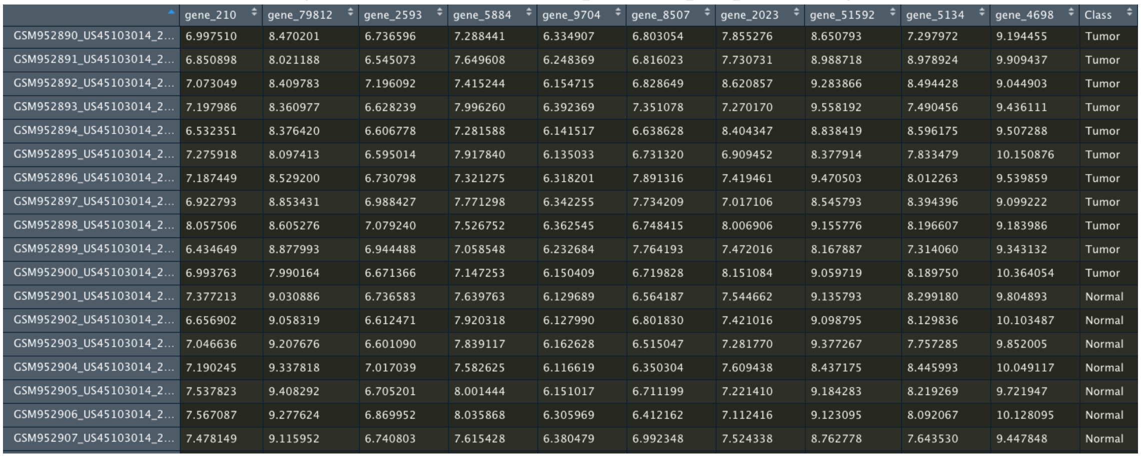
[2] <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>

[3] We thank Rodrigo Haas Bueno for his support with microarray data pre-processing.

studies, data quality analysis was performed with the *arrayQualityMetrics* package for R, followed by normalization with the Robust Multi-array Average (RMA) algorithm (IRIZARRY et al., 2003) through the *oligo* R package (CARVALHO; IRIZARRY, 2010). To model and to correct possible batch efects (source of variation mainly due to technical heterogeneity), the surrogate variable analysis (SVA) correction algorithm (LEEK; STOREY, 2007) was applied to data. Agilent studies were first assessed for data quality using the evaluation of MA-plots, background intensities boxplots, and PCA, as recommended by Limma user's guide. Pre-processing of expression values was performed using the limma R package (RITCHIE et al., 2015), applying background correction and between-array normalization using the quantile method. Probe to gene symbol annotation was performed using tables provided by the manufacturer or by the Bioconductor repository.

In microarray datasets, a gene may be identified by many ways depending on the reference identifier adopted. In order to avoid redundancy on gene identification for later comparison of gene sets, its symbols were mapped to its intrinsic *Entrez ID* - a unique gene identifier - using the R package *biomaRt* (DURINCK et al., 2005; DURINCK et al., 2009). Furthermore, the resulting genes not belonging to any other dataset were removed so as to create a unique set of common genes resulting in 7.897 genes in all datasets. An example of a dataset is presented in Figure 4.1 in which the rows represent the patients, the columns represent the genes, the values are the gene expressions and the last column is the target value (class). To present readable results in Figures and Tables, we represent the genes using their *Official Symbol* provided by HUGO Gene Nomenclature Committee (HGNC)[4].

Figure 4.1 – Dataset example after pre-processing.



	gene_210	gene_79812	gene_2593	gene_5884	gene_9704	gene_8507	gene_2023	gene_51592	gene_5134	gene_4698	Class
GSM952890_US45103014_2...	6.997510	8.470201	6.736596	7.288441	6.334907	6.803054	7.855276	8.650793	7.297972	9.194455	Tumor
GSM952891_US45103014_2...	6.850898	8.021188	6.545073	7.649608	6.248369	6.816023	7.730731	8.988718	8.978924	9.909437	Tumor
GSM952892_US45103014_2...	7.073049	8.409783	7.196092	7.415244	6.154715	6.828649	8.620857	9.283866	8.494428	9.044903	Tumor
GSM952893_US45103014_2...	7.197986	8.360977	6.628239	7.996260	6.392369	7.351078	7.270170	9.558192	7.490456	9.436111	Tumor
GSM952894_US45103014_2...	6.532351	8.376420	6.606778	7.281588	6.141517	6.638628	8.404347	8.838419	8.596175	9.507288	Tumor
GSM952895_US45103014_2...	7.275918	8.097413	6.595014	7.917840	6.135033	6.731320	6.909452	8.377914	7.833479	10.150876	Tumor
GSM952896_US45103014_2...	7.187449	8.529200	6.730798	7.321275	6.318201	7.891316	7.419461	9.470503	8.012263	9.539859	Tumor
GSM952897_US45103014_2...	6.922793	8.853431	6.988427	7.771298	6.342255	7.734209	7.017106	8.545793	8.394396	9.099222	Tumor
GSM952898_US45103014_2...	8.057506	8.605276	7.079240	7.526752	6.362545	6.748415	8.006906	9.155776	8.196607	9.183986	Tumor
GSM952899_US45103014_2...	6.434649	8.877993	6.944488	7.058548	6.232684	7.764193	7.472016	8.167887	7.314060	9.343132	Tumor
GSM952900_US45103014_2...	6.993763	7.990164	6.671366	7.147253	6.150409	6.719828	8.151084	9.059719	8.189750	10.364054	Tumor
GSM952901_US45103014_2...	7.377213	9.030886	6.736583	7.639763	6.129689	6.564187	7.544662	9.135793	8.299180	9.804893	Normal
GSM952902_US45103014_2...	6.656902	9.058319	6.612471	7.920318	6.127990	6.801830	7.421016	9.098795	8.129836	10.103487	Normal
GSM952903_US45103014_2...	7.046636	9.207676	6.601090	7.839117	6.162628	6.515047	7.281770	9.377267	7.757285	9.852005	Normal
GSM952904_US45103014_2...	7.190245	9.337818	7.017039	7.582625	6.116619	6.350304	7.609438	8.437175	8.445993	10.049117	Normal
GSM952905_US45103014_2...	7.537823	9.408292	6.705201	8.001444	6.151017	6.711199	7.221410	9.184283	8.219269	9.721947	Normal
GSM952906_US45103014_2...	7.567087	9.277624	6.869952	8.035868	6.305969	6.412162	7.112416	9.123095	8.092067	10.128095	Normal
GSM952907_US45103014_2...	7.478149	9.115952	6.740803	7.615428	6.380479	6.992348	7.524338	8.762778	7.643530	9.447848	Normal

[4] HUGO Gene Nomenclature Committee (HGNC): <https://www.genenames.org>

4.3 Ensemble Feature Selection Design

For each training dataset $T_i \in T$, we create an ensemble whose design is presented in Figure 4.2. As every dataset has a very limited number of samples, to generate stratified folds for cross-validation and to avoid using training samples for testing, we split the data into two - 60% for training and 40% for validation - stratified subsets. The training subset is resampled to perturb data for N bags. N is also the number of base selectors, one for each bag. We apply sampling through SMOTE to balance the class distribution in each bag. Otherwise, bootstrap wouldn't have the same effect for data perturbation once the samples in the minority class would be very similar due to over sampling and it wouldn't preserve the class distribution. According to Schubach et al. (2017), applying a sampling method in this step of the ensemble increases performance. Each base selector outputs a full ranking of scored genes. The set of rankings is aggregated by arithmetic mean defined in Equation 2.9. An ensemble feature selection outputs three objects: (i) local stability by Kuncheva Index (Equation 2.14) for the set of rankings; (ii) final aggregated ranking, also called as local ranking L_i; and (iii) local performance metrics (defined in Section 2.2.5) for the top K genes using 5-fold cross-validation, where K is the ranking threshold.

However, an ensemble feature selection output only measures performance based on data presented for a classifier, *i.e.,* T_i. Our interest lies in the stability and performance across several studies. To evaluate the ensembles' global stability, we applied the Kuncheva Index to the local rankings L_i for $i = \{1, 2, 3..., 7\}$ generated (Figure 4.3). The set of local rankings L is aggregated into a global ranking by arithmetic mean (Equation 2.9). We train a classifier using the global rank's top K genes and the datasets $T_i \in T$ as training data. To asses the model, we calculate the performance metrics presented in Section 2.2.5 using the test datasets $E_i \in E$ (Section 4.1) never applied for training to avoid data leakage. Finally, we have the global stability, global ranking and global performance metrics to compare the EFS model to other methods.

4.4 Meta-analysis

As one of ours goals in this work is to compare EFS with state-of-the-art meta-analysis, Figure 4.4 shows the same evaluation pipeline for EFS models. However, instead of the local ranking for a dataset T_i, meta-analysis calculates the effect size for each gene through differential expression analysis. The global ranking is generated by the

Figure 4.2 – Ensemble Feature Selection pipeline.

