Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
DATA SCIENCE AND COMPUTATION

Ciclo XXXIII

Settore Concorsuale: 09/H1 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

Settore Scientifico Disciplinare: ING-INF/05 - SISTEMI DI ELABORAZIONE DELLE INFORMAZIONI

Challenges and Opportunities of Machine Learning for Clinical and Omics Data

---

Presentata da: Luca Pestarino

Coordinatore Dottorato:
Prof. Andrea Cavalli

Supervisore:
Prof. Andrea Cavalli

Co-supervisore:
Dr. Sergio Decherchi

Esame finale anno 2022

"Cloudless everyday
You fall upon my waking eyes,
Inviting and inciting me to rise.
And through the window in the wall
Come streaming in on sunlight wings
A million bright ambassadors of morning.

And no one sings me lullabies
And no one makes me close my eyes
So I throw the windows wide
And call to you across the sky..."

"Echoes", Pink Floyd - 1971

Abstract

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA

Ph.D. in Data Science and Computation

Challenges and Opportunities of Machine Learning for Clinical and Omics Data

Luca Pestarino

Clinical and omics data are a promising field of application for machine learning techniques even though these methods are not yet systematically adopted in healthcare institutions. Despite artificial intelligence has proved successful in terms of prediction of pathologies or identification of their causes, the systematic adoption of these techniques still presents challenging issues due to the peculiarities of the analysed data. The aim of this thesis is to apply machine learning algorithms to both clinical and omics data sets in order to predict a patient's state of health and get better insights on the possible causes of the analysed diseases. In doing so, many of the arising issues when working with medical data will be discussed while possible solutions will be proposed to make machine learning provide feasible results and possibly become an effective and reliable support tool for healthcare systems.

Acknowledgements

First and foremost I am extremely grateful to my supervisor Professor Andrea Cavalli and my co-supervisor Sergio Decherchi, Ph.D. for their support, advice and guide during this Ph.D. and in the thesis development. I would also like to thank professor Paolo Vineis and his team whose help and ideas enlightened us in many research questions, and all the institutions I had the pleasure to collaborate with: the Italian Institute of Technology, the Bambino Gesù Children's Hospital, the Rizzoli Orthopaedic Institute and the Imperial College London. I also want to thank all my former and current colleagues at the Department of Pharmacy and Biotechnology: Daniele, Nicola, Marco, Dario, Martina, Matteo and Riccardo for all the fruitful discussions and time spent together during these years. I would like to give special thanks to Chiara, Erika and Laura, not simply colleagues but also friends and companions during this long journey. My working days would have not been the same without all of you.

I would like to express my gratitude to my family and my friends who probably never actually understood what I was working on but have always supported me in my choices and in my darkest times. I would also like to thank Alessio: even though you are the last arrived in my life, you are the only one who has ever truly understood me and you have always helped me and encouraged me in my most difficult choices.

Finally, a very special thanks goes to Asim. Although you are no longer part of my life you have always believed in me and all of this would have not been possible without you and your help.

Contents

List of Figures

List of Tables

List of Abbreviations

ADI	**A**utism **D**iagnostic **I**nterview
ADOS	**A**utism **D**iagnostic **O**bservation **S**chedule
AI	**A**rtificial **I**ntelligence
ASD	**A**utism **S**spectrum **D**isorder
AUC	**A**rea **U**nder the **C**urve
BMI	**B**ody **M**ass **I**ndex
BMD	**B**one **M**ineral **D**ensity
CBCL	**C**hild **B**ehavior **C**heck**L**ist
CNN	**C**onvolutional **N**eural **N**etwork
COPD	**C**hronic **O**bstructive **P**ulmonary **D**isease
CVD	**C**ardio-**V**ascular **D**isease
DGS	**D**isturbo **G**eneralizzato dello **S**viluppo
DSM	**D**iagnositc and **S**tatistical **M**anual of mental disorders
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
KNN	**K**-**N**earest **N**eighbours
IHD	**I**schemic **H**eart **D**isease
IQ	**I**ntelligence **Q**uotient
FPR	**F**alse **P**ositive **R**ate
ML	**M**achine **L**earning
MRI	**M**agnetic **R**esonance **I**maging
MSE	**M**ean **S**quared **E**rror
OI	**O**steogenesis **I**mperfecta
PCA	**P**rincipal **C**omponent **A**nalysis
PP	**P**rincipal **P**ath
PVB	**P**rimo **V**ocabolario del **B**ambino
RBF	**R**adial **B**asis **F**unction
RFE	**R**ecursive **F**eature **E**elimination
ROC	**R**eceiver **O**perating **C**haracteristic
SNE	**S**tochastic **N**eighbour **E**mbedding
SVM	**S**upport **V**ector **M**achine
t-SNE	**t**-distributed **S**tochastic **N**eighbour **E**mbedding
TNR	**T**rue **N**egative **R**ate
TPR	**T**rue **P**ositive **R**ate
TRAP	**T**raffic **R**elated **A**ir **P**ollution
VABS	**V**ineland **A**daptive **B**ehavior **S**cales

To my Elah...

Chapter 1

Introduction

In recent years, an always increasing amount of available data across all fields of science has been observed creating the need to employ and develop automated methods to automatically analyse them. Machine learning is a subfield of computer science that deals with the discovery of correlations and emerging patterns inside data and it is used to make predictions on new ones [1]. Machine learning techniques are particularly useful for several data analysis tasks, such as computer vision, computational biology, text processing, speech recognition, robotics and many others. Clinical data sets represent a promising field of application as clinical machine learning is not yet systematically adopted in organisations such as hospitals and clinics, even if successful applications are starting to appear [2].

1.1 Clinical machine learning

Machine learning techniques can be divided into two main categories: supervised and unsupervised learning techniques. Supervised learning is a predictive approach with the aim to learn a function relating the input with the output. In this case, the labels (the outcome of interest) are needed in order to learn the function and make predictions. On the other hand, unsupervised learning has the goal to find patterns or hidden structures in the data independently from the labels since this information is not available when performing this kind of analysis (i.e. there is not a known output to predict) [1]. Both supervised and unsupervised learning models can be fruitfully applied to clinical studies: machine learning techniques have been successfully used in several fields ranging from bioinformatics, genomics, medical informatics all the way to public health, bringing about a higher understanding of various diseases including cancer, heart failure, tissue classification and drug design [3]. Clinical machine learning analysis can be employed for different tasks according to the specific aim or research field. One of the most common is certainly the use of machine learning to perform a classification among patients in order to identify their diagnosis. Unsupervised learning, on the other hand, is typically more employed for precision medicine with the aim to redefine diseases and identify new paths to therapy by studying pathophysiologic mechanisms [4]. Clinical data analytics can bring many advantages to healthcare systems such as identifying the most effective treatments,

predict which individuals will benefit of a specific medication, or support prevention initiatives. All the potential benefits of analysing healthcare data include detecting a disease in its early stage in order to treat it more efficiently, decide whether a patient will benefit from surgery or not, identify risks for medical complications and many others. At the same time, machine learning research could lead to find new cost-effective ways to treat patients, improve clinical trials design and to detect disease patterns to improve public health surveillance and response [5]. Machine learning also allows to identify the features that better describe the analysed phenomenon. Since having too many features can cause overfitting problems and bring noise into the model, one generally wants to reduce the number of features by performing feature selection. By doing so, one can extract a subset of features that can be used to make the best prediction [6]. However, in order to do so in a reliable way, one needs to collect a large amount of unbiased data and possibly collect the same information on other independent cohorts to be used for testing the results. Unfortunately, this process is time consuming and typically data acquisition can be very expensive [4].

Although artificial intelligence (AI) began in 1956, its first application to clinical scenarios dates back to 20 years later, in 1976, when the first expert system in medicine (MYCIN) was developed to recommend antibiotic for different bacterial infections [7]. Due to its limitations, such as the large number of rules needed and the absence of integration between the system and the clinicians, MYCIN has never been used in practice. Despite machine learning can overcome some of the expert systems limitations, the first autonomous system approved by the FDA dates back only to 2018, with an AI-based diagnostic system for detection of diabetic retinopathy in primary care offices [8].

In recent years, deep learning methods have proved their efficacy in many prediction and classification tasks such as speech recognition and image classification. Deep learning is a subclass of machine learning which makes use of artificial neural networks (NN). NN have been particularly effective in studying and understanding genomics data, typically characterised by sparsity and high dimensionality. For example, deep convolutional neural networks (CNN) have been used to predict sequence function and activity in various cell types or to predict tissue functional activities from genomic sequences [9]. However, one of the most promising fields of application for clinical machine learning, is medical imaging. Recently, machine learning has shown the ability to see patterns that humans cannot perceive, increasing its popularity especially in radiology [6]. Other applications include predicting prognosis for patients with non-small cell carcinoma from histopathology images, diagnosing diabetic retinopathy from retinal fundus photographs [10], [11]. Authors in [2] applied a deep neural network to analyse three-dimensional optical coherence tomography scans for making referral recommendations, achieving the same or better performance of experts after training the algorithm on 14 884 scans. On the other hand, Authors in [12] developed a deep learning model built upon CNNs to analyse tumor images and extract informative features by optimising the partial likelihood

of a proportional hazards model. Another study on breast cancer has been described in [13], where deep learning and other machine learning algorithms, namely support vector machines, random forests and decision trees, have been used to predict the survival in 4902 patients records. In this case, however, clinical data have been employed rather then images, allowing Authors to extract the most important factors in survivability prediction. Other studies involved the use of brain magnetic resonance imaging (MRI) such as in [14], where a deep learning algorithm combining stacked auto-encoders and a softmax output layer has been implemented to early diagnose Alzheimer's disease, or in [15] where a deep belief network has been used to identify patterns of similarity on a data set of Alzheimer's disease and healthy samples.

The use of machine learning can thus provide better insights into medical analysis, showing possible new ways to address medical problems such as the prediction of pathologies or the identification of their causes while improving care, saving lives and lowering costs [5].

Even though these methods have provided successful results, the use of these techniques remains a challenging issue when it comes to their application to medical and biological data, due to the high dimensionality of the data as well as their heterogeneity, irregularity and temporal dependency[16]. These difficulties arise from the very nature of the data itself: clinical data sets often come from different sources of information such as biomedical data, experimental data and electronic health records [17]. The data stored can thus be expressed in different forms: they can be discrete or continuous, in image formats, tables, charts o clinical notes [18]. All of these data further increase the complexity and the dimensionality of the analysis creating a need to identify new ways to process the data such as the resort to cloud computing, multicore CPUs and GPUs [17]. Additionally, one of the main problems is the reliability of samples labels: clinicians are not always certain about the diagnosis or may disagree with other experts. Moreover, machine learning outputs (especially deep learning ones) are often difficult to interpret and new methods to understand how deep learning algorithms work are needed [9]. Furthermore, health systems do no completely trust the tasks performed by a machine, especially if the physicians can complete that task with a higher accuracy and explainability. Therefore, there is a need to integrate machine learning with physicians activity, overcoming the idea of machine learning as a "black box" [4]. A machine learning model is defined as a black box when the function is too complicated to be understood by humans and it is not possible to easily explain how it makes predictions [19]. For this reason, the main example of black box models are deep neural network since they are highly recursive and typically they are not able to produce comprehensible decisions making it difficult to trust their reliability for real-world problems [20]. In particular, the interpretability of a machine learning model is crucial when dealing with the medical field where the comprehension of how prediction are made is essential to understand the mechanisms behind the diagnosis [21]. Although the application of machine learning techniques presents a variety of challenges and difficulties, it also

shows great potential for delivering support tools for diagnosis and broadening the understanding of factors that may lead to the development of pathologies.

1.2 Overview of the thesis

The aim of this thesis is to apply machine learning techniques on different medical and biological data sets in order to predict a patient's health's state or to stratify them in sub-cohorts possibly exploiting both strictly clinical and omics data. Additionally, the identification of the features that are the most significant for the prediction of the health status and conducive for achieving better insights in possible causes of the disease will be addressed. Moreover, the issues regarding these analyses will also be considered and discussed in order to identify new ways to mitigate or solve them and get more reliable results. The data sets that will be used deal with patients affected by different disorders and diseases. In particular, the research will be focused on five data sets: four of these exhibit purely clinical features while the fifth one contains omics data, namely transcriptomics, adductomics, miRNA and metabolomics features. Different analyses will be performed according to the considered type of data. When dealing with clinical data sets the main goal is to use the clinical features to predict the diagnosis of the patients. Here the aim is not only to identify which clinical variables are more important for the prediction to simplify the diagnosis process, but also to understand the level of consistency of the labels assigned by the physicians. All the clinical data sets taken in consideration do not include healthy patients: this place a first limit in the analysis since discriminating healthy patients from sick ones is typically easier than classifying subgroups due to possible similarities between them. The first employed data set regards patients affected by different kind of disorders such as language disorders, intellectual disabilities and learning disorders including dyslexia, dyscalculia and dysgrafia. The second data set is about patients with various mental disorders such as autism spectrum disorder, Asperger syndrome and other language or cognitive disorders. The third and fourth data sets include patients affected by different types of osteogenesis imperfecta. For each of these data sets, both supervised and unsupervised learning analysis were performed in order to get more information about the diagnosis and its causes and possibly identify alternative subgroups of patients according to their similarity. When it comes to the fifth data set, it is an omics data set (known as Oxford Street II data set) originally designed to study the impact of air pollution on patients affected by different diseases or in healthy status. Having at disposal healthy patients data, it was possible to move away from the precision medicine design of the clinical data sets to perform an in-depth analysis of features affecting the disease. More specifically, some issues such as the feature selection stability, the presence of confounders, the causality between the features and the outcome, and the study of the disease progression will be addressed as well as the biological meaning of the features involved in the analysis. This thesis is organised as follows: the methods

employed for the analyses will be discussed in Chapter 2. Chapter 3 includes the description of each data set, the aim of each analysis and the associated results, while the issues encountered in the analyses and the final remarks will be discussed in Chapter 4.

Chapter 2

Methods

In this chapter, the known methods and algorithms employed for the analyses described in Chapter 3 will be discussed, whereas the methodological contributions will be examined along with the related results in the next chapter.

As previously mentioned, clinical data analysis can be performed employing all the variety of machine learning methods, ranging from classical supervised and unsupervised techniques to more elaborated deep learning models. However, in this thesis deep learning algorithms have not been employed. This choice stems from two specific reasons: the first is an intrinsic limit of the data sets analysed while the other is a limit of the methodology itself. Although in literature many examples of deep learning applications have been discussed, the data sets at disposal had very limited sample sizes making deep learning very difficult to implement practically. For this reason, other *classical* methods have been preferred to perform the analyses. Moreover, since one of the main goals is to identify possible new causes of diseases and thus better understand their mechanisms, one needs to employ highly interpretable models, automatically excluding all the "black box" algorithms such as deep neural networks. The exploited methods, including some of the classical machine learning algorithms, such as classification trees, lasso, support vector machine, etc. and other more recent methods such as the principal path [22], will be now described. The data pre-processing techniques, the employed stability indices and evaluation criteria will also be discussed in the following sections.

2.1 Supervised methods

Supervised learning are predictive models, thus the outcome of the analysis (the labels) has to be provided in order to build the model. According to the nature of the labels, supervised learning techniques can be further subdivided into two main categories: classification and regression methods [23]. Classification methods are employed when the labels are expressed as classes values, which can be binary (i.e. $y = \{0, 1\}$) or multiclass (i.e. the number of classes is larger than two). On the other hand, regression is employed when the labels are in form of continuous values. The algorithms/protocols used are the following:

- K-nearest neighbours.

- Lasso and elastic net.
- Support vector machine.
- Classification tree.
- Random forest.

2.1.1 K-nearest neighbours

The nearest-neighbour is a very simple algorithm which classifies samples according to the observations in the training set which are closest to the data point x [24]. The k-nearest neighbours (KNN) fit $\hat{Y}$ according to:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{2.1}$$

where y_i is the data point label and $N_k(x)$ is the neighbourhood of x identified by the k closest points to x_i in the training set. The algorithm simply performs an average of the closest points to x according to a metric that typically is set to Euclidean distance. The k-nearest neighbours has been used in this thesis when relabelling the samples with the principal path approach as described in section 3.4.7 when dealing with the continuous diagnosis issue.

2.1.2 Lasso and elastic net

The lasso (least absolute shrinkage and selection operator) was originally introduced in [25] as a mean to perform feature selection in high dimensionality scenarios. It can be used for both classification and regression problems.
Given p predictors $\mathbf{x}_1, ..., \mathbf{x}_p$, the outcome $\mathbf{y}$ can be predicted by:

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + ... + \mathbf{x}_p\hat{\beta}_p. \tag{2.2}$$

This linear model can be estimated by using the ordinary least squares (OLS), by minimising the residual sum of squares. While OLS estimates have often low bias, they typically have large variance. This issue can be mitigated by shrinking some coefficients towards 0. One possible solution is to employ the ridge regression [26]: by shrinking the coefficients, it improves the stability but it does not set any coefficient exactly to 0 and thus it does not perform a proper feature selection. The lasso, on the other hand, can set some of the coefficients to 0, keeping only the most important ones. The lasso solution is obtained by minimising the following cost function:

$$\min_{\beta_0, \beta} \quad \frac{1}{N} \sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \lambda \left\| \beta \right\|_1 \tag{2.3}$$

where $l(y_i, \beta_0 + \beta^T x_i)$ is the square loss function for the i^{th} observation and λ is the tuning parameter controlling the shrinkage penalty. The ridge regression has the same functional of (2.3) with a different penalty:

$$\min_{\beta_0,\beta} \quad \frac{1}{N}\sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \lambda \|\beta\|_2^2 . \tag{2.4}$$

Fig. 2.1 shows the difference between lasso and ridge regression estimates with two parameters. The shaded areas represent the constraint regions, while the ellipses are the contours of the OLS error function centered on the OLS estimates. As one can see, the solutions are in the points where the ellipses hit the shaded area, however, differently from the ridge, the lasso area has a diamond shape and when the ellipse hits the corner one parameter is equal to zero.

Thanks to its sparse representation, the lasso is thus more appealing in those contexts where there is a large number of predictors and one wants to perform feature selection. Nonetheless, the lasso shows some limitations. First of all, if there are some highly correlated variables the lasso tends to choose one of those while excluding the others in a random way [27]. Secondly, due to its convex optimisation problem, if the number of predictors is larger than the number of samples ($p > n$), the lasso can select at most n features. Moreover, in situations where $n > p$ with high correlations among the variables, ridge regression typically performs better than lasso. In order to mitigate these issues, another regularisation technique named elastic net has been proposed in [27]. Elastic net can simultaneously perform feature selection by also selecting groups of correlated features, by combining the penalties of both ridge regression and lasso, overcoming their respective limitations. In other words, it performs feature selection with the stability and prediction ability typical of Tikhonov regularization [26].

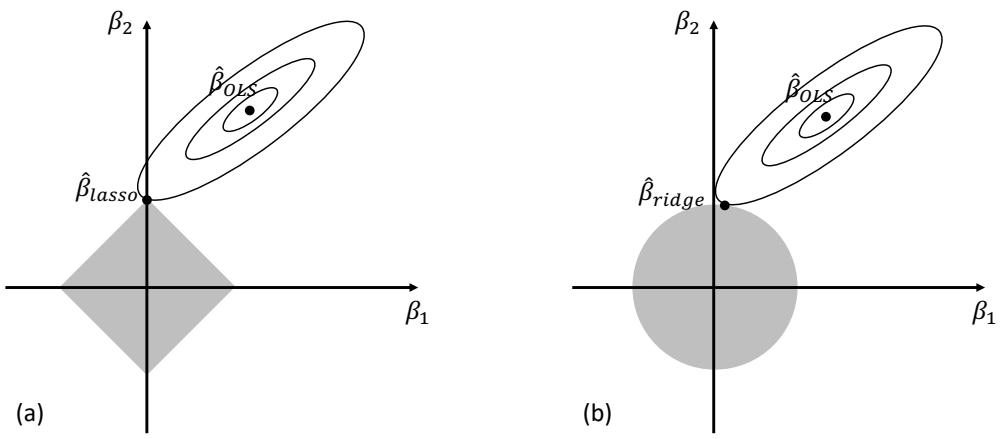


FIGURE 2.1: Lasso (a) and ridge regression (b) coefficients estimation.

The elastic net solution is obtained by solving the following minimisation problem:

$$\min_{\beta_0,\beta} \quad \frac{1}{N}\sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \tag{2.5}$$

where λ_1 and λ_2 are the penalties associated with lasso and ridge regression respectively. Let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, then solving (2.5) is equivalent to:

$$\min_{\beta_0,\beta} \quad \frac{1}{N}\sum_{i=1}^{N} l(y_i, \beta_0 + \beta^T x_i) + (1-\alpha) \|\beta\|_2^2 + \alpha \|\beta\|_1 . \tag{2.6}$$

As before, $l(y_i, \beta_0 + \beta^T x_i)$ is the square loss function for the i^{th} observation while α controls the elastic net penalty. Elastic net can thus be viewed as a generalisation of the lasso or ridge regression in the special cases $\alpha = 1$ and $\alpha = 0$ respectively.

When it comes to identifying the lasso and elastic net solution, it is important to specify that there is not a closed form solution to the lasso problem and it has to be solved using an iterative approach such as the coordinate descent algorithm [28]. When $\lambda_2 > 0$, elastic net does not suffer of the same issue, having a strictly convex penalty function and thus a unique solution. However, it can be proved that in case of an orthonormal input matrix, minimising (2.6) is the same as minimising a lasso problem. In particular, the elastic net solution is as follows:

$$\hat{\beta}_{(elastic\,net)} = \frac{(|\hat{\beta}_i(OLS)| - \lambda_1/2)_+}{1+\lambda_2} sgn\{\hat{\beta}_i\}(OLS) \tag{2.7}$$

where $\hat{\boldsymbol{\beta}}(OLS) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ whereas the lasso solution would have been:

$$\hat{\beta}_{(lasso)} = (|\hat{\beta}_i(OLS)| - \lambda_1/2)_+ sgn\{\hat{\beta}_i\}(OLS). \tag{2.8}$$

Lasso and elastic net are thus two very useful algorithms when working in clinical scenarios with more predictors than samples. Despite lasso's sparsity advantage, elastic net would be preferred in those cases with highly correlated features as in the case of omics data. For the same reason, elastic net has proved to be more stable in terms of selected features when dealing with correlated features scenarios.

2.1.3 Support vector machine

The support vector machine (SVM) for classification is a well-established and powerful supervised learning method. It identifies a maximal margin hyperplane between two classes [29]. SVM algorithm uses kernels to generalise the hyperplane concept to a nonlinear separating surface.

As described in [29], given a set of samples $\{\mathbf{x}_i, y_i\}$, $i = 1, ..., n$, $y_i \in \{-1, 1\}$, $\mathbf{x}_i \in \mathbb{R}^d$, one supposes to have a separating hyperplane which can separate positive samples from negative. All the point lying on the hyperplane must satisfy $\mathbf{w} \cdot \mathbf{x} + b =$

0, where $\mathbf{w}$ is normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the distance from the hyperplane to the origin and $\|\mathbf{w}\|$ is the Euclidean norm of $\mathbf{w}$. If the data is linearly separable all the data points satisfy the following relations:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \; for \; y_i = +1 \tag{2.9}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \; for \; y_i = -1 \tag{2.10}$$

that can be combined into:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \; \forall i. \tag{2.11}$$

The points for which the equality in (2.9) and (2.10) hold, lay on the two hyperplanes respectively, with distance $|1-b|/\|\mathbf{w}\|$ and $|-1-b|/\|\mathbf{w}\|$ from the origin, thus the two distances are the same and equal to $1/\|\mathbf{w}\|$ while the margin is $2/\|\mathbf{w}\|$. The two hyperplanes are thus parallel and there are no points between them. Therefore, it is possible to identify the hyperplanes for which the margin is maximum by minimising $\|\mathbf{w}\|^2$ subject to (2.11). The points laying on the hyperplanes are called support vectors since their removal would change the solution. If one expresses the problem in a Lagrangian form one gets:

$$\min_{\mathbf{w},b} \quad L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{n} \alpha_i \qquad \text{s.t.} \quad \alpha_i \geq 0 \tag{2.12}$$

where α_i, $i = 1, ..., n$ are positive Lagrange multipliers. Being this a convex quadratic programming problem, it is possible to solve the strongly dual problem, that is maximising L_P. Given the following optimal relations:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{2.13}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.14}$$

one can then substitute these values into (2.12) to get the dual problem:

$$\begin{aligned} \max_{\alpha_i} \quad & L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{s.t.} \quad & \alpha_i \geq 0 \\ & \sum_i \alpha_i y_i = 0. \end{aligned} \tag{2.15}$$

The equations shown until now hold for the linear separable case. However, linear separability does not always occur in practice, hence this algorithm would not identify a feasible solution. To solve this issue, it is possible to relax the constraints (2.9) and (2.10) by introducing positive slack variables ξ_i, $i = 1, ..., n$ as

in [30]:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \xi_i \quad for\ y_i = +1 \tag{2.16}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i \quad for\ y_i = -1 \tag{2.17}$$

$$\xi_i \geq 0\ \forall i. \tag{2.18}$$

A way to introduce an extra cost for errors is to change the objective function from $\|\mathbf{w}\|^2/2$ to $\|\mathbf{w}\|^2/2 + C(\sum_i \xi_i)^k$, where C is a parameter regulating the errors weight. When C is very high, even a small value of ξ_i determines a large increase in the cost function. At the same time, with a small value of C the cost function is less sensitive to an increase in the value of ξ_i. The dual formulation now becomes:

$$\begin{aligned} \max_{\alpha_i} \quad & L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \\ & \sum_i \alpha_i y_i = 0 \end{aligned} \tag{2.19}$$

with solution:

$$\mathbf{w} = \sum_{i=1}^{N_S} \alpha_i y_i \mathbf{x}_i \tag{2.20}$$

where N_S is the number of support vectors. Fig. 2.2 shows an example of linear separating hyperplanes with separable data (a) and non-separable data (b) in a two-dimensional space: the circled dots represent the support vectors.

Moreover, this formula can be generalised to the case where there is not a linear decision function. Suppose one can map the data to another space H, by using a function $\phi()$. By defining a kernel function K as a function of sample pairs only $K(\mathbf{x}_i, \mathbf{x}_j)$ and the Mercer condition holds [31], namely the induced kernel matrix is

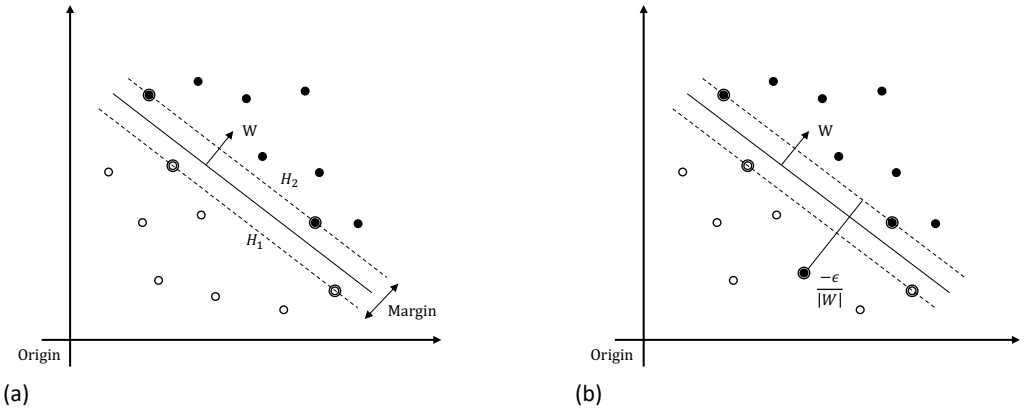


FIGURE 2.2: Linear separating hyperplanes with separable data (a) and non separable data (b).

positive semi-definite, then the kernel can be expressed as a dot product in a proper, possibly infinite dimensional, space $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Hence one can only use K without explicitly knowing the form of Φ. By doing so, the algorithm only depends on kernel evaluations. (2.19) now becomes:

$$\begin{aligned} \max_{\alpha_i} \quad & L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq C \\ & \sum_i \alpha_i y_i = 0. \end{aligned} \tag{2.21}$$

If the kernel is a linear function, then (2.21) is the same of (2.19), however one can decide to employ different kernel functions. One of the most popular ones is the Gaussian RBF (Radial Basis Function) kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 / 2\sigma^2}. \tag{2.22}$$

Hence one is still performing a linear separation but in a different space. One of the advantages of using a linear kernel is a more interpretable model. Moreover, thanks to its linearity, it is possible to evaluate the weights of each feature and order them to identify the most important ones. Finally, despite all the above described algorithms are discussed for the two classes case they can be extended to multiclass problems in a one-versus-all way.