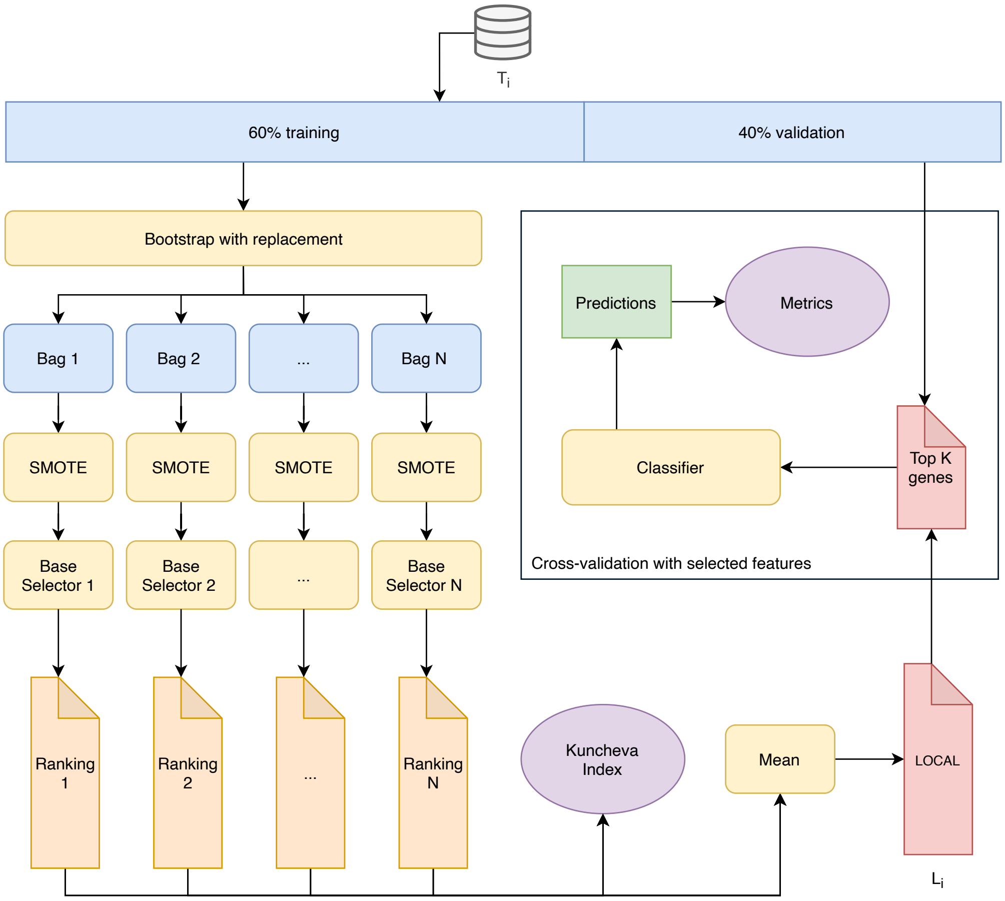
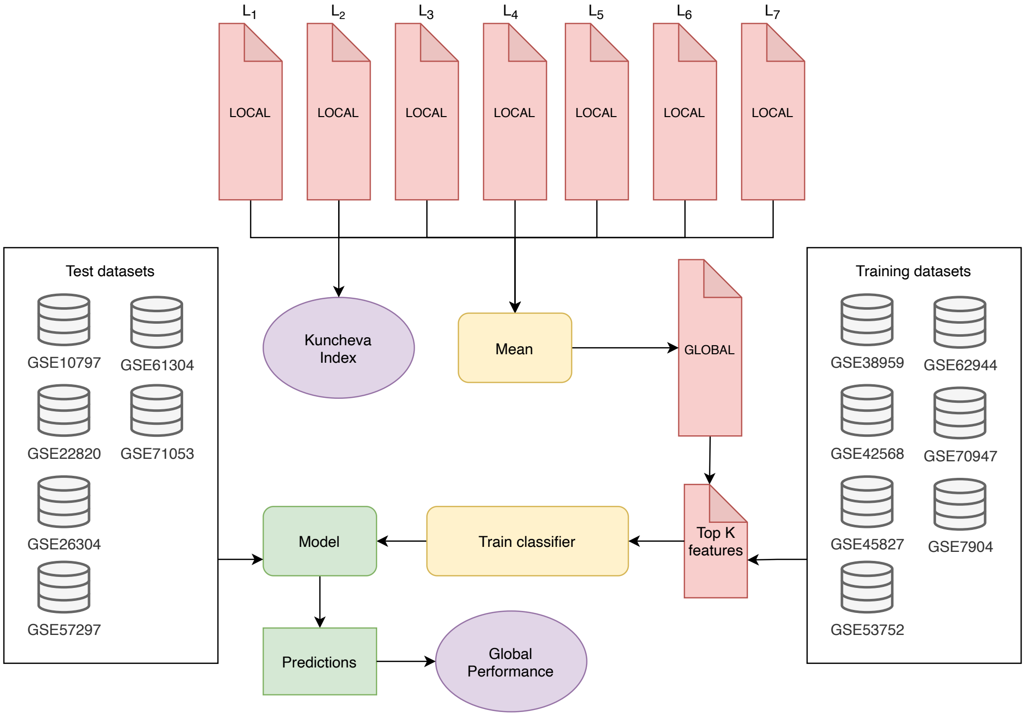
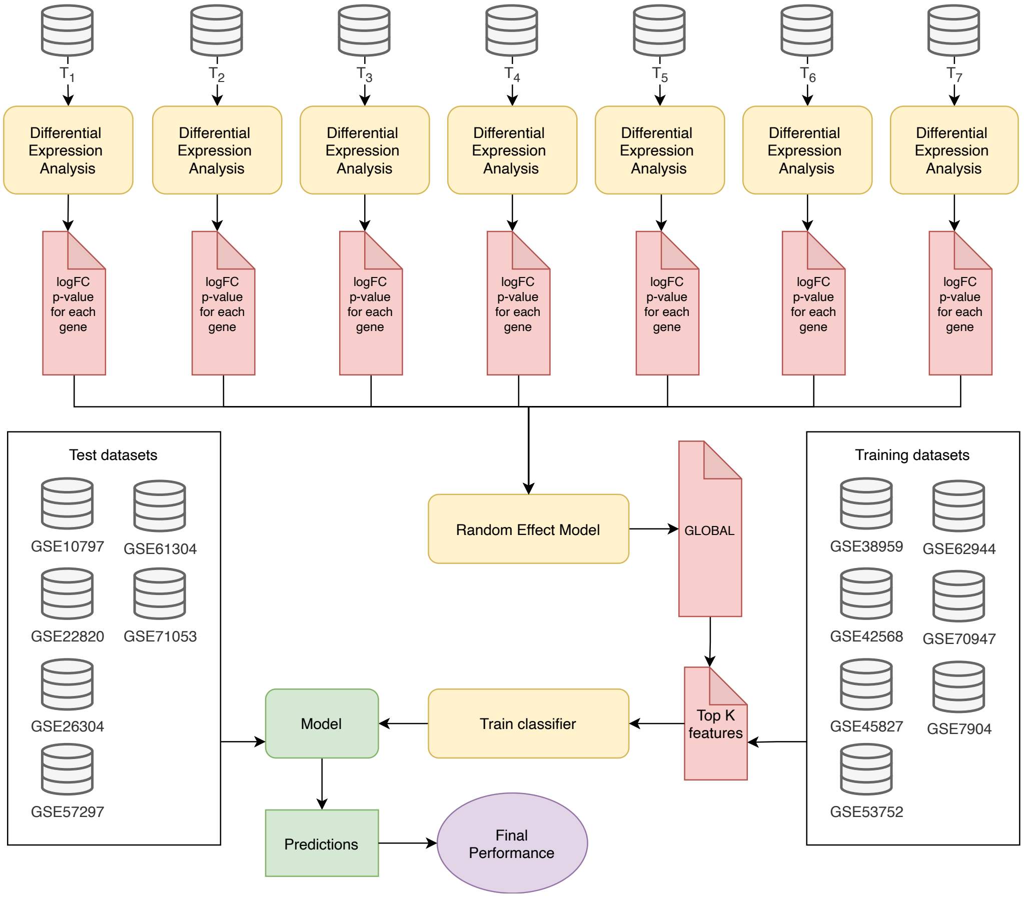


Figure 4.3 – Ensemble Feature Selection evaluation pipeline.



meta-analysis method called Random Effect Model (REM) defined in Section 4.4. 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.

Figure 4.4 – Meta-analysis evaluation pipeline.



4.5 Experimental Setup

As discussed, there are many parameters to take into account. For every experiment, however, we fix the number of folds in k-fold cross-validation to 5, the training set size to 60% of the data and the aggregation method to *mean*. The classifiers' specific parameters were automatically optimized by the *train* function implemented in the *caret* package in R (KUHN, 2008). For this work purpose, we experiment with every remaining parameter combination to have a broad view of its impact in stability and performance. Sampling methods and threshold, for example, could impact performance and stability. On the other hand, we expect that the number of bags impacts stability, but not performance. And, the main object of study, feature selection algorithms could have a decisive impact on stability, while classifiers could have on performance. Therefore, we experiment every combination of the following parameters and its possible values:

- **Number of top genes selected**: 5, 10, 15, 20, 25, 30, 50 75, 100, 150, 200, 250, 500
- **Number of bags**: 5, 10 ,25, 50, 100
- **Sampling method**: SMOTE, down sampling, no sampling
- **Classifier**: SVM, J48, KNN, NNET
- **Base selector**: information gain, *chi* squared (ϕ_c), symmetrical uncertainty, mRMR, variance

5 RESULTS

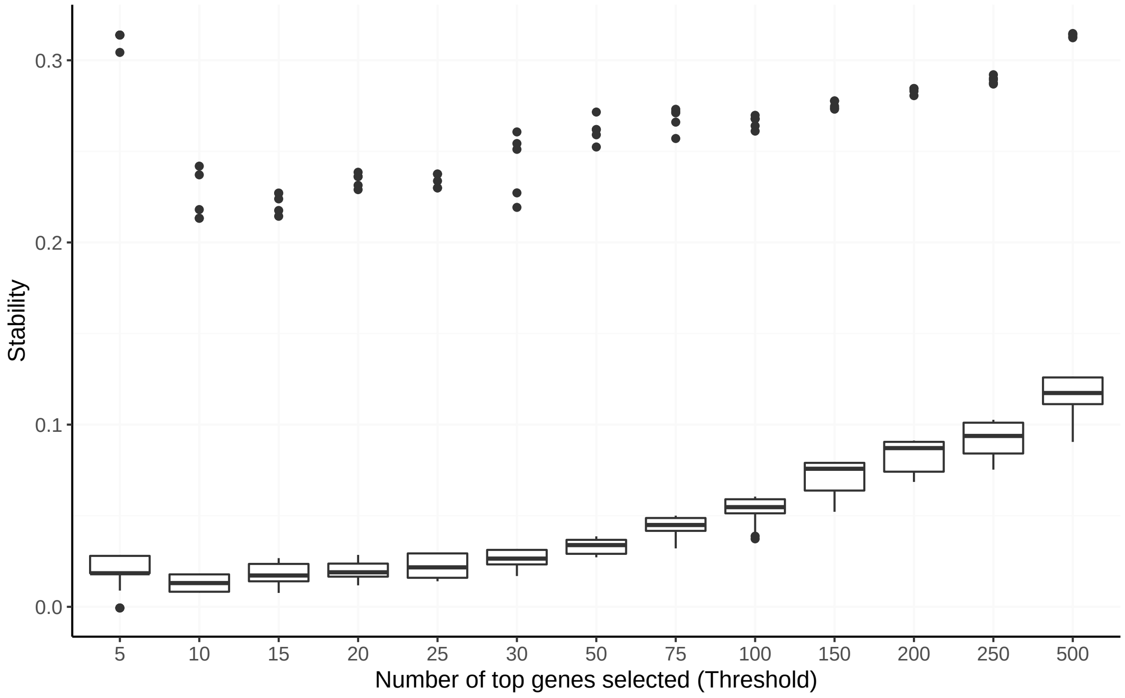
To ease our analyses, we divide this chapter into sections dedicated to only one object of investigation. Furthermore, we adopted a top-down approach in which we begin the analysis with an overview of the results and we end it with the most granular level. The boxplots present a series of results considering all parameter values. When we specify a parameter, the others are included with all its possible values. Section 5.1 presents the stability results to identify the set of parameters that increases stability. Section 5.2 presents the classifiers' predictive performance leading to Section 5.3 to compared the presented results with meta-analysis. Finally, Section 5.4 analyze the selected genes' biological functions.

5.1 Stability analysis

High stability for biomarker identification in ensemble feature selection methods means the most informative genes for one study are also the most informative for another, *i.e.,* it reflects the quality and robustness of the selected genes across the studies. The ideal scenario is a stability close to 1, indicating that almost all the same genes were selected by feature selection methods across different studies. The bigger the subset of selected genes, the higher the stability. Figure 5.1 shows the expected behavior for global stabilities, *i.e.,* stability increases with the number of selected genes. Note that global stabilities, in general, are very low, which indicates that genes among each subset evaluated (top X genes selected) show a considerable level of divergence. These differences observed for distinct studies may be due to the molecular characteristics embedded in dataset used for feature selection (FS), as well as to distinct FS methods and classifiers adopted in the methodology. On the other hand, we can see a slight increase in stability for the top 5 genes. Therefore, local ensembles are most agreeing in ranking top positions.

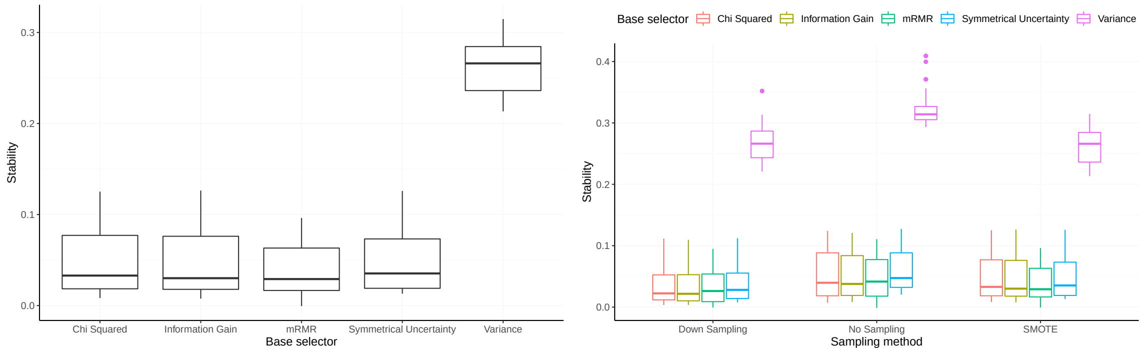
As seen in Figure 5.1, outliers outperform - in terms of stability - for every threshold. Breaking by base selectors, however, it is clear that *variance* presents an overall higher stability in comparison to other base selectors (Figure 5.2a), which explains the mentioned outliers. Figure 5.2b corroborates to establish variance as a robust method for biomarker identification in microarray datasets and shows the impact of sampling methods over stability. As the number of bags (Figure 5.3), sampling methods have insignificant impact in the overall global stability. Therefore, stability is mainly impacted by the se-

Figure 5.1 – Stability for different number of genes selected from the global ranking.



lected genes' subset size and the base selector bias. Since we want to selected a minimum set of informative genes - the biomarkers - high stability lies in the choice of the base selector.

Figure 5.2 – Variance outperforms other base selector methods independently of other parameters.

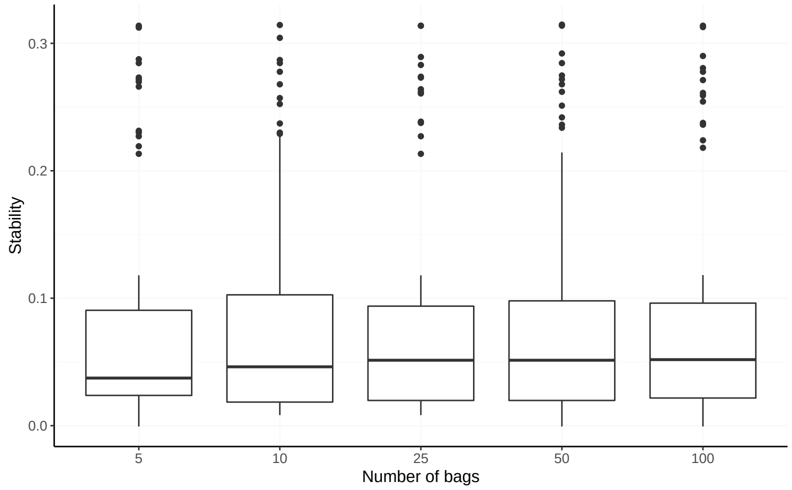


(a) Stability by base selector.

(b) Stability by sampling method break by base selector.

The global stabilities presented is at the top in the range of 1 to -1 from Kuncheva index. According to Kuncheva (2007), stability is considered high when above 0.5. However, Figure 5.4 shows higher stabilities results when compared locally within different datasets. In GEO, different datasets are collected from distinct studies and show varied patients and tumor characteristics, which may explain our finding of local stabilities higher than global stabilities. In this work, we will not address the biological implications

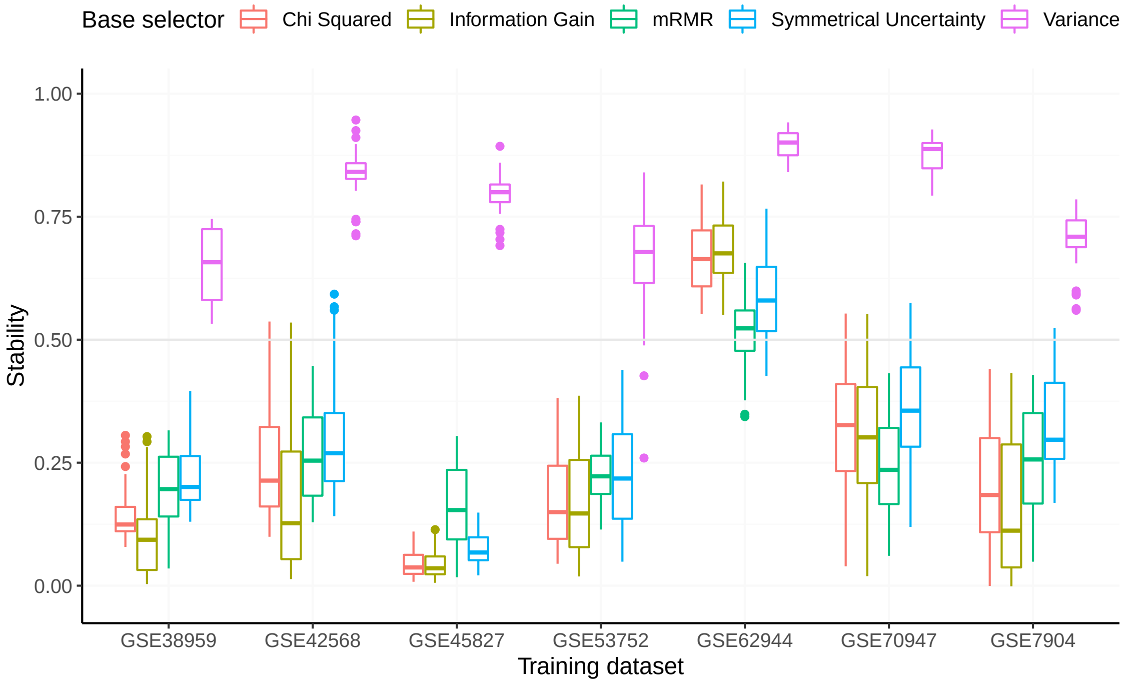
Figure 5.3 – Stability for different number of bags regardless of base selector.



of different sources of data, but this could be interesting to investigate in future analyses.