2.1.4 Support vector regression

The Support vector machine was originally conceived to solve classification tasks. However, a variant for regression problems has been proposed in [32]. The regression version of SVM is based on the so called epsilon intensive loss function:

$$L_\epsilon(y, \hat{y}) = \begin{cases} 0 & \text{if } |y - \hat{y}| < \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise} \end{cases} \tag{2.23}$$

where ϵ indicates the ϵ-tube within which the prediction error is not considered. Therefore the objective function becomes:

$$J = C \sum_{i=1}^{N} L_\epsilon(y_i, \hat{y}_i) + \frac{1}{2} \|\mathbf{w}\|^2 \tag{2.24}$$

where $\hat{y}_i = f(x_i) = \mathbf{w}^T \mathbf{x}_i + b$ and C is the usual inverse regularisation parameter. To solve this problem, one can add the slack variables ξ_i to the objective function as constraints in order to measure the deviation of training samples from the ϵ-tube [33]:

$$
\begin{aligned}
J &= C\sum_{i=1}^{N}(\xi_i + \xi_i^*) + \frac{1}{2}\|\mathbf{w}\|^2 \\
\text{s.t.} \quad & y_i \leq f(\mathbf{x}_i) + \epsilon + \xi_i \\
& y_i \geq f(\mathbf{x}_i) - \epsilon + \xi_i^* \\
& \xi_i \geq 0 \\
& \xi_i^* \geq 0
\end{aligned}
\tag{2.25}
$$

This primal formulation, as before, can be put in a dual form whose solution is given by:

$$\hat{\mathbf{w}} = \sum_{i}^{N}(\alpha_i - \alpha_i^*)\mathbf{x}_i \tag{2.26}$$

where α_i and α_i^* are Lagrange multipliers. Here the support vectors are defined as the $\mathbf{x}_i$ for which α_i and α_i^* are non-zero, which are the points lying on or outside the ϵ-tube. Plugging in (2.26) into $f(\mathbf{x}) = \hat{\mathbf{w}}^T\mathbf{x} + b$ one gets the dual version:

$$f(\mathbf{x}) = \sum_{i}^{N}(\alpha_i - \alpha_i^*)\mathbf{x}_i^T\mathbf{x} + b. \tag{2.27}$$

From this, as did with the standard SVM, one can easily apply kernel functions:

$$f(\mathbf{x}) = \sum_{i}^{N}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i, \mathbf{x}) + b. \tag{2.28}$$

SVR has the advantage of allowing the use of SVM formalism for regression but it has the disadvantage of introducing another regularisation parameter (ϵ). In this thesis SVR has been employed in section 3.4 to compare the results of binary and *continuous* diagnosis.

2.1.5 Classification tree

Classification and regression trees (CART) are a non-parametric technique which represent an instance of decision trees. These models make predictions according to simple decision rules and recursively partition the data space [1]. In a regression scenario, the model can be described as follows: given N samples and p predictors that is (x_i, y_i) for $i = 1, ..., N$, with $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, a decision tree partitions the space into M regions $R_1, R2, ..., R_M$:

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m) \tag{2.29}$$

where c_m is the response modeled as a constant. Identifying the optimal partitioning is typically computationally infeasible, therefore it is possible to do so by employing

a greedy algorithm such that:

$$\min_{(j,s)} \left[\min_{s \in S_j} \text{cost}(x_i, y_i : x_{ij} \leq s) + \text{cost}(x_i, y_i : x_{ij} > s)\right] \tag{2.30}$$

where j is the splitting variable and s is the split point and S_j is the set of possible thresholds for the feature j. By doing so, it is possible to easily scan all the data and identify the best pair (j, s). The cost function can be chosen according to the type of analysis (e.g. the squared error in case of a regression analysis). If the analysis involves a classification, the algorithm will consider the proportion of class k in each node m:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k) \tag{2.31}$$

where R_m represents a region and N_m is the number of observations. The observations are classified to the class according to the majority class in that node $k(m) = \arg\max_k \hat{p}_{mk}$. As in the regression case, different measures can be employed such as the misclassification error, the Gini index or the cross-entropy [23]. Once the best split has been found, the process is repeated until a specific criterion is satisfied. Typically the tree can be grown until a minimum node size is reached, but it can also be pruned before it reaches the final leaves. The idea behind the pruning is to prevent overfitting by stopping the growing if the error reduction is not enough to justify the complexity of the tree.

The main advantage of decision tree is their high interpretability: an example can be seen in Fig. 2.3, where according to each rule the decisions are make (one moves left when the condition is true) until the classification is performed. Moreover, trees can handle different data input requiring less preprocessing than other machine learning techniques (there is no need to normalise the data or performing one-hot encoding to deal with categorical variables), they are robust to outliers and automatically perform feature selection. However, they are typically weaker than other algorithms and, more importantly, tend to overfit the data (especially when not pruned) and to have high variance since they are very sensible to changes in the input data [1]. Nonetheless, properly managed classification trees have been widely

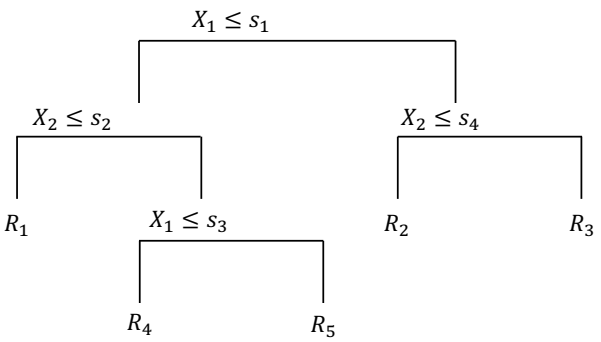


FIGURE 2.3: An example of decision tree with two inputs.

employed in this thesis since their high interpretability is often very useful in clinical scenarios.

2.1.6 Random forest

Random forest (RF) is a machine learning method that uses the bagging technique on classification and regression trees [23]. Bagging is an ensemble method which combines the prediction of various learning algorithms to get a more accurate prediction than the single model. The main idea is to average various noisy and unbiased models (such as classification trees) to reduce their variance:

$$f_{rf}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \tag{2.32}$$

where $f_{rf}(x)$ is the prediction function, $f_t(x)$ is the t^{th} tree and T is the total number of trees. However, the bagging approach can lead to correlated predictors limiting the variance reduction. For this reason, random forest also select a random subset of features (e.g. the square root of the number of features) to build each tree so that each tree is independent from the others. Despite the use of random forest leads to a less interpretable model than simple decision trees, it is more powerful since it allows to make each tree grow to its maximum getting low bias and high variance and then average the results reducing the variance. Moreover, it is also possible to obtain a list of the most important features: at each split in each tree the importance of the splitting variables is the improvement in terms of cost of the split. This values is accumulated over all the trees for each feature.

Another powerful ensemble method, the gradient boosting [23] (such as XGBoost [34]), can also be employed with decision trees in comparison with random forest. In this thesis, however, random forest has been chosen as preferred ensemble method to deal with decision trees limitations, mainly because of its good prediction performance, robustness to noise, tuning simplicity and its ability in handling highly non-linear biological data [35].

2.2 Unsupervised methods

The main aim of unsupervised machine learning is to identify structures in the data, therefore it does not need the outcome of the analysis to be defined a priori as in the supervised case: the models have the form of $p(\mathbf{x}_i|\theta)$ instead of $p(y_i|\mathbf{x}_i, \theta)$ where p indicates the probability distributions, $\mathbf{x}_i$ is the matrix of available features and y_i is the label for the i-th sample. One of the most popular example of unsupervised learning is clustering that is the process of identifying coherent groups among samples. Other algorithms are used to perform dimensionality reduction of high dimensional data to both improve the performance of a supervised algorithm or simply

for data visualisation. The unsupervised algorithms employed in this thesis are the following:

- t-distributed Stochastic Neighbour Embedding (t-SNE).
- K-means.
- K-medoids.
- Spectral clustering.
- Principal path.

2.2.1 t-SNE

t-SNE (t-Distributed Stochastic Neighbour Embedding) is a dimension reduction technique which allows to visualise high dimensional data in a two or three dimensional space and has been created as a variation of SNE (Stochastic Neighbour Embedding) [36]. SNE calculates the similarity between data points using the Euclidean distance. Given two data points x_i and x_j in the original data space, one can define $p_{j|i}$ as the probability that x_j would be the neighbour of x_i according to a Gaussian centered at x_i. At the same time, one can define $q_{j|i}$ as the same probability for the low-dimensional counterparts y_i and y_j. SNE minimises the Kullback-Leibler divergence over all the data points; this is defined by:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}} \tag{2.33}$$

where P_i and Q_i represent the conditional probability distributions given all data points x_i and projected points y_i respectively. The Kullback-Leibler divergence is not symmetric, thus different types of errors are not equally weighted. t-SNE uses a symmetric version of input points probabilties and a Student-t distribution instead of Gaussian to compute the similarity between the points, that is p_{ij} is used instead of $p_{j|i}$ where $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$ and:

$$p_{i|j} = \frac{\exp(-\left\|x_i - x_j\right\|^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(-\left\|x_k - x_l\right\|^2 / 2\sigma^2)} \tag{2.34}$$

Hence one can define a new divergence measure cost function:

$$C = \sum_i KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{2.35}$$

where $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji} \quad \forall i, j$ this last quantity is defined by:

$$q_{ij} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq l} (1 + \left\|y_k - y_l\right\|^2)^{-1}} \tag{2.36}$$

where q_{ij} is using a Student-t distribution with a single degree of freedom. t-SNE is very helpful for visualising data but it suffers of some weaknesses such as its sensitivity to the curse of intrinsic dimensionality of the data (the local linearity assumption made by using Euclidean distance might be violated) and the non-convexity of the cost function. The last one in particular implies that the algorithm might fall into a local minimum with a need to set different optimisation parameters which can bring different results according to their values. Of these, the most important one is the so called perplexity which can be seen as a smooth measure of the number of neighbours.

Due to the high-dimensionality of the analysed data set, t-SNE has been widely used as visualisation technique, especially for the unsupervised analysis. However, one should remark that t-SNE projection is just a rough approximation of the true data space and the two-dimensional representation might not necessarily correspond to the original one, still they can provide some qualitative insights on the data.

2.2.2 K-means

K-means is one of the most popular clustering algorithm; it can use any proper metric or distance and often employs the Euclidean distance as similarity measure [37]. The objective of k-means is to minimise the dissimilarity of observations from the cluster mean defined by the points in that cluster:

$$\min_{\mathbf{C}} \quad \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} \|x_i - m_i\|^2 \tag{2.37}$$

where m_i is the mean of the k-th cluster C. The number of clusters is often decided a priori and represents a hyper-parameter. The within-cluster sum of squares is often referred to as inertia. Since computing all the possible combinations of samples and means would be infeasible for large data sets, typically in a first step the points are assigned randomly to the k clusters. Then, for each cluster the mean (the centroid) is computed and the points are assigned to the closest cluster mean. This procedure is repeated until the assignments do not change. This algorithm is a typical instance of Expectation Maximization method [38]. However, k-means algorithm finds a local minimum, thus results highly depend on the starting random assignation and the procedure should be repeated to identify the solution for which the objective function is smallest [23]. A possible solution to this issue is to resort to other centroids initialisation methods such as *kmeans++* which initialise the centroids to be distant from each other, generally achieving better results than random initialisation [39]. In this thesis, k-means has been largely employed as clustering algorithm in order to identify possible coherent groups of patients.

2.2.2.1 Elbow criterion

The elbow criterion is a heuristic that can be used to identify a reasonable number of clusters [40]. It is based on the idea to plot the clustering cost function (the sum of squared errors) or another variability measure, as a function of the number of clusters. The *optimal* number of clusters is identified as the value corresponding to the elbow of the curve. In turn this elbow is often identified via percentage change of the derivative. The elbow criterion has been used to identify the number of clusters when using k-means and k-medoids (see next paragraph).

2.2.3 K-medoids

The K-medoids is another clustering algorithm. It shares the same idea of k-means but it has two main differences. First the cluster centres (the medoids) are actual points, i.e. they are not calculated as means as in the case of k-means centroids. Moreover, k-medoids is found to be often more resilient to outliers [41]. k-medoids approximately minimises the same k-means cost but with the centers being restrained to be samples. As in the case of k-means, it is possible to employ techniques such as the elbow criterion to identify the optimal number of clusters. In this thesis, k-medoids has been employed to group features in the attempt to increase the feature selection stability as later described in section 3.4.7. In particular, the Pearson correlation has been used as distance function to cluster features and to remove redundancies before the supervised learning (and further feature selection) step.

2.2.4 Spectral clustering

Spectral clustering is another and rather powerful clustering technique. It can be seen as generalisation of standard clustering methods that can be used even when the clusters shapes are non-convex [23]. Given a set of data points $x_1, ..., x_n$ and some similarity $s_{ij} \geq 0$ between the data points x_i and x_j, it is possible to represent these similarities in form of a similarity graph $G = (V, E)$ where the vertices v_i represent the data points which are connected if the similarity is positive or larger than some threshold. It is thus possible to group the samples such that the edges between different groups have low weight while the samples within the same group have high weight [42]. The similarity measure can be expressed in different ways but the two most popular ones are based on the RBF kernel and the K-nearest neighbours graph. Given the similarity matrix, one can define the adjacency matrix $\mathbf{W} = \{w_{ij}\}$ as the matrix of edge weights: if $w_{ij} = 0$, than the vertices v_i and v_j are not connected. At the same time one can define the degree matrix $\mathbf{D}$ as a diagonal matrix with the degrees on the diagonal. The degree of a vertex v_i is defined as:

$$d_i = \sum_{j=1}^{n} w_{ij}. \tag{2.38}$$

Given **W** and **D** one can define the graph Laplacian matrix as:

$$\mathbf{L} = \mathbf{D} - \mathbf{W}. \tag{2.39}$$

This is the key ingredient of spectral clustering. Indeed computing m eigenvectors corresponding to the m smallest eigenvalues of **L** gives access to a usually well behaved space where performing a common clustering algorithm is particularly easy. In a sense spectral clustering is more a projection method than a clustering method per se. Indeed in this space, by using a clustering algorithm such as k-means, it is possible to group the rows of the eigenvectors matrix and obtain a clustering on the original samples. In spectral clustering one needs to select the type of similarity graph and its related parameters (the k number of neighbours or the value of σ in the kernel case), the number of eigenvectors to extract and more importantly, the number of clusters.

2.2.4.1 Self-tuning spectral clustering

One way to address some of these issues of spectral clustering is to employ the so-called self-tuning spectral clustering proposed in [43]. In this variation Authors propose to calculate a local scaling parameter σ_i for each sample s_i. Given the distance between data points s_i and s_j defined as $d(s_i, s_j)/\sigma_i$ and its converse $d(s_j, s_i)/\sigma_j$, one can define the affinity between two points as:

$$\hat{A}_{ij} = \exp\left(\frac{-d^2(s_i, s_j)}{\sigma_i \sigma_j}\right). \tag{2.40}$$

In particular, by studying the local neighbourhood of point s_i one can define:

$$\sigma_i = d(s_i, s_K) \tag{2.41}$$

where s_K is the K-th neighbourhood of point s_i typically set to $K = 7$ without further model selection. Once the value of σ has been defined, it is also possible to identify the optimal number of clusters. This can be done by observing the structure of eigenvectors. Given the matrix Z resulting by rotating the eigenvector matrix and $M_i = \max_j Z_{ij}$, one can define the following cost function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2} \tag{2.42}$$

where C is the number of groups. By minimising (2.42) over all possible rotations, one can automatically identify the number of clusters as the one associated with the minimal cost.

In this thesis, spectral clustering and its self-tuning variant have been employed to perform unsupervised analysis together with k-means.

2.2.5 Principal path

The Principal Path (PP) is a novel learning approach inspired by the minimum free energy path concept in statistical mechanics [22]. It can be considered a local version of the principal curve, or an out-of-sample extension of Dijsktra shortest path with a proper metric. It can be used to study the evolution of a phenomenon over a manifold such as the one between healthy and sick patients. Given a starting and an ending point it builds a morphing curve that connects the two objects. In order to build the path, the algorithm tries to simultaneously get a smooth path and try to maximize the probability of laying inside the high density regions of the data distribution. Formally the principal path minimises the following functional:

$$\min_{W,u} \sum_{i=1}^{N} \sum_{j=1}^{N_c} \|\phi(x_i) - w_j\|^2 \delta(u_{i,j}) + s \sum_{i=0}^{N_c} \|w_{i+1} - w_i\|^2 \tag{2.43}$$

where N is the number of samples, N^c is the number of waypoints (clusters), $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ is the (possibly non linear) mapping of the d-dimensional input space, x_i is a sample of the $N \times d$ matrix X arranged in a row-wise fashion, w_j is a waypoint of the $N \times d'$ matrix W arranged in row-wise fashion, and $\delta(u_i, j)$ is the Kronecker delta where u_i are cluster memberships. This function represents a regularised version of the k-means clustering algorithm, where the hyperparameter s regulates the trade-off between smoothness and the data fitting of the generated path: the first and last cluster/waypoint are kept fixed while the others are evolved according to the regularisation term. The principal path has been successfully employed in many applications, starting from the study of free energy of a drug binding to a protein [44], the study of the sentic paths between concepts [45], to the study of evolution of music and visual art [46]. Fig.2.4 shows an example of the principal path first application in understanding chemical processes (bound/unbound protein/ligand complex). In this thesis, it has been used to investigate the progression of a disease. Given two classes (e.g. healthy and diseased) the principal path tries to give a surrogate of the time evolution of a disease trying to capture the changing between the two states.

2.3 Feature Selection Techniques

Feature selection is the process of identifying a subset of relevant features to be used in the model prediction. Feature selection can improve the prediction performance and more importantly enhance the model interpretability [47]. Some feature selection methods directly return a subset of the most important features (e.g. lasso) while others simply identify a ranking among the features according to their prediction relevance (e.g. random forest). Some machine learning methods can directly perform feature selection: this is the case of lasso, elastic net and random forest. In other cases, one can exploit other techniques in order to achieve similar results. In

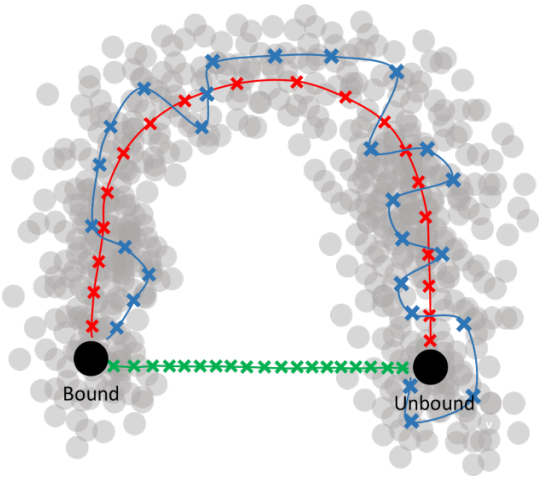


FIGURE 2.4: A principal path in red connecting a start and ending point, a straight line in green and in blue the trivial connection of clusters from k-means [44]. The red line tries to move following high probability regions (and hence minimum free energy regions).

this section some of the feature selection methods employed in this thesis will be described.

2.3.1 Recursive feature elimination

Recursive feature elimination (RFE) is a feature selection strategy that deletes the worst-performing features of a particular model in a stepwise fashion [48]. The process is executed recursively until the best subset of features is achieved. RFE works following these steps:

1. Train a classifier.
2. Define a ranking for all the features (possibly taking into account the model weights w_i).
3. Remove the last ranked feature and restart from the first step.

In some cases it could be useful to remove more than one feature at a time in order to improve the computational performance. Linear SVM is a perfect algorithm to use in combination with this feature selection technique and has been employed in this thesis in order to identify a feature subset whenever it was useful to reduce the number of features.

2.3.2 Unsupervised feature selection

Here a simple approach that has been used in several activities in this thesis is reported. When performing unsupervised analysis to identify unknown groups of samples, one is able to subdivide data points according to their similarity. However, one typically does not know *why* the samples are similar and thus grouped together as clustering does not contribute with an explanation. Since one of the aims

of this thesis is to identify the most important features behind a classification rule, one would like to do so even in the case of unsupervised learning. Here, one wants to find alternative groups of samples and understand why they are formed, rather than identifying the patients based on the already available labels (e.g. the patients diagnosis). As a matter of fact, when implementing any of the clustering algorithms already described, one can obtain a new labelling for the samples according to fact they belong to a specific group. These new labels can be used with any classification algorithm that can perform feature selection to identify which variables define the samples grouping. More specifically, a classification tree has been employed to classify the new clustering-induced classes and therefore identify the most relevant features. Hence, the used classifier is only instrumental to the feature selection activity as the classification accuracy of clusters is obviously high by construction. This choice has been made due to the high interpretability of the decision tree result. Moreover, this approach also allows to identify the threshold values used for the grouping process. These two aspects are of notable importance when dealing with clinical scenarios where approaches based on thresholds are typically used in diagnosis processes.

2.4 Pre-processing methods

Data pre-processing methods refer to all those techniques that can be implemented before the actual analysis in order to make it possible or improve its performance. In this section the pre-processing techniques employed in this thesis will be described. These include the missing data handling and the manipulation of both continuous and categorical features.

2.4.1 Imputation of missing values

In some cases, a data set may contain missing values and therefore some information may not be available for all the samples. Most machine learning algorithms cannot handle missing data and therefore it is necessary to pre-process the data to make it readable by the algorithms. The simplest procedure to handle missing values is to simply remove all the columns or rows in which the data are not present. Obviously, this strategy is not always suggested since it may cause a huge loss of information if the number of missing values is large. In order to avoid this, one can impute the missing data and infer them using the other available information. This can be done by employing different approaches such as replacing the empty values with the mean, the median or the mode of each feature. In other cases the values can be imputed as a function of the other features [49]. In this thesis, however, the number of missing values (when present) was very large and data imputation techniques could have introduced bias in the analysis [50], [51]. For this reason, all the missing values have been removed from the data sets. Since the sample size was very limited in the beginning, in order to preserve the maximum number of samples as possible,

the features with a high number of missing values (e.g. 90%) have been removed as a first step, afterwards only the samples with any missing values for the remaining features have been dropped.

2.4.2 Features encoding

In many cases, not all the features are expressed as numeric values: this is the case of categorical variables. Some algorithms, such as classification trees, can easily deal with categorical features while others require that all the features are expressed as numerical values. In order to convert the features to numerical values, a procedure called one-hot encoding has been applied for all the analysis where categorical features were present. One-hot encoding simply creates a binary column for each category. Obviously, a feature with only two categories will be transformed to a single binary feature where 0 and 1 represent the two classes. However, if the number of categories is greater than two, a new number of variables will be created according to the number of categories. On the other hand, if an ordinal relationship among the classes exist, using one-hot encoding would not consider this order. Therefore, an ordinal encoding should be implemented by transforming the original variable to a numeric feature of integers (e.g. from 1 to the number of classes of the considered feature).

2.4.3 Features scaling

Contrary to categorical features, continuous variables are totally readable by any machine learning algorithm. However, in some cases it might be needed to transform the original data in order to improve the models performance or simply reduce the computing time (this is especially true when the input features are not expressed in the same scale) [52]. In this thesis, all the features where scaled to make them lie between zero and one by applying the following transformation:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.44}$$

where X is the single feature, while X_{max} and X_{min} are the corresponding maximum and minimum value of the considered feature.

2.5 Performance metrics

In this paragraph, the performance metrics employed to evaluate the results of the analysis are described. These include those related to both classification and regression (for the progressive diagnosis case):

- Accuracy and balanced accuracy;
- ROC and AUC;

- MSE and R^2.

2.5.1 Accuracy and balanced accuracy

Accuracy is the easiest way to assess the performance of a classification algorithm. It is calculated as the the proportion of number of correct predictions made by the model [53]. Formally:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2.45}$$

where $\hat{y}_i$ is the predicted value of the outcome y and $I(y_i \neq \hat{y}_i)$ is an indicator variable. Accuracy values are always comprised in the range $[0, 1]$ with 1 being the best. Despite its popularity, accuracy measure is not very suited for unbalanced data sets. In these cases, the algorithm could reach high value of accuracy even by classifying all the samples as the majority class. To mitigate this issue, it is possible to employ a slightly different measure which is the balanced accuracy. Both accuracy and balanced accuracy can be used with binary or multiclass classification. Given two classes, the "positive" (P) and the "negative" (N) ones, one can define the true positive samples as the number of positive samples correctly identified as positive. Therefore the true positive rate (TPR) can be calculated as:

$$TPR = \frac{TP}{P} \tag{2.46}$$

which is also known as *recall* or *sensitivity*. In the same way, one can define the true negative rate as (TNR):

$$TNR = \frac{TN}{N} \tag{2.47}$$

which is also known as *specificity*. Given these values one can finally calculate the balanced accuracy as:

$$BA = \frac{TPR + TNR}{2} = \frac{Sensitivity + Specificity}{2}. \tag{2.48}$$

The balanced accuracy is more reliable in case of unbalanced data since it takes into account the relative per class accuracy. For this reason it has been employed in comparison to accuracy when dealing with unbalanced data. By using the values of TP, TN and the false positive samples (FP), other performances metrics can also be calculated such as the *precision* and the F_1 *score*:

$$Precision = \frac{TP}{TP + FP} \tag{2.49}$$

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{2.50}$$

2.5.2 ROC and AUC

The Receiver Operating Characteristic (ROC) Curve is a graphical representation used to evaluate the performance of a classification algorithm [54]. It is obtained by plotting the false positive rate (FPR) against the true positive rate (TPR) at different thresholds of the classification function where the false positive rate can be calculated as:

$$FPR = 1 - TNR = 1 - Specificity. \tag{2.51}$$

Since each of these values can be in in the range $[0, 1]$, the total area of the plot is 1.0. The plot can be divided in two equal parts by a diagonal line of equation $y = x$ representing the random guess: if the curve lies in the upper part of the plot, the algorithm performs better than the random choice whilst it performs worse if it lies under the diagonal line, therefore the point with coordinate $(0, 1)$ represents the perfect classification. The curve is plotted by taking into consideration the probability or score obtained as result of the classification which represents the degree a sample belongs to a specific class. By setting a threshold it is possible to assign each sample to a binary class according to their score or probability value. For each possible threshold it is thus possible to represent a point of the ROC curve given by the values of FPR and TPR for that specific threshold. The ROC curve is therefore produced by observing how these values vary according to different thresholds. Since the curve itself is not an objective way to evaluate an algorithm (except in the extreme cases), one can calculate the area under the curve (AUC) to have a clear number defining the performance of the classifier [55]. The AUC can also be seen as the probability that the algorithm will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Because the plot has an unit square area, a random guess would produce an area of 0.5 while a perfect classifier will score an area of 1.0. The use of AUC is also useful to compare different algorithms performances, and it is more reliable to evaluate performances in those cases where simple accuracy might give misleading results (e.g. with unbalanced data). The ROC curve and AUC have been used in this thesis to evaluate the performances of many classifiers as well as accuracy and the balanced accuracy.

2.5.3 MSE and R^2

When it comes to regression tasks, accuracy and ROC curves cannot be used to evaluate the performance of the analysis since regression involve the use of continuous values. Given the true value of the analysis y_i and its associated predicted value $\hat{y}_i$, one can define the mean squared error (MSE) as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{2.52}$$

The MSE will be small when the prediction is very close to the true value and it will be higher when the prediction is less correct [53]. Since MSE is expressed in the unit of the data, it is not always easy to interpret, especially when comparing different models. Nevertheless, it is possible to calculate its normalised version known as coefficient of determination or simply R^2 which is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} \tag{2.53}$$

where $\bar{y}$ is defined as the sample average of y such that $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The numerator is also known as residual sum of squares (RSS) whose complement is the explained sum of squares (ESS), and the denominator is the total sum of squares (TSS). R^2 can thus be seen as the portion of variance explained by the model, and since it is normalised its value belongs to the interval $[0, 1]$ with 1 being the best possible value. MSE and in particular R^2 have been used to evaluate the performance of regression algorithms when considering the continuous diagnosis case (see section 3.4.7).

2.6 Feature selection stability indices

The stability of feature selection can be defined as the sensitivity of selected features to small perturbations in the training set [56]. In other words, if the features are robust to changes, an algorithm would select them even if the underlying data is slightly different. In order to understand the level of stability of an algorithm feature selection, it is necessary to define an objective way to measure it. In recent years, different indices have been proposed in the literature as the stability of feature selection for machine learning algorithm is still an open issue especially when working with omics and biological data, particularly in the small sample regime. Here the high redundancy and correlation between features can further increase the instability not only between different models from different algorithms but also within the method itself [57].