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 the 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.

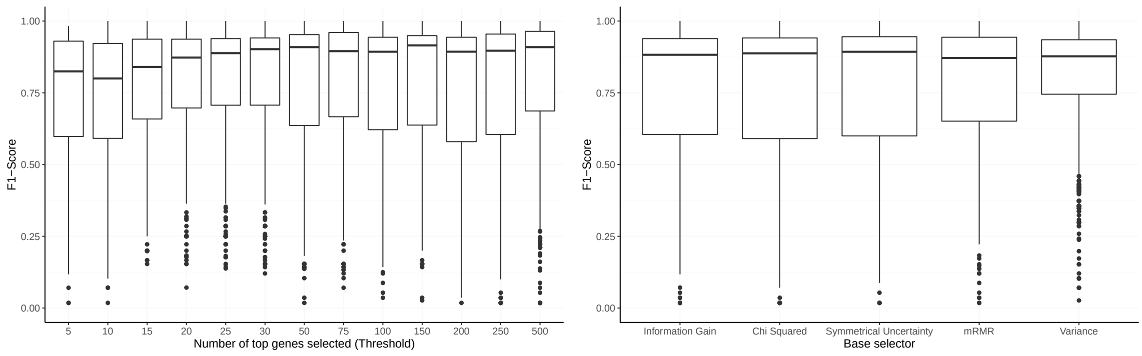
Figure 5.4 – Local stability for training datasets break by base selector.



5.2 Predictive performance analysis

Classifiers predictive performance can also guide the identification of biomarkers. We can hypothesize that the most informative genes will yield a better performance in classification. However, several other parameters could also impact performance. As Figure 5.5 shows, the number of genes selected and base selectors don't impact performance significantly. We can see a slight increase on performance until 50 genes are selected. However, in general, the top 5 genes already presented high F1-Score values (Figure 5.5a). Among the base selectors, variance presented results with less variability, which corroborates with the stability results. Therefore, we will focus our next analysis on results presented by variance as base selector and the top 20 genes as threshold.

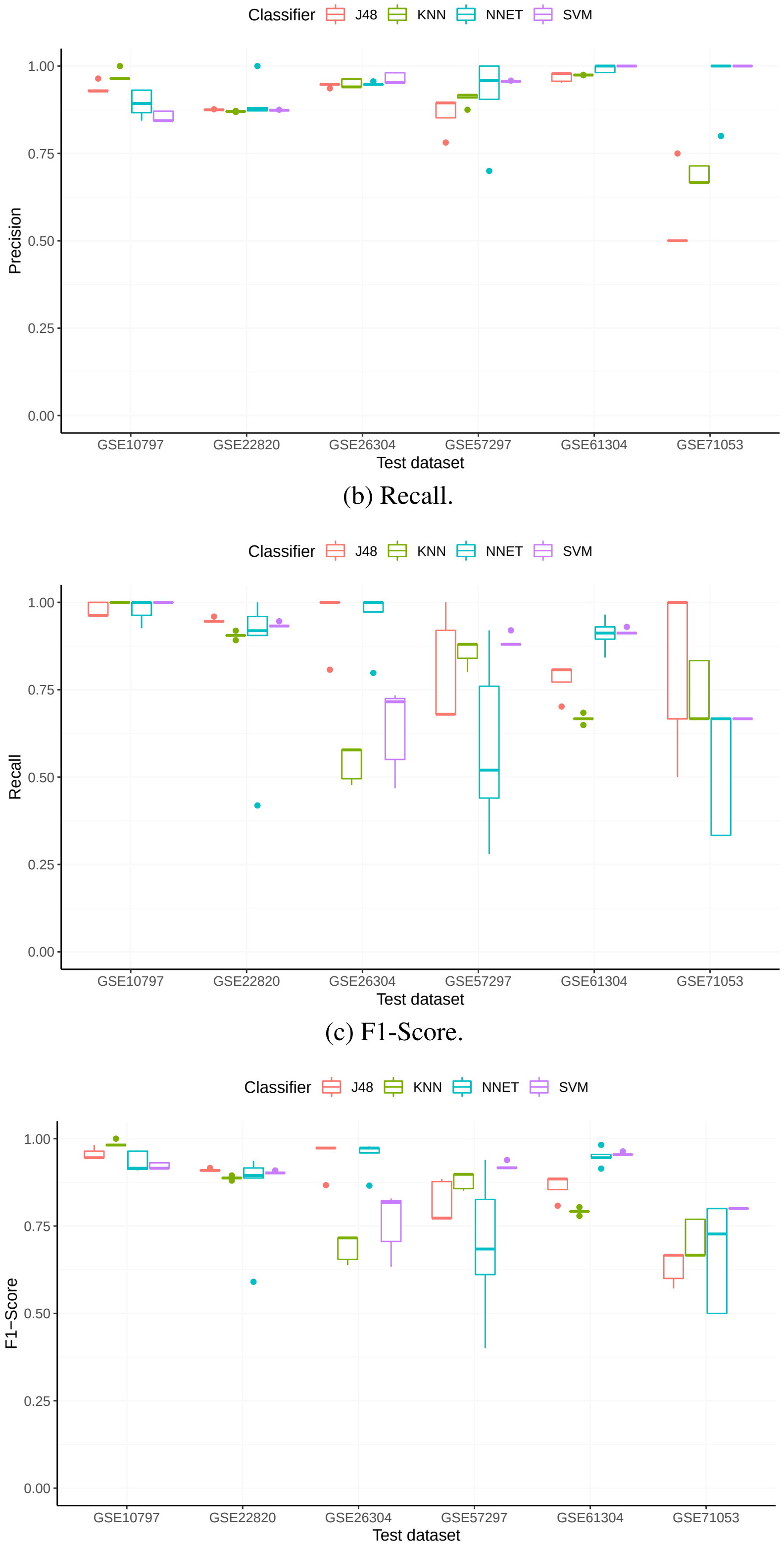
Figure 5.5 – Both threshold and base selector have no impact on performance.


(a) (b)

In general, the ensemble learning feature selection method presented satisfactory performance. Figure 5.6c shows that there is an insignificant difference between the highest median results within test datasets. On recall (5.6b), we can see higher variance in results while on precision (5.6a) we see less variable results. This effect could be caused by a disproportionately smaller number of negative samples compared to the number of positive samples making recall more sensitive to prediction variations. Nonetheless, F1-Score results mostly vary between different test datasets due to the heterogeneity between them (different studies).

Table 5.1 shows the performance summary. There are 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. For example, MAE for GSE10797 is $[(0.963 - 0.928) + (0.963 - 0.925) + (0.963 - 0.924)]/3 \approx 0.04$. MAE can be seen as a relative measure of how much better the best classifier is among the others. Note that when SVM

Figure 5.6 – Classification performance across test datasets using variance and the top 20 genes.



(a) Precision.

(b) Recall.

(c) F1-Score.

- with $\sigma = 0.04$ and $C = 1$ - outperform other classifiers within test datasets, its MAE is higher compared to the other cases. In this way, generally, SVM would be the best choice. On the other hand, in terms of recall, KNN - with $K = 5$ - seems to be a more robust choice (Table 5.2). Due to KNN inductive bias, we can raise the hypothesis that the genes selected by variance clustered well true samples reducing false negative predictions.

Table 5.1 – Classifiers' mean F1-Score by test dataset for *variance* and the top 20 genes. Bold values indicate the highest performance achieved for a dataset.

	Classifier				
Dataset	SVM	J48	KNN	NNET	MAE
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 5.2 – Classifiers' mean Recall by test dataset for *variance* and the top 20 genes. Bold values indicate the highest performance achieved for a dataset.