According to [58], given M subsets of features, each of size d, one can represent them as a matrix Z of size $M \times d$. It is thus possible to calculate the frequency $\hat{p}_j$ of each feature as the mean of the j^{th} column of Z:

$$\hat{p}_j = \frac{1}{M}\sum_{i=1}^{M} z_{i,j} \tag{2.54}$$

This is the starting information to evaluate the stability of each feature subset by using different stability measures which have been proposed in recent years. Some of these are simpler than others, such as the Jaccard Index [58], while others are more elaborate like the Adjusted Stability Measure [59] or the Sorensen-Dice Index [60]. Authors in [58] stated that there rare five desirable properties that a stability index should satisfy in order to be considered reliable, but none of the cited indices satisfy all of them simultaneously:

- **Strict monotonicity**: The stability index should be a strictly decreasing function of the sample variance of the variables.
- **Fully defined**: The stability index should be defined for any collection of feature sets.
- **Maximum stability**: The stability index should reach its maximum value if and only if all the feature sets are identical.
- **Bounds**: The stability index should have an upper and a lower bound not dependent on the number of features.
- **Correction by chance**: The value of the stability index should be constant whenever the features have been drawn at random.

Nonetheless, the same Authors also proposed a new index satisfying all these properties. Given a matrix Z of size $M \times d$, the stability index can be expressed as:

$$\hat{\Phi}(Z) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}(1 - \frac{\bar{k}}{d})} \tag{2.55}$$

where $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ is the unbiased variance of feature f, $\bar{k} = \frac{1}{M}\sum_{i=1}^{M}\sum_{f=1}^{d} z_{i,f}$ is the average number of features selected within the M feature subsets.

This kind of index should be used for proper feature selection methods, i.e. methods returning a specific subset of features such as lasso or elastic net. This index can also be used for other methods such as SVM and random forest, although it requires setting a specific threshold to convert the ranking into a subset. Alternatively, one can use the index reported in [61] which is specific for ranking feature selection methods.

Given a matrix R of size $M \times d$, where r_i is the feature ranking on the i^{th} subset, the stability index can be defined as:

$$\hat{\Phi}(R) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{V_r} \tag{2.56}$$

where s_f^2 is the unbiased sample variance of the rank of the f^{th} feature and $V_r = \frac{d^2-1}{12}$.

However, one should consider that in presence of highly correlated or redundant features (such as the case under consideration) these indices might underestimate the stability. For this reason, an index taking into account this aspect has been proposed [57]:

$$\hat{\Phi}_C(Z) = 1 - \frac{tr(\mathbf{CS})}{tr(\mathbf{C\Sigma^0})} \tag{2.57}$$

where $\mathbf{S} = \hat{Cov}(Z_f, Z_{f'}) = \frac{M}{M-1}(\hat{p}_{f,f'} - \hat{p}_f\hat{p}_{f'})$, $\mathbf{C}$ is a matrix describing the feature relationships, and $\mathbf{\Sigma^0}$ is a matrix normalising the measure. This index can be seen as

a generalisation of (2.55): when the matrix $\mathbf{C}$ is equal to the identity matrix, the two indices are the same.

Due to their properties and advantages, these indices have been employed when analysing the Oxford Street II data set (see section 3.4.8). In particular, (2.55) has been used for filter feature selection methods such as lasso and elastic net and (2.56) has been used for ranking feature selection techniques (random forest and SVM). Finally, (2.57) has been used for all the methods in order to take into account of the effective stability (for ranking feature selection models one has to set a threshold to convert the ranking into a subset).

Chapter 3

Results

In this chapter, the analysis, the data sets employed and the related results will be shown, as well as the methodological contributions produced. After describing the approaches applied to clinical data sets and their related challenges, a more in-depth analysis regarding an omics data set, where different issues will be addressed and discussed, will be described. The considered data sets are the following:

- Clinical data sets:
  - Dyslexia data set.
  - Autism data set.
  - Osteogenesis imperfecta I data set.
  - Osteogenesis imperfecta II data set.
- Omics data set:
  - Oxford Street II data set.

3.1 Dyslexia

Dyslexia is defined as a learning disability with neurobiological origin. It consists in difficulties in word recognition and poor spelling and decoding abilities [62]. This may lead to problems in reading comprehension, writing and spelling and can impede the growth of vocabulary and background knowledge. This disorder affects around 7% of school-aged children and it usually persists in adulthood affecting career, academic performance and quality of life in general [63]. Learning is a complex process, hence it is difficult to quantify and identify the associated disorders. Learning difficulties may affect understanding, acquisition, retention and organisation of information, thus they can be diagnosed through IQ measures, assessments of difficulties in reading, oral and written speech [64]. Reliable diagnoses can be determined only behaviourally after some years of education: in this period, the difference between normal cognitive and reading abilities is more observable. As alternative, many researches leveraged the identification of biomarkers of dyslexia by using functional (and non functional) magnetic resonance imaging (fMRI, MRI) [65]. Moreover, thanks to functional imaging, it is possible to analyse brain functions

during the performance of cognitive tasks. When performing these tasks, specific brain regions are activated, thus the changes in the neural activity can be measured with fMRI or magnetoencephalography (MEG). Evidence has shown a failure of left hemisphere posterior brain systems during reading in adult dyslexic readers [62].

Given the large number of factors related to the learning difficulties identification, it is not trivial to provide a reliable diagnosis process. Therefore, it is possible to implement an automatic way to provide diagnosis support by using machine learning techniques [64]. Different studies have been conducted in learning disabilities areas by employing various techniques or involving diverse kinds of applications.

When it comes to the methodology, the algorithms employed vary from decision trees to SVMs or other custom algorithms specifically designed to deal with the case under study [66]. The work in [67] made use of different classification algorithms to predict reading disabilities on a data set of 356 first graders students with AUC higher than 90% using SVM. In [64], Authors employed PCA and Self Organising Maps (SOM) to successfully identify learning difficulties on a sample of 134 children. Authors in [68] implemented a K-nearest neighbours classifier (KNN) to classify children with no dyslexia and those who seem to be dyslexic in spelling and reading by using 857 school children scores in different tests; they also validated the results with a human domain expert. Other approaches involve the use of imaging as in the case of [65] where SVM was employed to distinguish 21 subjects with dyslexia from 21 subjects without dyslexia. Although the Authors did not achieve a high value of accuracy, they concluded that not all dyslexia symptoms are present at the same cortical level of organisation and thus that dyslexia is not a uniform syndrome. On the other hand, Authors in [63] were able to achieve higher accuracy values with a linear support vector machine on MRI features related to a sample of 28 dyslexic children and 33 controls.

Here, the results of an analysis involving similar approaches as those shown in the literature will be shown, highlighting different problems and issues related to this application field.

3.1.1 The data set

The data set employed for this analysis was collected by Professor Stefano Vicari's research group at the Bambino Gesù Children's Hospital in Rome, Italy. The data set included 1321 patients affected by different kind of disorders such as language disorders, intellectual and learning disabilities, and many others. Many possible diagnoses such as intellectual disabilities, attention-deficit/hyperactive disorder (ADHD), learning disorders including dyslexia, dyscalculia and dysgraphia are present for each patient. On the other hand, the data set is composed of 256 features including general information such as age, education level, familiarity with the disorder and other more specific variables like IQ test scores, visual motor integration (VMI) test results, Child Behavior Checklist (CBCL) results and other test scores.

3.1.2 Results

The main goal of this analysis was to classify patients according to their diagnosis. In order to do this, different classification algorithms have been employed. At the same time an unsupervised approach has been implemented to identify possible alternative categories among patients. Despite the apparent wide dimensionality of the data set, the data suffered of a very high incidence of missing values and many variables and/or subjects had to be dropped from the analysis. More specifically, the features with more than 90% of missing values have been dropped, afterwards only the samples with any missing values for the remaining features have been removed in order to minimise the data loss in both terms of samples and variables. Nevertheless, this drastically reduced the number of available samples and features consequentially limiting many machine learning approaches such as the use of deep neural networks. Due to the large amount of missing values, data imputation has not been considered a safe choice since most of the samples would have got imputed values adding a bias to the analysis. Therefore, the final data set only included 399 patients and 12 variables. Fig. 3.1 shows the data set missing value distribution: from the plot it is clear that most of the features are available for only few samples and they have to be dropped in order to preserve the sample size. The features employed for the analysis and their related type are reported in Table 3.1. Here, in addition to age, education level and intelligence quotient (IQ), the CBCL variables represent the scores related to the Child Behavior Checklist which is a questionnaire used to asses behavioural and emotional problems in children and adolescents [69]. It includes different syndrome scales such as anxious_depressed (CBCL_Ansia/Dep), social problems (CBCL_soc), withdrawn_depressed (CBCL_Chius/Dep), somatic complaints (CBCL_Lam somat) and thought problems (CBCL_Pr pens).

TABLE 3.1: List of features employed for the analysis on the dyslexia data after removing those with too many missing values.

Dyslexia Data Features List		
Feature Name	**Feature Description**	**Feature Type**
Age	Patient age	Continuous
Education level	Patient education level	Ordinal
IQ	Patient intelligence quotient	Continuous
CBCL_attiv	CBCL - Activities	Continuous
CBCL_soc	CBCL - Social relations	Continuous
CBCL_scol	CBCL - School	Continuous
CBCL_tot comp	CBCL - Total competence	Continuous
CBCL_Ansia/Dep	CBCL - Anxious/depressed	Continuous
CBCL_Chius/Dep	CBCL - Withdrawn/depressed	Continuous
CBCL_Lam somat	CBCL - Somatic complaints	Continuous
CBCL_Pr social	CBCL - Social problems	Continuous
CBCL_Pr pens	CBCL - Thought problems	Continuous

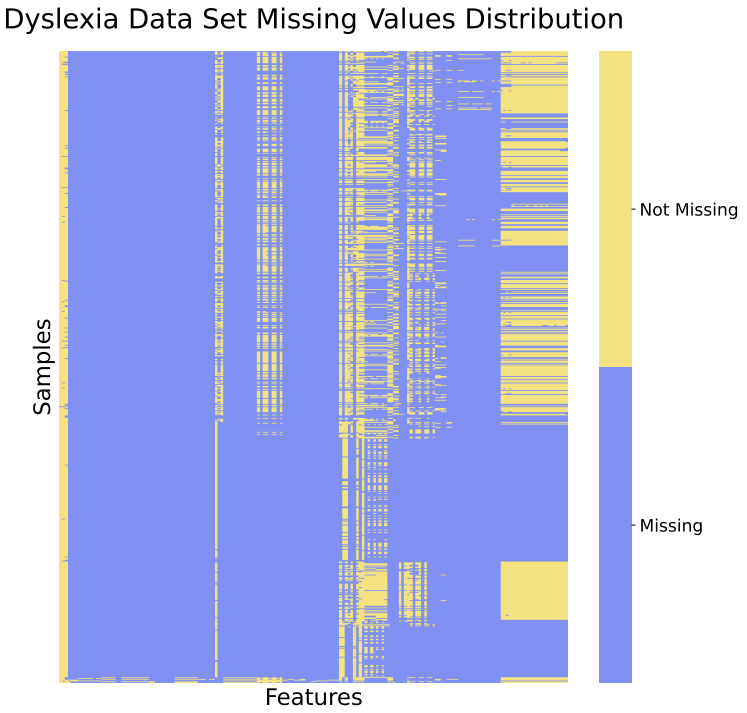


FIGURE 3.1: Missing values distribution in the dyslexia data set.

As mentioned before, the data set includes many possible diagnosis types, however many of these have a very low frequency making a classification and training task very hard to be pursued. Additionally, the data set does not include control samples, thus a possible classification is even more difficult to perform. Given the high number of classes available, and the poor multi-class classification performances, one possible solution was to transform the multi-class classification problem into a binary task. In particular, the analysis was focused on the "easiest" class to discriminate from the others, namely "profound intellectual disability", in order to evaluate whether machine learning was able to solve this classification problem. In this case the classes are unbalanced, therefore proper countermeasures were properly applied. Table 3.2 shows the results in terms of accuracy, balanced accuracy and area under the ROC curve of the classification algorithms that have been employed on the data set. These algorithms include SVM with both linear and RBF kernels, random forest and elastic net. The data set was split in two parts: 75% was used as training (299 samples) and the remaining 25% as test set (100 samples). Moreover, for each method, model selection was performed by using a 3-fold cross-validation on the training set in order to identify the best hyper-parameters. The algorithms were implemented by using scikit-learn libraries [70] for the known models applied. Furthermore, all the data was normalised in the range [0, 1]. A grid search approach has been employed in order to find the best hyper-parameters for each model whose values are shown in Table 3.3.

Even by adjusting for unbalanced classes and using balanced accuracy, the results are not reliable since the algorithms keep selecting the majority class. Fig. 3.2

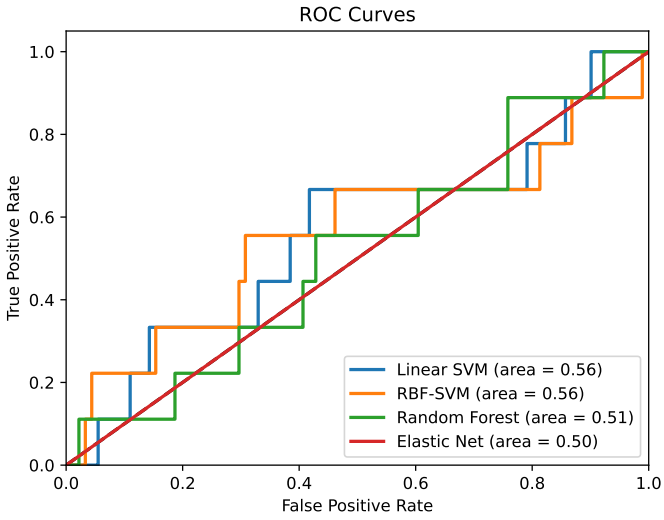


FIGURE 3.2: ROC curves for each of the classification algorithms employed for the analysis of dyslexia data.

shows the ROC curves and the related areas for the methods employed. By observing these results, one can hypothesize that there is a lack of correlation between the diagnosis and the features collected by the physicians. Since the kind of diagnosis is based on parameters and test results included in the data set, contrarily to what it is observed a classification task should not be too hard. This may point out to a possible weakness of the employed methods for this kind of data or the possible existence of some form of incoherence inside the data set. This putative incoherence may come from two sources. The first could be the incoherence among the labels (different physicians give different diagnoses). This in turn might be due to the fact that physicians do not simply rely on test and clinical information (reported in the data set) to diagnose some disease or disturb but they also rely on their knowledge and opinion that might not be the same of their colleagues. The second source of incoherence, tightly related to the first one, may be the fact that the true decision variables which lead to the diagnoses are not present in the data set, hence an AI algorithm cannot work in this case. All these considerations might be the cause of failure of the employed classification algorithms in predicting the labels.