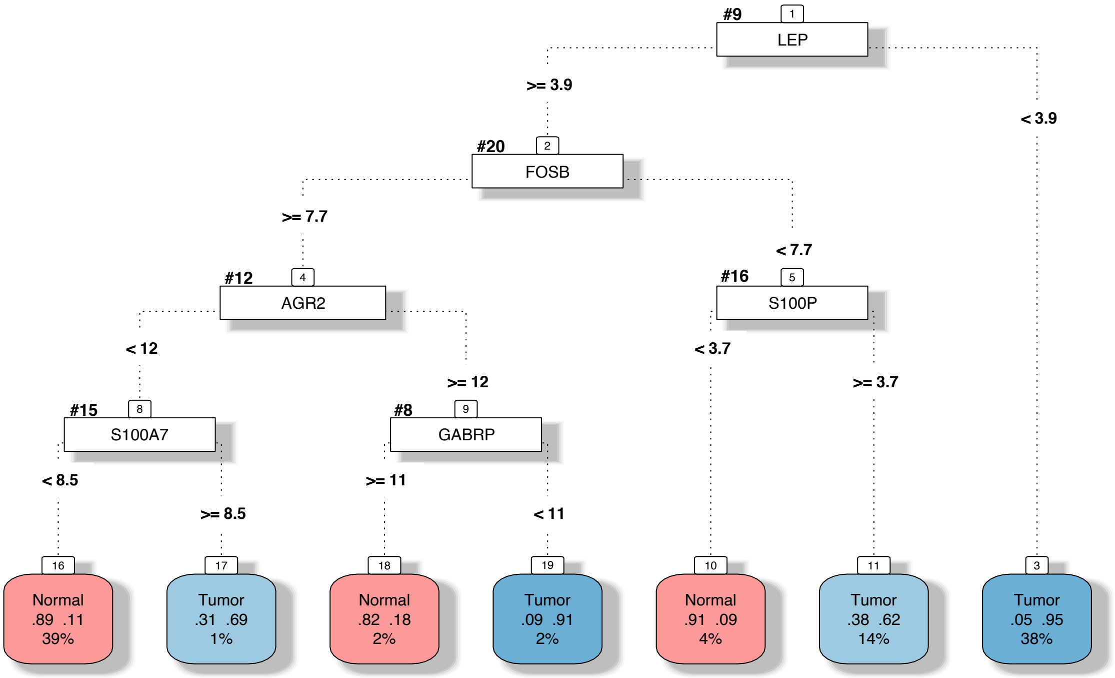
	Classifier				
Dataset	SVM	J48	KNN	NNET	MAE
GSE10797	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.04	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.02	0.03
GSE22820	0.71 ± 0.37	0.87 ± 0.10	$\mathbf{0.89 \pm 0.09}$	0.86 ± 0.23	0.08
GSE26304	0.54 ± 0.43	$\mathbf{0.85 \pm 0.22}$	0.65 ± 0.10	0.81 ± 0.28	0.18
GSE57297	$\mathbf{0.89 \pm 0.09}$	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

An interesting case happens with GSE26304 in which J48 classifier - with $C = 0.5$ and $M = 3$ - presents both the highest F1-Score and recall among other classifiers. To investigate the difference in performance, we plot the model's representation in Figure 5.7. The most informative genes - the ones in and near the root - are not even in the top 6 genes selected by variance. We can see that the gene LEP - the tree's root - is at the ranking's 9^{th} position. J48 also performs satisfactorily for other cases, which makes the classifier an important candidate to understand genes' informative power in a dataset due to its easy interpretation.

5.3 EFS versus Meta-analysis

To validate the proposed EFS method, we compared it with results generated by one of the state-of-the-art methods for candidate biomarkers identification - meta-analysis of transcriptome data. It is important to analyze their differences in performance to understand the scenarios in which one method outperform the other. For Figure 5.8, the EFS

Figure 5.7 – J48 model trained with the top 20 genes selected by variance. Rules' values are expressed in log_2 scale.



results were generated by variance and both methods applied the top 20 genes. The classifiers were able to perform satisfactorily for different sets of genes - one for each method. However, we can see a slightly better performance for EFS. Within each test dataset, it is clear that EFS methodology outperform meta-analysis in most cases (Figure 5.9).

Table 5.3 summarizes the performance of the both methods which corroborates with the presented results. Note that, for GSE613040 and GSE71053, EFS outperform meta-analysis by 0.331 and 0.139, respectively. However, the interesting case lies on GSE26304. As mentioned, J48 performed satisfactorily for GSE26304. We saw that the classifier was not applying the top genes in the model. Instead, J48 identifies genes at middle positions in the ranking as being the most informative. Due to heterogeneity between the rankings and the datasets, GSE26304 case must be further investigated.

Both methods presented satisfactory performance. The biggest difference in performance was found between the test datasets. Meta-analysis presented a high performance for GSE26304. However, in four out of the six test datasets, EFS outperforms the state-of-the-art method. When comparing the top 20 ranking generated by EFS and meta-analysis, a curious finding is that the overlap among them is very low. Only one gene was common to both rankings, namely *S100A7*. Gene *S100A7* is also present in the J48 trained model (Figure 5.7) for GSE26304, which highlights its informative power to this

Figure 5.8 – Classifier general performance comparison between EFS and meta-analysis.

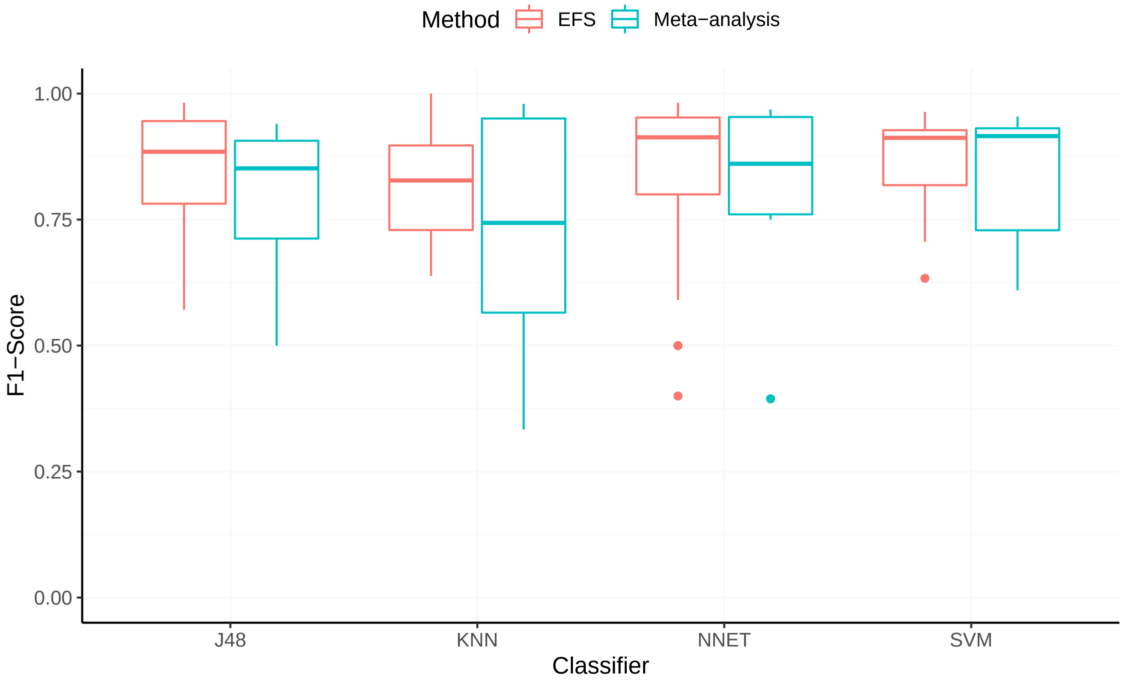
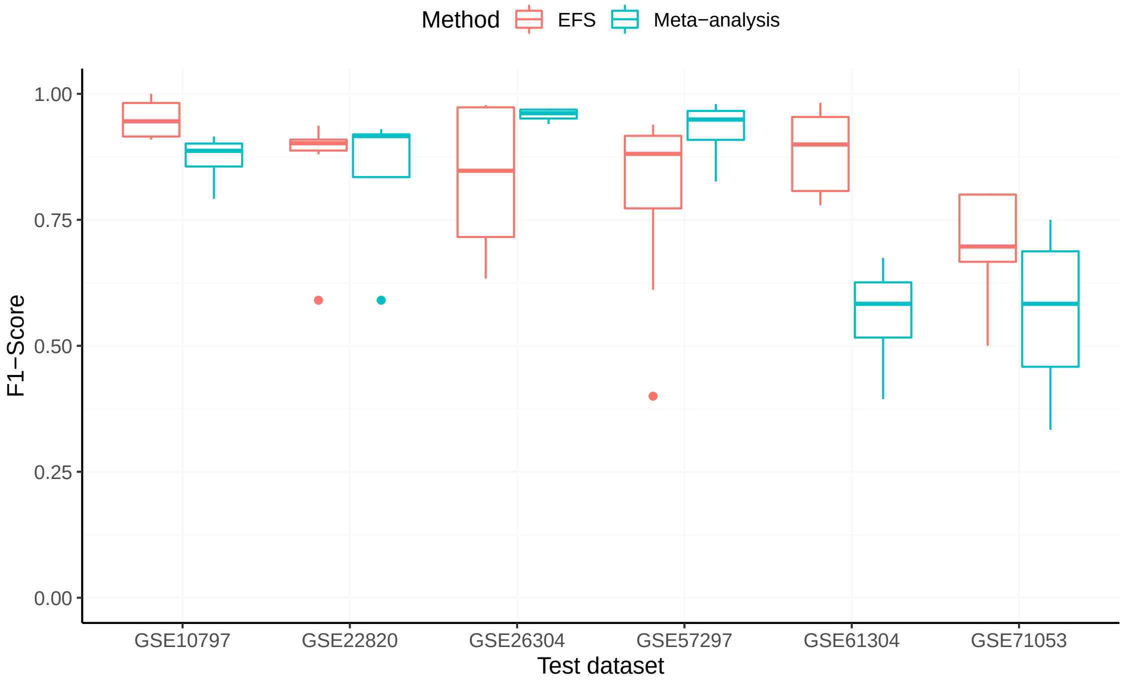


Figure 5.9 – Comparison of EFS and meta-analysis performance within test datasets regardless of classifier.



specific study. According to literature, *S100A7* is highly expressed in breast cancer and may play a role in early tumor progression (EMBERLEY; MURPHY; WATSON, 2004). The high correlation among genes may justify to some extent the low overlap - a factor

Table 5.3 – Methods' mean F1-Score by test dataset for *variance* and the top 20 genes regardless of classifier. Bold values indicate the highest performance achieved for a dataset.

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

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[1]. 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.

Table 5.4 – Top 20 genes found by EFS and meta-analysis ordered by ranking position (#). Bold values indicates matched IDs.

	Gene Symbol	
#	EFS	Meta-analysis
1	SCGB2A2	CXCL10
2	PIP	**S100A7**
3	SCGB1D2	FMOD
4	TFF1	POP1
5	LTF	OXA1L
6	ADIPOQ	ILK
7	KRT14	IGF1
8	GABRP	MX1
9	LEP	MOAP1
10	KRT15	APOD
11	CPB1	BTG1
12	AGR2	ASPM
13	DHRS2	KIF14
14	LPL	CHML
15	**S100A7**	OAS2
16	S100P	CKS2
17	CALML5	MYL9
18	CIDEC	AHCYL1
19	PCOLCE2	SCO2
20	FOSB	BARD1

[1] https://www.genecards.org/

5.4 Functional analysis of most informative genes

To interpret the rankings produced by our EFS approach and the meta-analysis method in terms of biological plausibility, we performed a functional enrichment analysis of results. In Bioinformatics, functional enrichment analyses are carried out to investigate the functional role of a set of genes of interest and extract hypotheses about their relation with the condition under study. These methods aim at detecting pathways or functions over-represented (*i.e.,* significantly associated) in the set of genes relative to what is expected by chance using common statistical tests. For ranked list of genes, the Gene Set Enrichment Analysis (GSEA) is particularly suitable, since it aims at determining whether genes participating in a given biological pathway tend to occur towards the top (or bottom) of the ranked list, suggesting that this pathway is correlated with the disease or condition investigated (SUBRAMANIAN et al., 2005). Therefore, it is not necessary to filter the ranking by using a specific threshold such as the top K genes.

Table 5.5 – KEGG pathways enriched (FDR < 0.05) for the ranking generated by our EFS approach. NES: normalized enrichment score. setSize: number of genes in the ranking that were associated with a given pathway.

ID	Description	setSize	NES	pvalue	FDR
hsa04061	Viral protein interaction with cytokine and cytokine receptor	70	1.635	9.9990e-05	0.0048
hsa04151	PI3K-Akt signaling pathway	264	1.342	9.9990e-05	0.0048
hsa04512	ECM-receptor interaction	71	1.642	9.9990e-05	0.0048
hsa04657	IL-17 signaling pathway	69	1.681	9.9990e-05	0.0048
hsa03320	PPAR signaling pathway	54	1.800	1e-04	0.0048
hsa04974	Protein digestion and absorption	64	1.650	1e-04	0.0048
hsa05150	Staphylococcus aureus infection	42	1.823	0.0001	0.0048
hsa04060	Cytokine-cytokine receptor interaction	189	1.409	0.0001	0.0075
hsa04915	Estrogen signaling pathway	101	1.532	0.0001	0.0075
hsa04610	Complement and coagulation cascades	61	1.579	4e-04	0.0135
hsa00982	Drug metabolism - cytochrome P450	40	1.636	0.0008	0.0245
hsa05144	Malaria	41	1.607	0.0014	0.0394
hsa00350	Tyrosine metabolism	30	1.649	0.0016	0.0417

Table 5.6 – KEGG pathways enriched (FDR < 0.05) for the ranking generated by the REM meta-analysis method. NES: normalized enrichment score. setSize: number of genes in the ranking that were associated with a given pathway.

ID	Description	setSize	NES	p-value	FDR
hsa04110	Cell cycle	99	1.540	9.9990e-05	0.0170
hsa03030	DNA replication	28	1.971	0.0001	0.0170
hsa03410	Base excision repair	28	1.782	0.0002	0.0226

To run this analysis, we used the R package *clusterProfiler* (YU et al., 2012) and the functional annotations (*i.e.,* gene sets) provided by the KEGG Pathway database (KANEHISA; GOTO, 2000). The function *gseKEGG* was adopted, considering pathways composed of three to 800 genes. The function computes an enrichment score (ES), which reflects the degree to which the genes in a gene set are over-represented at the provided ranking, either in top or bottom of it. The normalized enrichment score (NES) is an adjustment of the ES to account for differences in the gene set sizes, allowing a comparison of results across gene sets. We ran 10.000 permutations to assess the statistical significance of the ES and used the False Discovery Rate (FDR) method to adjust p-values for multiple hypothesis testing. Pathways with an adjusted p-value FDR < 0.05 were considered statistically significant in our functional enrichment analysis.

Results for this analysis are shown in Tables 5.5 and 5.6 for EFS and REM rankings, respectively. A total of 13 pathways were enriched for the EFS ranking, whereas three were found significantly associated with the meta-analysis ranking. An interesting finding is that there is no overlap between the pathways enriched in the rankings produced by the two methods. Moreover, in both rankings we can observe pathways that have been previously implicated in cancer. Regarding results obtained for the analysis of EFS ranking, we observed the enrichment of the PI3K-Akt signaling pathway, involved in growth, proliferation, survival, motility, metabolism, and immune response regulation, and also in cancer cell resistance to antitumor therapies (ORTEGA et al., 2020). We also found the signaling pathway by interleukin-17 (IL-17), a family of proinflammatory cytokines with both pro and antitumor effects depending on the conditions (FABRE et al., 2018), significantly associated with this ranking. In Bai et al. (2019), authors also observed the enrichment of the PPAR signaling pathway, ECM-receptor interactuion, IL-17 signaling pathway, Complement and coagulation cascades, and Tyrosine metabolism in their list of differentially expressed genes derived from three breast cancer gene expression datasets.

Regarding the meta-analysis ranking, the three pathways enriched are related to DNA damage and repair, and to cell cycle. The cell cycle pathway is essential for cell growth, proliferation, and reproduction. Deregulation of the cell cycle is a common feature of cancer, enabling limitless cell division and promoting increased susceptibility to the accumulation of additional genetic alterations (MALUMBRES; BARBACID, 2009). In addition, defects in DNA damage and repair machinery are an underlying cause for the development and progression of several types of cancer, including breast cancer. Interestingly, the two most studies genes in breast cancer, BRCA1 and BRCA2, whose mutations are known to predispose to breast and ovarian cancer, transcriptionally regulate some genes involved in DNA repair and cell cycle (YOSHIDA; MIKI, 2004).

The visualization of the GSEA results is provided in Figure 5.10 for the EFS ranking and in Figure 5.11 for the meta-analysis ranking, considering two selected pathways among the most enriched ones.

Figure 5.10 – Significantly enriched pathways in the EFS-based ranking according to the GSEA analysis using KEGG as the annotation database. For both pathways, many genes concentrated in the top of the ranking have participation in the given biological process.

(a) (b)

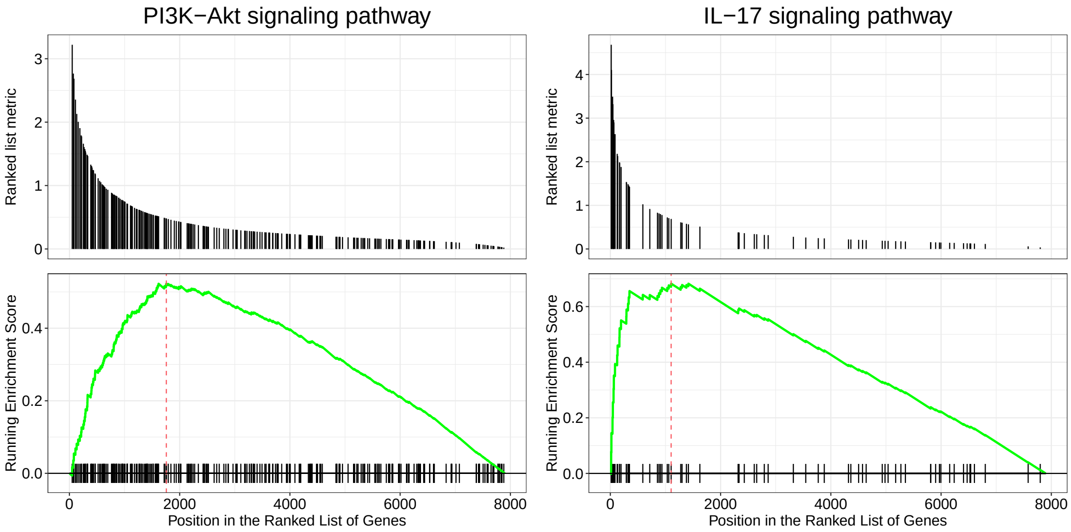
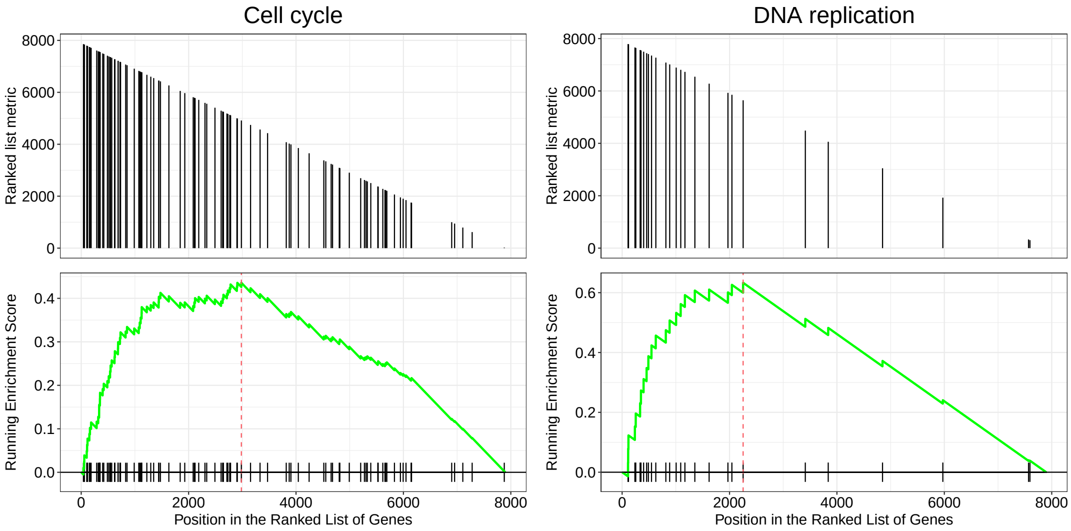