TABLE 3.2: Supervised algorithms performances on dyslexia data. The table shows the values of area under the ROC curves, accuracy and balanced accuracy for each of the applied models.

Methods	**Test Accuracy**	**Test Balanced Accuracy**	**Test AUC**
Elastic Net	0.91	0.50	0.50
Linear SVM	0.50	0.43	0.56
RBF-SVM	0.50	0.43	0.56
Random Forest	0.91	0.50	0.51

TABLE 3.3: Supervised algorithms best hyper-parameters for dyslexia data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Elastic Net	Linear SVM	RBF-SVM	Random Forest
C: 0.0001	C: 10 000	C: 10 000	No. of trees: 100
L1 ratio: 0.1		Gamma: 0.001	Split criterion: Entropy
			Max depth: 3
			Max features: Square root

Consequently, in order to identify possible new patterns other than those described by the diagnosis, an unsupervised approach has been applied too. In this case, the aim is to find groups of patients that are similar to each other according to their clinical information. Before using a clustering algorithm, it is also possible to observe the data distribution thanks to a dimensionality reduction algorithm such as t-SNE. As shown in Fig. 3.3, one can see that the samples are mixed all together without a clear distinction of possible clusters. The two labels in the plot (1 and 0) correspond respectively to patients with profound intellectual disabilities and all the others. Although no clear cluster shapes emerge from the plot, one can identify two possible groups of patients by using the k-means algorithm. Fig. 3.4 shows the two groups of patients identified by k-means. Using other clustering algorithms such as the spectral clustering, the result remains similar, therefore there is no significant margin between the clusters. A possible consecutive step is to identify which features are responsible of this new labelling induced by the clustering. To do so, a classification tree has been employed using the k-means groups as labels. A classification tree has been used since one would like to have a clear rule describing which features are involved in the task and for which label it is possible to classify the samples as described in section 2.3.2. However, since the clusters are not particularly well separated, the resulting classification tree is not easy to interpret and it overfits the data. Fig. 3.5 clearly shows the complexity of the classification tree and how much it tends to over-fit in the attempt to include each sample in a specific group. When reading a classification tree output it is important to specify that whether the condition in the node is true one has to consider the left branch of the tree. On the other hand, if the condition is false, one has to follow the right branch. Even though the tree does not generalise well, the most important variable defining the first split is CBCL - anxious/depressed (CBCL_Ansia/Dep) with a threshold value of 64.5, followed by other CBCL variables such as CBCL - social problems (CBCL_Pr social) and CBCL - thought problems (CBCL_Pr pens.)

In conclusion, in this clinical data set (including the parameters employed by the physicians) it is not possible to get reliable results through both supervised and unsupervised methods to correlate with the diagnose. Supervised classification algorithms fail to select the minority class with low balanced accuracy and AUC values. At the same time, by using unsupervised learning it is not possible to identify clearly distinct alternative groups of patients and by using a classification tree one cannot

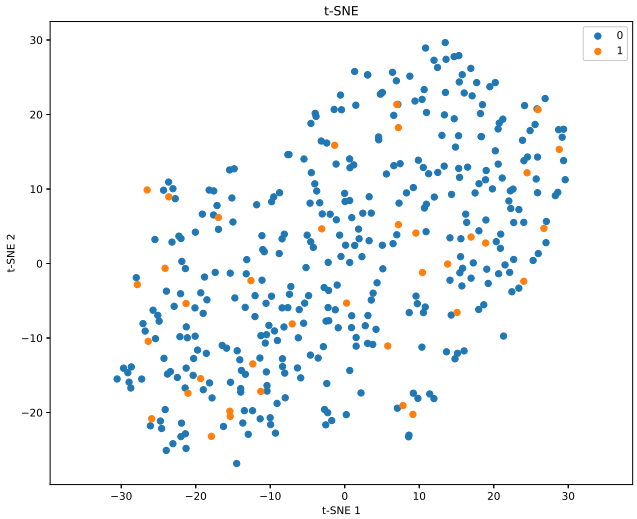


FIGURE 3.3: t-SNE two dimensional projection of dyslexia data. Ones represent patients affected by profound intellectual disabilities while zeros are all the others.

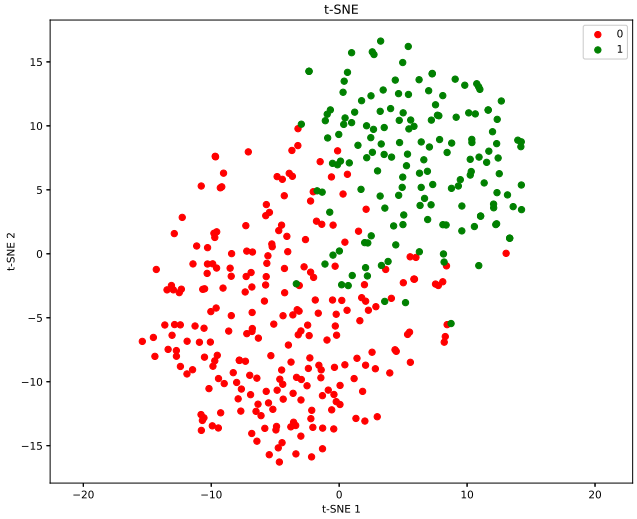


FIGURE 3.4: t-SNE two dimensional projection of clustering on dyslexia data. Zeros and ones represent the two clusters classes identified by k-means.

identify features that can describe the new groups accurately. These results suggest that despite the promising performance that machine learning can achieve in identifying learning disabilities diagnosis, the difficulties that may arise in practice are far from trivial. The presence of many missing values, the degree of consistency in assigning the labels, and the presence or absence of clinical features, heavily affect the classification performance leading to poor results that do not support the patients stratification process beyond the healthy/diseased dichotomy.

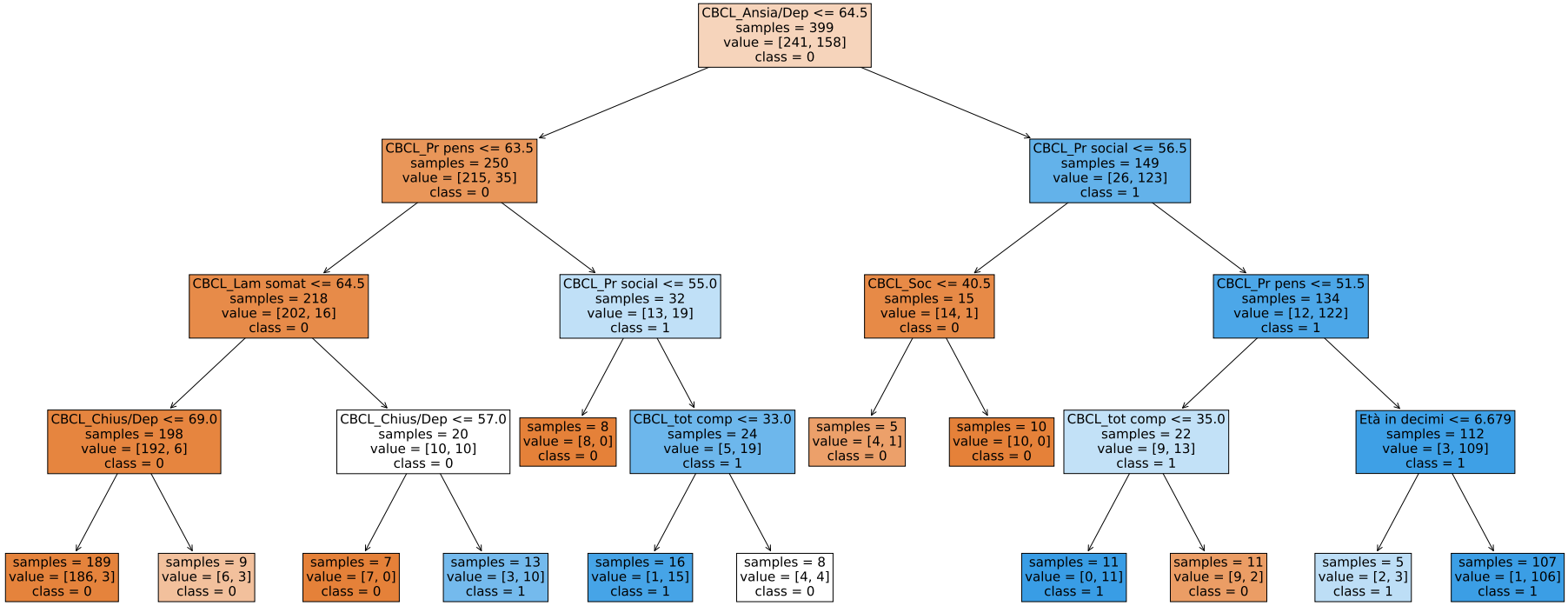


FIGURE 3.5: Classification tree rules based on the labels induced by k-means algorithm on dyslexia data. The tree rules are read as follows: if the condition expressed in the node is true, one has to consider the left branch, otherwise the condition is false and one should move on the right branch.

3.2 Autism

Autism spectrum disorder (ASD) is a neurodevelopmental disorder which includes deficits in social communication and social interaction, characterised by repetitive and restricted behaviours and interests. Due to this, children affected by ASD develop difficulties in attention, social reciprocity and in the use of both verbal and non-verbal communication. In the past, the Diagnositc and Statistical Manual of Mental Disorders 4 (DSM-IV) written by the American Psychiatric Association, subdivided autism in four different categories namely: Asperger syndrome, autistic disorder, pervasive developmental disorder not otherwise specified (PDD-NOS) and childhood disintegrative disorder (CDD). However, the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-5) now unites all these categories into the autism spectrum disorder [71]. ASD affects around 1% of the population, a number that is increasing worldwide, with a 4 times larger prevalence in males than females and its diagnosis is mainly established via clinical interviews and behaviour observations.

The symptoms of autism are easier to identify in children around 2-3 years old, thus it is crucial to diagnose ASD as early as possible since this will lead to earlier intervention. Currently, there are many different ways to identify ASD. Some of the most employed clinical methods are the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). Many other screening tools are available as reported in [72], where 37 different methods are critically analysed in order to suggest possible new research fields in the matter. These screening methods have been subdivided in three categories: for infants and children, adolescents and adults, or hybrid, but none of these seem to perform completely well in terms of the considered parameters. Despite these techniques are reliable and standardised diagnostic instruments, they have a major disadvantage: they require a lot of time to be processed and to interpret the results [71].

Nonetheless, this kind of approach may cause wrong diagnosis, preventing the possibility of intervention on children with ASD [73]. This leads to the need to develop an objective way to diagnose ASD. In this context, artificial intelligence, machine learning and imaging techniques such as magnetic resonance imaging (MRI) bring a great support as diagnostic tools for ASD. Moreover, the use of machine learning can help reducing the diagnostic time allowing patients to receive earlier intervention. Typically in these cases, the use of machine learning implies performing a classification between people affected by ASD and those who are not, while also performing feature selection in order to identify and choose the most relevant features for the classification process [71]. Many researches have been conducted in this field aiming to support the ASD diagnosis process. The approaches employed vary from the most simple ones to other more complex. For instance, Authors in [74] used a simple logistic regression analysis to classify patients coming from two data sets based on the autism quotient (AQ)-10 adult and AQ-10 adolescent screening

methods. The analysis revealed some effective features related to screenings regarding communication and social behaviour. On the other hand, more elaborated approaches involve the use of deep learning as in the case of [75] where a feed-forward artificial neural network has been employed to limit the error-prone human interpretation. Moreover, Authors in [76] were able to identify a subset of behavioural features from the ADOS Module 4, allowing to reduce the amount of information required for the analysis while achieving similar prediction performances. On the other hand, despite the positive prediction results, Authors in [77] applied various supervised learning classifiers to predict ASD showing that in most cases using the available data is not enough to achieve perfect accuracy and thus more information is needed to fill the gap between the algorithm and the clinicians opinion. Other researches involve the use of new different ways to perform classification. For instance, a technique called Rules-Machine Learning (RLM) has been proposed in [78]. This method allows automatic classification while identifying autism features and being easily interpreted by all users including parents and teachers. The Authors were able to achieve as high predictive accuracy as other well known techniques.

Moving away from typical clinical data, different family studies have shown that genetic disorders are associated with ASD, pointing to the fact that genetic factors play an important role in ASD etiology. Rather than employing clinical data, some researches have been made on the prediction of autism risk genes. Authors in [79] performed a binary classification of ASD risk genes by using brain developmental gene expression data to extract long non-coding RNA (lncRNA) using Haar wavelet transform (HWT) and the Bayes network classifier to test the model. On the other hand, Authors in [80] made use of a machine learning approach to predict ASD while identifying risk genes using features from spatiotemporal gene expression patterns in human brain and other gene variation features. This allowed the Authors to select genes with strong prior evidence as well as potential novel candidates.

Other approaches involve the use of imaging data such as those coming from MRI, structural magnetic resonance imaging (sMRI) or functional magnetic resonance imaging (fMRI). These techniques provide an objective measurement of the brain morphology and activity, allowing to investigate the neurological underpinnings of ASD; these tools are essential for clinical diagnosis. Both fMRI and sMRI data contain relevant information concerning biomarkers and features such as early circumference enlargement and volume overgrowth of the brain. The autism prediction can then being performed by using standard machine learning methods such as SVM, random forest and gradient boosting machine, while others involve the use deep learning methods like convolutional neural networks or autoeconoders [81]. These approaches are gaining more and more attention, achieving promising results as stated in [82] where a meta-analysis involving sMRI subgroups have been performed. Nonetheless, this kind of studies still have many limitations and challenges, such as the limited amount of available data and the compliance requirements for data collection, with a clear need for new studies on the matter [73], [82].

Machine learning techniques have thus been widely used in ASD research, nevertheless several issues still affect the reliability of the analysis such as limited sample size, improper sampling methods, feature redundancy and labels imbalance [71]. Some of these issues also affect the here proposed analyses as described in the following sections.