Figure 5.11 – Significantly enriched KEGG pathways in the ranking extracted with the meta-analysis approach according to the GSEA analysis.

(a) (b)



The PI3K-Akt and IL-17 signaling pathways show a maximum deviation of the enrichment score from zero (indicated by the red dashed line) before the top 2000 genes. We may also visually note the larger number of genes involved in the PI3K-Akt signaling pathway, as indicated by the rug plot in the bottom of the plots showing the running enrichment score. In the cell cycle and DNA replication pathways over-represented in the meta-analysis ranking (Figure 5.11), we note that although more genes are involved in cell cycle, the maximum running score for DNA replication is found at a higher position of the ranking. This indicates that genes involved in DNA replication tend to have higher relevance according to the criterion adopted by the REM meta-analysis method to consider genes more informative for breast cancer detection.

6 CONCLUSIONS

Genomics has been an increasing field of study in the last decades. From biological processes to pathological ones, genomics helps to understand the implications of diseases and treatments in a gene level. The human cell holds approximately 20.000 genes. In this sense, high-throughput technologies, such as microarray, have been applied to extract the expression profiles from each one of the 20.000 genes. The set of expression profiles constitutes a high-dimensional dataset. To extract information, and consequently knowledge, from data, advanced computational algorithms are required since state-of-the-art methods no longer perform satisfactorily under these conditions. Feature selection and, mainly, ensemble feature selection approaches have been studied as an alternative tool for knowledge discovery in high-dimensional datasets. EFS has been applied to identify biomarkers in microarray datasets. In this work, we apply an homogeneous EFS to increase stability while maintaining performance in breast cancer biomarker identification. We compare the results with the state-of-the-art method, meta-analysis.

Across all datasets, homogeneous EFS achieved stabilities higher than 0.3 in a range of -1 to 1. Locally, within each dataset, EFS achieved stabilities close to 1, mainly for EFS based on variance feature selection method. Stability findings evidence the heterogeneity between the datasets. Nonetheless, globally and locally, homogeneous EFS based on variance clearly achieved the highest stabilities. Specially for the top 5 genes, variance achieved a global stability close to 0.4 which indicates EFS agrees on the five most informative genes across all datasets.

While stability reflects the quality of selected genes, performance is equally important to assure genes' informative power. Throughout the experiments, we show that the threshold and the type of base selector have no significant impact on performance. On the other hand, SVM outperform other classifiers in most cases in terms of F1-Score for the top 20 genes selected by variance. To increase recall, however, KNN showed better results among the classifiers. We also compared the EFS performance to meta-analysis results. Overall, homogeneous EFS outperfom meta-analysis. In GSE61304, for example, the difference in performance was, on average, 0.331 for F1-Score.

Interestingly, for both rankings we found a strong enrichment of pathways previously implicated in cancer pathogenesis, although more terms were retrieved for the EFS-based ranking. Moreover, no overlap was found among terms enriched for the EFS and meta-analysis rankings. The lack of overlap among over-represented pathways for the rankings derived from distinct analytical strategies may suggest that the EFS-based method proposed in our work reveals complementary mechanisms to the traditional meta-analysis regarding the molecular basis of breast cancer. In other words, each approach may be sensitive to detect distinct types of molecular alterations, which may help elucidating different faces of cancer development or progression. Further exploration of these results is needed to better understand their differences and how they could be jointly explored in the search for candidate cancer biomarkers.

Despite our promising results, important questions still remain unanswered. Are there biological implications within each dataset that can explain the difference in stability and performance between them? Does variance as base selector cluster tumor cases better so KNN is able to increase recall? Are there significant difference between EFS and meta-analysis top 20 genes? Does feature selection methods were able to reduce redundancy between the top genes? Could genes be grouped in order to reduce redundancy? As an important health issue, we must address these important questions in future studies to assure the quality and safety of our approach for practical usage.

REFERENCES

ABEEL, T. et al. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. **Bioinformatics**, v. 26, n. 3, p. 392–398, 2009. ISSN 14602059.

AERTS, S. et al. Gene prioritization through genomic data fusion. **Nature Biotechnology**, v. 24, n. 5, p. 537–544, 2006. ISSN 1546-1696. Available from Internet: <https://doi.org/10.1038/nbt1203>.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. Instance-based learning algorithms. **Machine Learning**, v. 6, n. 1, p. 37–66, 1991. ISSN 1573-0565.

ALEJANDRO, L.-R. et al. Ensemble Feature Selection and Meta-Analysis of Cancer miRNA Biomarkers. **bioRxiv**, p. 353201, 2018.

ALI, M. et al. uEFS: An efficient and comprehensive ensemble-based feature selection methodology to select informative features. **PLOS ONE**, Public Library of Science, v. 13, n. 8, p. 1–28, 2018. Available from Internet: <https://doi.org/10.1371/journal.pone.0202705>.

ANG, J. C. et al. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 13, n. 5, p. 971–989, 2016. ISSN 1557-9964.

ARNOLD, K.; GOSLING, J.; HOLMES, D. **The Java programming language**. [S.l.]: Addison Wesley Professional, 2005.

BAI, J. et al. Screening of core genes and pathways in breast cancer development via comprehensive analysis of multi gene expression datasets. **Oncology letters**, Spandidos Publications, v. 18, n. 6, p. 5821–5830, 2019.

BARRETT, T. et al. NCBI GEO: archive for functional genomics data sets—update. **Nucleic Acids Research**, Oxford University Press, v. 41, n. D1, p. D991–D995, 2012.

Ben Brahim, A.; LIMAM, M. Robust ensemble feature selection for high dimensional data sets. In: **Proceedings of the 2013 International Conference on High Performance Computing and Simulation, HPCS 2013**. [s.n.], 2013. p. 151–157. ISBN 9781479908363. Available from Internet: <https://ieeexplore.ieee.org/abstract/document/6641406/>.

Ben Brahim, A.; LIMAM, M. Ensemble feature selection for high dimensional data: a new method and a comparative study. **Advances in Data Analysis and Classification**, p. 1–16, 2017. ISSN 18625355. Available from Internet: <https://link.springer.com/article/10.1007/s11634-017-0285-y>.

BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Ensembles for feature selection: A review and future trends. **Information Fusion**, v. 52, p. 1–12, 2019. ISSN 1566-2535. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1566253518303440>.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. An ensemble of filters and classifiers for microarray data classification. **Pattern Recognition**, v. 45, n. 1, p. 531–539, jan 2012. ISSN 00313203. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0031320311002718>.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. Feature selection for high-dimensional data. **Progress in Artificial Intelligence**, Springer Verlag, v. 5, n. 2, p. 65–75, may 2016. ISSN 21926360. Available from Internet: <https://link.springer.com/article/10.1007/s13748-015-0080-y>.

BOLÓN-CANEDO, V. et al. A review of microarray datasets and applied feature selection methods. **Information Sciences**, Elsevier Inc., v. 282, p. 111–135, oct 2014. ISSN 00200255. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0020025514006021>.

BREIMAN, L. et al. **Classification and Regression Trees**. [S.l.]: Wadsworth and Brooks, 1984.

BUENO, R. H.; RECAMONDE-MENDOZA, M. Meta-analysis of transcriptomic data reveals pathophysiological modules involved with atrial fibrillation. **Molecular Diagnosis & Therapy**, Springer, v. 24, n. 6, p. 737–751, 2020.

CARVALHO, B. S.; IRIZARRY, R. A. A framework for oligonucleotide microarray preprocessing. **Bioinformatics**, Oxford University Press, v. 26, n. 19, p. 2363–2367, 2010.

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. **J. Artif. Int. Res.**, AI Access Foundation, El Segundo, CA, USA, v. 16, n. 1, p. 321–357, jun. 2002. ISSN 1076-9757.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. ISSN 1573-0565.

Cover, T. M. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. **IEEE Transactions on Electronic Computers**, EC-14, n. 3, p. 326–334, 1965.

CRAMER, H. **Mathematical methods of statistics (PMS-9), volume 9**. [S.l.]: Princeton University Press, 1999. ISBN 9780691005478.