3.2.1 The data set

The data set employed for this analysis was collected by Professor Stefano Vicari's research group at the Bambino Gesù Children's Hospital in Rome, Italy. This data set is composed of 1386 samples and 724 features. The patients are labelled based on various mental disorders such as autism spectrum disorder (ASD), Asperger syndrome, and other language and cognitive disorders. As in the case of the dyslexia data set, there are no control samples, but only people affected by the disorders, placing this work in a precision medicine regime. Each patient has been observed at eight consecutive periods of time, ranging from T0 to T7 included. For each period of time, various test scores and results have been collected including cognitive tests, Kiddie Schedule for Affective Disorders and Schizophrenia (K-SADS) results, Child Behaviour Checklist, *primo vocabolario del bambino* (child's first vocabulary (PVB) forms), Autism Diagnostic Observation Schedule (ADOS), and other general information about the patients such as age, sex, parents and other relatives information. For each period of time the diagnosis related to that time span is also available. However, most of this data is not available for all the patients since only few of them have information related to all the periods of time. In addition to this, there are many other missing values even when considering the first period of time (T0). As in the case of the dyslexia data set, the missing values imputation had to be avoided in order to prevent the introduction of an arbitrary bias in the analysis.

3.2.2 Results

The first aim of this analysis was to limit the loss of data (the number of features and samples) due to the missing values. In particular, firstly the features with over than 90% of missing values have been dropped and then the samples with any missing values for the remaining features have been removed. Fig. 3.6 shows the data set missing value distribution. As shown in the figure, the amount of missing values is very high and it tends to increase by moving to the right side of the plot. In fact, as the observation time increases from T0 to T7, the missing values increase as well. Moreover, most of the information available at T>1 is not very useful for the analysis since it is related to general patients characteristics such as age and education.

After having dealt with missing values, both supervised and unsupervised analysis have been performed on different subsets of the data set. For each analysis the subset was split in two parts: 50% was used as training set and the remaining 50% was used as test set. Given the reduced sample size, one wants to have

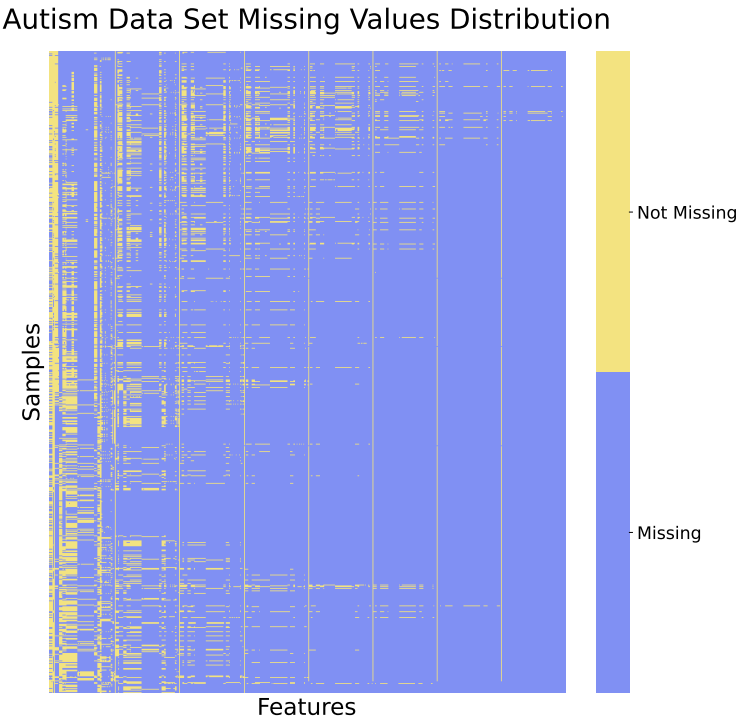


FIGURE 3.6: Missing values distribution in the autism data set.

enough samples not only in the training set but also in the test set. Therefore an increased test size has been chosen in comparison with the one selected in the previous section. A 3-fold cross-validation has also been applied on the training set for hyper-parameters tuning using a grid search approach. All the tested models were implemented through scikit-learn libraries [70]. All the data was normalised in the range [0, 1] and all categorical variables were treated using one-hot encoding. According to the type of analysis different subsets of data were selected to maximise the number of features/samples. The subsets were different from each other by the number of features and samples. Five different subsets have been analysed:

- T0 data with T0 diagnosis as label.
- T0 data with T1 diagnosis as label.
- T0 data with T0 PVB form (child's first vocabulary) as label.
- T0 data with T1 diagnosis as label (excluding cognitive test).
- T0 selected features with T1 diagnosis as label.

The list of features employed in the analysis after having removed the missing values is reported in Table 3.4. The autism diagnostic interview (ADI) is an interview, usually undertaken with the parents of the child, designed to get a information about the functioning of the patients. The Autism Diagnostic Observation Schedule (ADOS) is a standardized semi-structured observational play and activity assessment of the child [83] while the parenting stress index (PSI) is a measure designed to

TABLE 3.4: List of features employed for the analysis on the autism data after removing those with too many missing values.

Autism Data Features List		
Feature Name	**Feature Description**	**Feature Type**
Decimal Age	Patient age	Continuous
ADI_A	Autism diagnostic interview - A	Continuous
ADI_B	Autism diagnostic interview - B	Continuous
ADI_C	Autism diagnostic interview - C	Continuous
ADI_D	Autism diagnostic interview - D	Continuous
RICOVERO	Patient hospitalisation	Binary
ADOS_LING/COMUN	ADOS - Language/communication	Continuous
ADOS_INT.SOC.	ADOS - Social interaction	Continuous
ADOS_TOT	ADOS - Total score	Continuous
PSI_PD	PSI - Parental distress	Continuous
PSI_P-CDI	PSI - Dysfunctional interaction	Continuous
PSI_DC	PSI - Difficult child	Continuous
PSI_RISP.DIF	PSI - Defensive responding	Continuous
PSI_STRESST.	PSI - Total stress	Continuous
CBCL_TIPO	CBCL scale	Categorical
CBCL_INT.PROBL.	CBCL - Internalising problems score	Continuous
CBCL_EXT.PROBL.	CBCL - Externalising problems score	Continuous
CBCL_TOT.PROBL.	CBCL - Total problems score	Continuous
VABS_COMUN.	VABS - Communication	Continuous
VABS_AB.QUOT.	VABS - Daily living	Continuous
VABS_SOC.	VABS - Socialisation	Continuous
VABS_AB.MOT.	VABS - Motor skills	Continuous
QI/QS	Intelligence / Strategic quotient	Continuous
ADOS_MODULO	ADOS module	Categorical
TEST_COGNITIVO	Cognitive test type	Categorical

evaluate stress in parent-child relationships [84]. The Vineland Adaptive Behavior Scale (VABS) is a semi-structured interview of the parent regarding different domains namely communication (VABS_COMUN.), daily living (VABS_AB.QUOT.), socialisation (VABS_SOC.) and motor skills (VABS_AB.MOT.) [85]. All these features represent the score for each of test/interview and their related subcategories. RICOVERO is a binary feature describing if the patient has been hospitalised or not, CBCL is the Child Behavior Checklist already seen for the dyslexia data set and ADOS MODULO is a categorical feature representing the module employed for the ADOS. Finally the cognitive test (TEST_COGNITIVO) is a categorical feature stating the kind of cognitive test performed.

3.2.2.1 T0 data with T0 diagnosis as label

The first subset of data included only samples at T0. After having removed all the columns and/or patients with missing values, the subset includes 90 samples and 25 features (those reported in Table 3.4). This shows the first limitation of this data set, that is the large number of missing values. As already discussed for the dyslexia

case, this problem highly affects the analysis preventing the use of deep learning methods as well as the reliability of the results. This recurrent issue is probably due to the original size of the data set (in both terms of number of patients and time span analysed) which may have caused different physicians to define the diagnosis and filling the information about the patients.

The aim of the analysis performed on this first subset is to classify the patients according to a binary label. A binary label has been used instead of the original multi-label classification due to the large number of missing values and the presence of few samples with specific rare disorders. Transforming the labels also reduced the original imbalance among the classes. For this reason, the label 1 has been assigned to patients affected by development disorder autism (DGS - *disturbo generalizzato dello sviluppo*) and the label 0 to all the others. Despite having grouped the labels, the two classes are still unbalanced with DSG having twice the number of the other samples.

In order to classify patients according to their diagnosis group, SVM with both linear and RBF kernels have been employed. These models have been chosen in order to classify patients using both linear and non-linear approaches. Table 3.5 shows the vales of AUC, accuracy and balanced accuracy for the two classifiers employed in the analysis, while Fig. 3.7 shows the results related to the performance in terms of ROC curve and AUC for the same models. The best parameters identified for the two models are C=1 and C=100 and gamma=0.1 respectively. In both cases the C parameter has been adjusted according to the class weight to take the data set imbalance into consideration. As one can see, the models did not reach very high values of accuracy or AUC, but considering the difficulty of the task these values are still acceptable and the results might be useful in guiding the diagnosis process.

In addition to this supervised approach, the unsupervised method described in section 2.3.2 has been applied to verify whether other groups of patients were present and which features determine those groups. Fig. 3.8 clearly shows that while the patients are all mixed together, independently of their diagnosis, the presence of at least three different clusters is evident as shown in Fig. 3.9 where spectral clustering has been used to group the patients (in the original space). The number of clusters used was three in order to avoid the identification of smaller and less relevant groups. One also has to consider that the data representation in Fig. 3.9 is merely a two-dimensional approximation made with t-SNE, hence the samples might be closer in the original space where the actual spectral clustering was performed. The

TABLE 3.5: Linear and RBF SVM performances on the prediction of T0 diagnosis with autism data. The table shows the values of area under the ROC curves, accuracy and balanced accuracy for each of the applied models.

Methods	**Test Accuracy**	**Test Balanced Accuracy**	**Test AUC**
Linear SVM	0.69	0.65	0.69
RBF-SVM	0.62	0.58	0.69

classification tree in Fig. 3.10 shows the rules according to which the groups are formed. In particular, the values of the ADOS module employed (whether ADOS MODULO is 1 or not), the kind of cognitive test (TEST COGNITIVO type 2 or not) and VABS motor skills (VABS AB. MOT) all at T0 bring a different classification than the original one. In spite of the results, further investigations are needed in order to understand if these results might be useful in a real world clinical scenario.

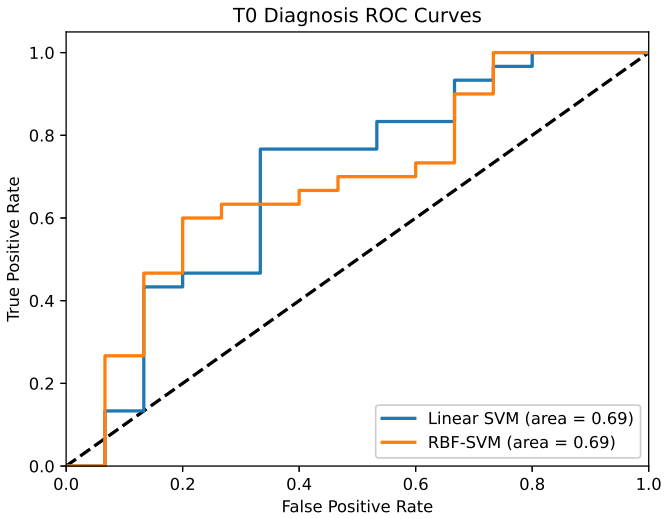


FIGURE 3.7: ROC curve for each of the classification algorithms employed for the prediction of T0 diagnosis with autism data.

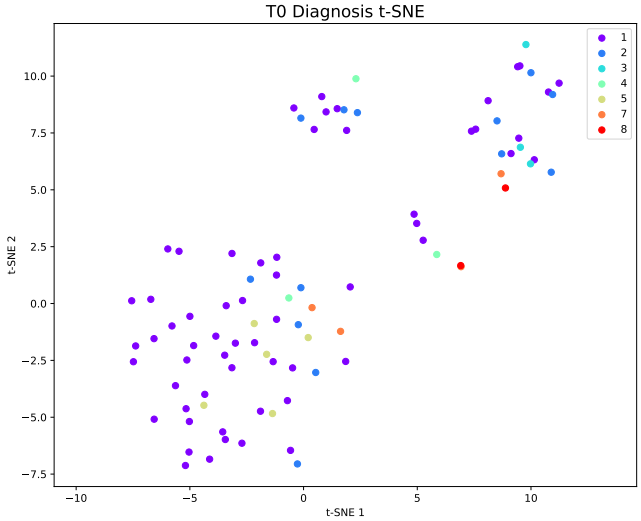


FIGURE 3.8: t-SNE two dimensional projection of autism data at T0. Each label represents one of the possible available diagnosis for the patients.

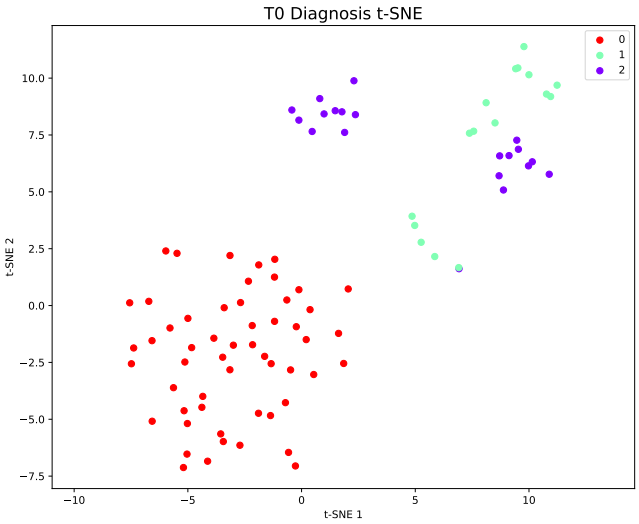


FIGURE 3.9: t-SNE two dimensional projection of clustering on autism data at T0. Zeros, ones and twos represent the three clusters classes identified by spectral clustering.