DAS, A. K.; DAS, S.; GHOSH, A. Ensemble feature selection using bi-objective genetic algorithm. **Knowledge-Based Systems**, v. 123, p. 116–127, 2017. ISSN 09507051. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S0950705117300801>.

DERSIMONIAN, R.; KACKER, R. Random-effects model for meta-analysis of clinical trials: An update. **Contemporary Clinical Trials**, v. 28, n. 2, p. 105–114, 2007. ISSN 1551-7144. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1551714406000486>.

DURINCK, S. et al. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. **Bioinformatics**, v. 21, p. 3439–3440, 2005.

DURINCK, S. et al. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. **Nature Protocols**, v. 4, p. 1184–1191, 2009.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: NCBI gene expression and hybridization array data repository. **Nucleic Acids Research**, Oxford University Press, v. 30, n. 1, p. 207–210, 2002.

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap**. [S.l.]: Chapman and Hall/CRC, 1994. ISBN 9780429246593.

EMBERLEY, E. D.; MURPHY, L. C.; WATSON, P. H. S100a7 and the progression of breast cancer. **Breast Cancer Research**, Springer, v. 6, n. 4, p. 1–7, 2004.

FABRE, J. A. S. et al. The interleukin-17 family of cytokines in breast cancer. **International Journal of Molecular Sciences**, Multidisciplinary Digital Publishing Institute, v. 19, n. 12, p. 3880, 2018.

FACELI, K. et al. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Grupo Gen - LTC, 2011. ISBN 9788521618805. Available from Internet: <https://books.google.com.br/books?id=4DwelAEACAAJ>.

FDA-NIH Biomarker Working Group and others. BEST (Biomarkers, endpoints, and other tools) resource [Internet]. Food and Drug Administration (US), 2016. Available from Internet: <https://pubmed.ncbi.nlm.nih.gov/27010052>.

FENG, Y. et al. Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis. **Genes & Diseases**, Elsevier, v. 5, n. 2, p. 77–106, 2018.

GINSBURG, G. S.; PHILLIPS, K. A. Precision Medicine: From Science To Value. **Health affairs (Project Hope)**, v. 37, n. 5, p. 694–701, may 2018. ISSN 1544-5208. Available from Internet: <https://pubmed.ncbi.nlm.nih.gov/29733705https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5989714/>.

HAIDICH, A. B. Meta-analysis in medical research. **Hippokratia**, LITHOGRAPHIA Antoniadis I.-Psarras Th. G.P., v. 14, n. Suppl 1, p. 29–37, dec 2010. ISSN 1790-8019. Available from Internet: <https://pubmed.ncbi.nlm.nih.gov/21487488>.

Hanchuan Peng; Fuhui Long; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1226–1238, 2005.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132733501.

HE, Z.; YU, W. Stable feature selection for biomarker discovery. **Computational Biology and Chemistry**, Elsevier, v. 34, n. 4, p. 215–225, aug 2010. ISSN 14769271. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S1476927110000502>.

HODSON, R. Precision medicine. **Nature**, v. 537, n. 7619, p. S49–S49, 2016. ISSN 1476-4687. Available from Internet: <https://doi.org/10.1038/537S49a>.

IRIZARRY, R. A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. **Biostatistics**, Oxford University Press, v. 4, n. 2, p. 249–264, 2003.

KANEHISA, M.; GOTO, S. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Research**, Oxford University Press, v. 28, n. 1, p. 27–30, 2000.

KENT, J. T. Information gain and a general measure of correlation. **Biometrika**, v. 70, n. 1, p. 163–173, 04 1983. ISSN 0006-3444. Available from Internet: <https://doi.org/10.1093/biomet/70.1.163>.

KHAIRE, U. M.; DHANALAKSHMI, R. Stability of feature selection algorithm: A review. **Journal of King Saud University - Computer and Information Sciences**, King Saud bin Abdulaziz University, jun 2019. ISSN 13191578. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S1319157819304379>.

KOLDE, R. et al. Robust rank aggregation for gene list integration and meta-analysis. **Bioinformatics**, v. 28, n. 4, p. 573–580, 01 2012. ISSN 1367-4803. Available from Internet: <https://doi.org/10.1093/bioinformatics/btr709>.

KUHN, M. Building predictive models in r using the caret package. **Journal of Statistical Software, Articles**, v. 28, n. 5, p. 1–26, 2008. ISSN 1548-7660. Available from Internet: <https://www.jstatsoft.org/v028/i05>.

KUNCHEVA, L. A stability index for feature selection. In: **Proceedings of the IASTED International Conference on Artificial Intelligence and Applications, AIA 2007**. [S.l.: s.n.], 2007. p. 421–427.

KUNCHEVA, L. I.; RODRÍGUEZ, J. J. On feature selection protocols for very low-sample-size data. **Pattern Recognition**, v. 81, p. 660–673, 2018. ISSN 0031-3203. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S003132031830102X>.

LEEK, J. T.; STOREY, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. **PLoS Genetics**, Public Library of Science, v. 3, n. 9, p. e161, 2007.

LIU, H.; LIU, L.; ZHANG, H. Ensemble gene selection by grouping for microarray data classification. **Journal of Biomedical Informatics**, v. 43, n. 1, p. 81–87, 2010. ISSN 1532-0464. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S1532046409001117>.

MALUMBRES, M.; BARBACID, M. Cell cycle, CDKs and cancer: a changing paradigm. **Nature Reviews Cancer**, Nature Publishing Group, v. 9, n. 3, p. 153–166, 2009.

ORTEGA, M. A. et al. Signal transduction pathways in breast cancer: the important role of pi3k/akt/mtor. **Journal of Oncology**, Hindawi, v. 2020, 2020.

PEARSON, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. **The London, Edinburgh, and Dublin Philosophical**

Magazine and Journal of Science, Taylor Francis, v. 50, n. 302, p. 157–175, 1900. Available from Internet: <https://doi.org/10.1080/14786440009463897>.

PES, B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. **Neural Computing and Applications**, 2019. ISSN 1433-3058. Available from Internet: <https://doi.org/10.1007/s00521-019-04082-3>.

PES, B.; DESSÌ, N.; ANGIONI, M. Exploiting the ensemble paradigm for stable feature selection: A case study on high-dimensional genomic data. **Information Fusion**, v. 35, p. 132–147, 2017. ISSN 15662535. Available from Internet: <https://www.sciencedirect.com/science/article/pii/S1566253516300847>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2020. Available from Internet: <https://www.R-project.org/>.

RECAMONDE-MENDOZA, M.; BAZZAN, A. L. Social choice in distributed classification tasks: Dealing with vertically partitioned data. **Information Sciences**, Elsevier Inc., v. 332, p. 56–71, 2016. ISSN 00200255.

RITCHIE, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic acids research**, v. 43, n. 7, p. e47, apr 2015. ISSN 1362-4962 (Electronic).

SAHA, S. K.; SARKAR, S.; MITRA, P. Feature selection techniques for maximum entropy based biomedical named entity recognition. **Journal of Biomedical Informatics**, v. 42, n. 5, p. 905–911, 2009. ISSN 15320464.

SCHUBACH, M. et al. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. **Scientific Reports**, v. 7, n. 1, p. 2959, 2017. ISSN 2045-2322. Available from Internet: <https://doi.org/10.1038/s41598-017-03011-5>.

SEIJO-PARDO, B. et al. Ensemble feature selection: Homogeneous and heterogeneous approaches. **Knowledge-Based Systems**, Elsevier B.V., v. 118, p. 124–139, 2017. ISSN 09507051. Available from Internet: <http://dx.doi.org/10.1016/j.knosys.2016.11.017>.

Shannon, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948.

SHARIFI, S. et al. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. **PLOS ONE**, Public Library of Science, v. 13, n. 2, p. 1–18, 2018. Available from Internet: <https://doi.org/10.1371/journal.pone.0191227>.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 102, n. 43, p. 15545–15550, 2005.

SUROWIECKI, J. **The Wisdom of Crowds**. Knopf Doubleday Publishing Group, 2005. ISBN 9780307275059. Available from Internet: <https://books.google.com.br/books?id=hHUsHOHqVzEC>.

THEIL, H. A note on certainty equivalence in dynamic planning. **Econometrica**, [Wiley, Econometric Society], v. 25, n. 2, p. 346–349, 1957. ISSN 00129682, 14680262. Available from Internet: <http://www.jstor.org/stable/1910260>.

TORO-DOMÍNGUEZ, D. et al. A survey of gene expression meta-analysis: methods and applications. **Briefings in Bioinformatics**, v. 00, n. January, p. 1–12, feb 2020. ISSN 1467-5463. Available from Internet: <https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbaa019/5753843>.

YOSHIDA, K.; MIKI, Y. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to dna damage. **Cancer Science**, Wiley Online Library, v. 95, n. 11, p. 866–871, 2004.

YU, G. et al. clusterprofiler: an R package for comparing biological themes among gene clusters. **Omics: A Journal of Integrative Biology**, Mary Ann Liebert, Inc., v. 16, n. 5, p. 284–287, 2012.

YU, L.; LIU, H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In: **Proceedings, Twentieth International Conference on Machine Learning**. [S.l.: s.n.], 2003. v. 2, p. 856–863. ISBN 1577351894.

ZHANG, X.; JONASSEN, I. An Ensemble Feature Selection Framework Integrating Stability. **Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019**, p. 2792–2798, 2019.