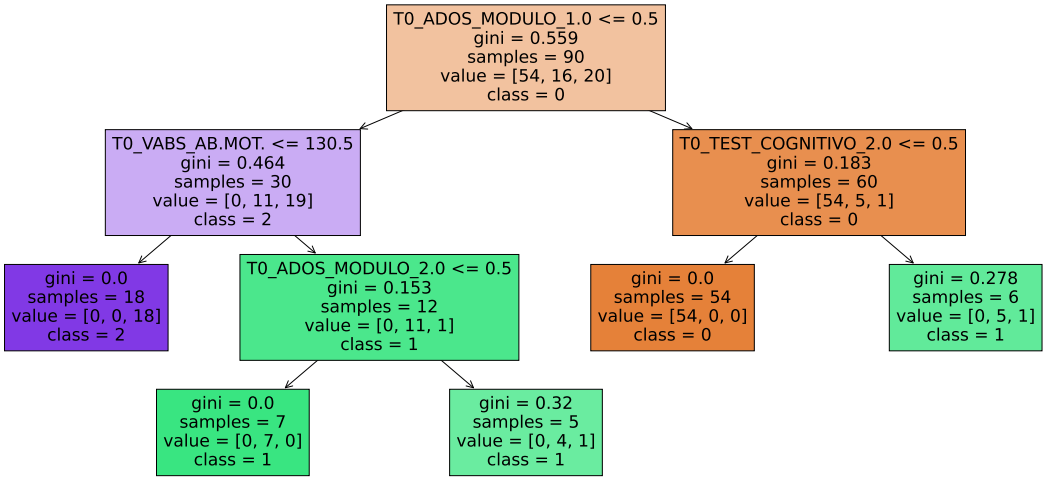


FIGURE 3.10: Classification tree rules based on the labels induced by spectral clustering algorithm on autism data at T0.

3.2.2.2 T0 data with T1 diagnosis as label

A second analysis was done using the features related to T0 with the aim to predict the diagnosis at time T1. The idea was to predict what would have been the diagnosis according to some information related to the previous time window. Since only the patients with available data at T1 were considered, the number of missing values increased as stated before. As a result, in this case the data set is composed of 55 patients and 25 features (the list is reported in Table 3.4). As in the previous section, the diagnosis types were grouped to convert the multi-class problem to a binary one. The same approach was also applied but in this case the results were better, showing that information at time T0 can better predict the diagnosis at T1 in comparison with T0 as shown in Fig. 3.11 and Table 3.6. In this case, the best hyper-parameters identified by cross-validation were C=1 for the linear case, while C=1 and gamma=0.01 were used for the RBF kernel. In both cases the C parameter has been adjusted according to the class weight. When it comes to the unsupervised analysis, even if the patients are different than before, three clusters are still visible by using k-means algorithm as shown in Fig. 3.12, while the rules describing these groups depends on the ADOS module (ADOS MODULO = 1 or not) and VABS socialisation score (VABS SOC.) as described in the classification tree shown in Fig. 3.13.

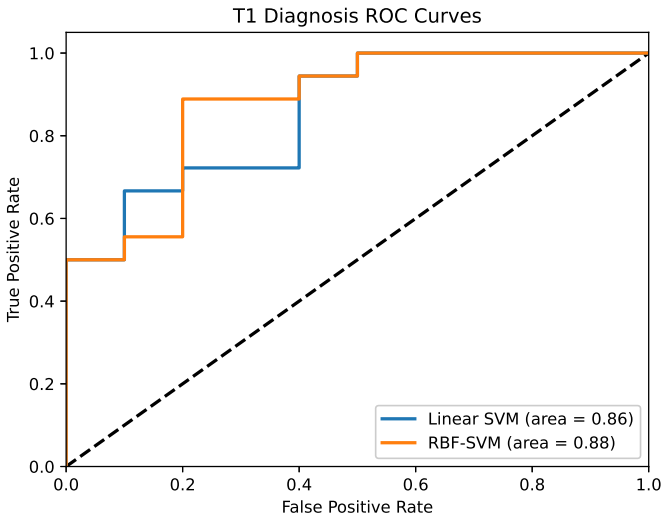


FIGURE 3.11: ROC curve for each of the classification algorithms employed for the prediction of T1 diagnosis with T0 autism data.

TABLE 3.6: Linear and RBF SVM performances on the prediction of T1 diagnosis with T0 autism data. The table shows the values of area under the ROC curves, accuracy and balanced accuracy for each of the applied models.

Methods	Test Accuracy	Test Balanced Accuracy	Test AUC
Linear SVM	0.71	0.73	0.86
RBF-SVM	0.68	0.73	0.88

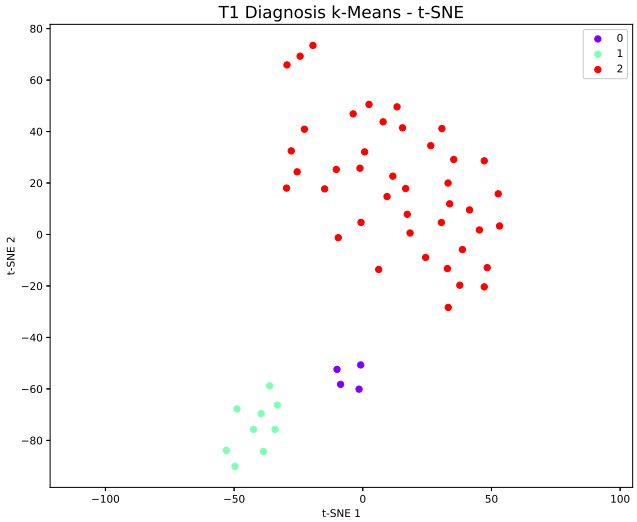


FIGURE 3.12: t-SNE two dimensional projection of autism data at T0 with T1 diagnosis. Zeros, ones and twos represent the three clusters classes identified by k-means.

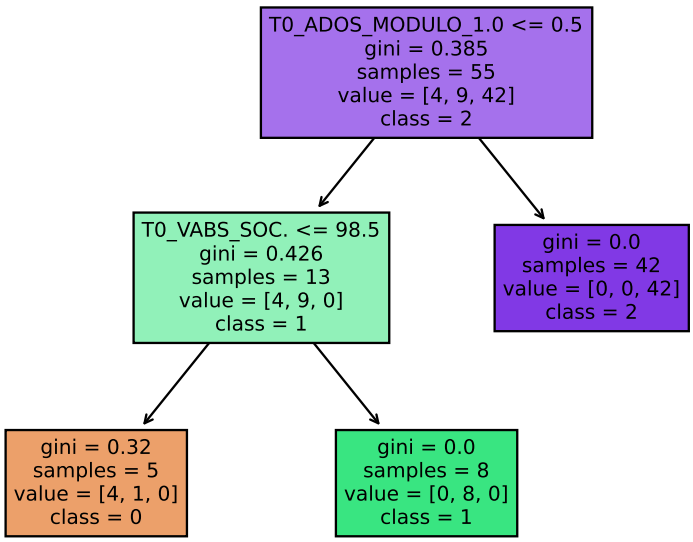


FIGURE 3.13: Classification tree rules based on the labels induced by k-means clustering algorithm on autism data at T0 with T1 diagnosis.

3.2.2.3 T0 data with T0 PVB form as label

The third analysis involved the use of a different variable as label: instead of using the original diagnosis information, one of the features has been used. In particular the PVB form (child's first vocabulary) was chosen. This feature can assume 4 different values: gestures and words - short form, gestures and words - long form, words and sentences - short form and words and sentences - long form. These values have been converted to binary and two different labelling have been used: short forms versus long forms and gestures and words versus words and sentences. In this case, the samples without missing values were even less, therefore a combination of 42 patients and 25 features (as reported in Table 3.4) were used (decreasing the number of features did not give any significant improvement). As shown in Fig. 3.14 and Fig. 3.15 the data allow to better predict gestures vs words form rather then short vs long forms. The best hyper-parameters identified by cross-validation for the two analyses are reported in Table 3.7.

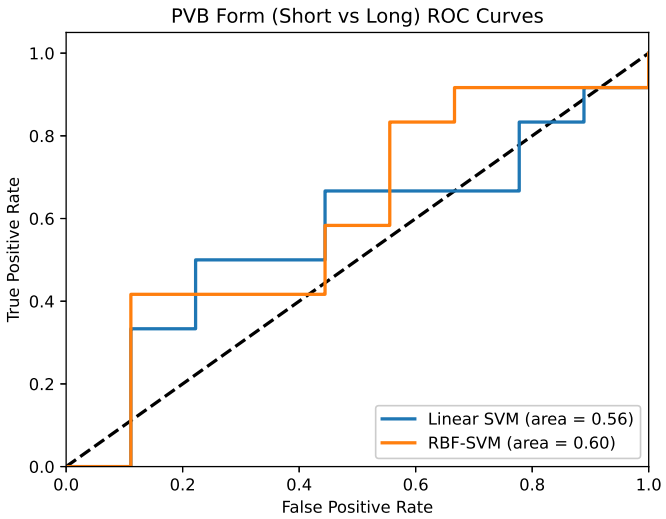


FIGURE 3.14: ROC curve for each of the classification algorithms employed for the prediction of the PVB form (short vs long) with autism data.

TABLE 3.7: Supervised algorithms best hyper-parameters for prediction of PVB forms on autism data. The table shows the best hyper-parameters identified by a grid search cross-validation for the applied models and the related PVB form prediction.

Short vs Long		**Gestures vs Words**	
Linear SVM	RBF-SVM	Linear SVM	RBF-SVM
C: 1	C: 100	C: 1	C: 10
Balanced weights	Gamma: 0.001	Balanced weights	Gamma: 0.1
	Balanced weights		Balanced weights

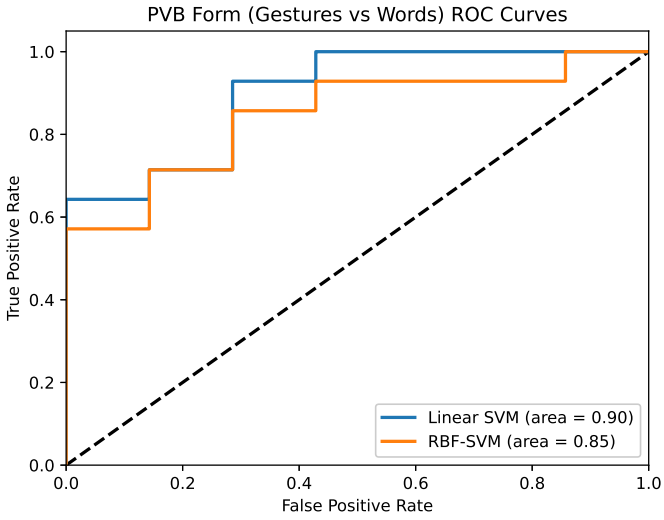


FIGURE 3.15: ROC curve for each of the classification algorithms employed for the prediction of the PVB form (gestures vs words) with autism data.

When it comes to the unsupervised analysis, there are no clusters clearly visible as shown in Fig. 3.16, however by using the spectral clustering algorithm one can identify two clusters (see Fig. 3.17). The rules describing this new classification are shown in the classification tree in Fig. 3.18: ADOS total score (ADOS TOT) and VABS communication (VABS COMUN.) define the two groups of patients.

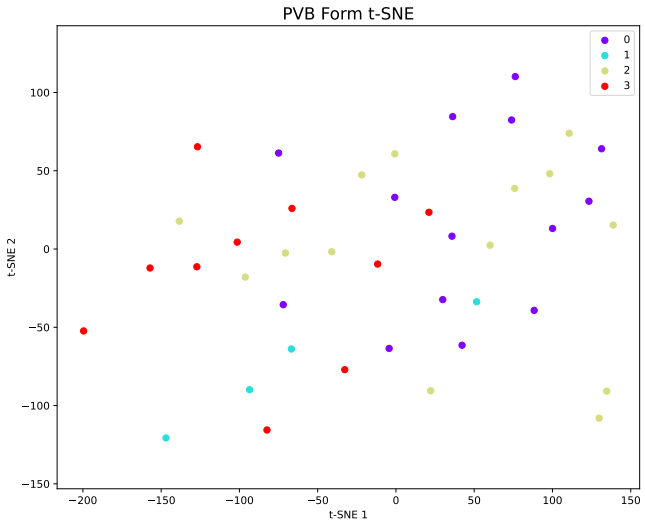


FIGURE 3.16: t-SNE two dimensional projection of autism data (PVB form). Each label represents one of the PVB forms: gestures and words (short form), gestures and words (long form), words and sentences (short form) and words and sentences (long form) respectively.

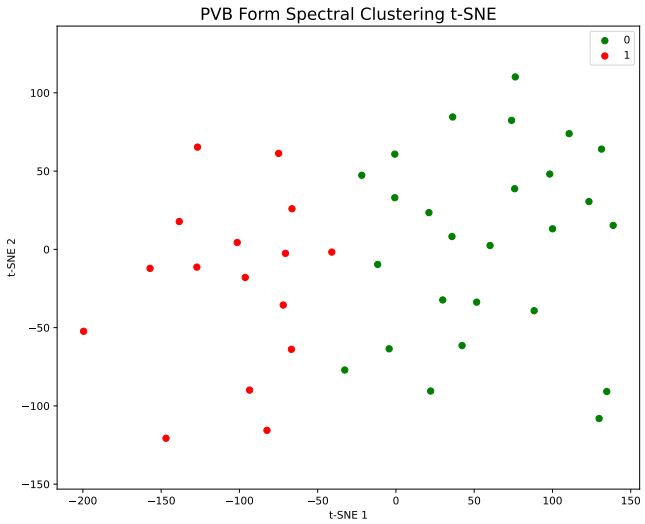


FIGURE 3.17: t-SNE two dimensional projection of clustering on autism data (PVB form). Zeros and ones represent the two clusters classes identified by spectral clustering.

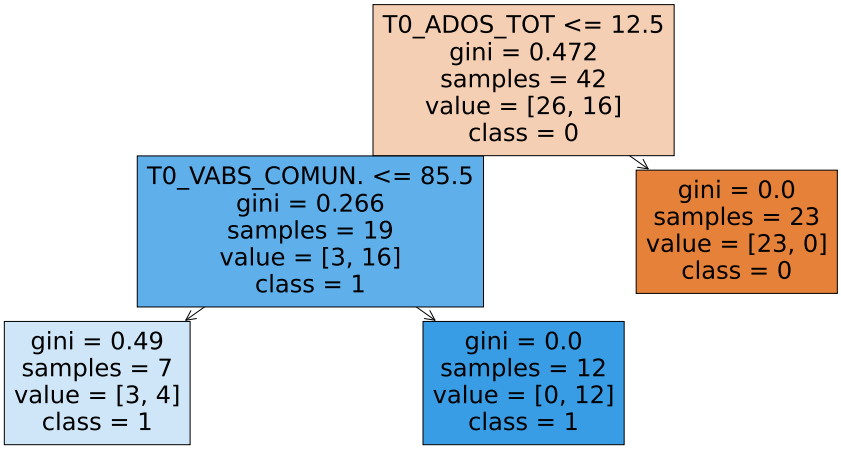


FIGURE 3.18: Classification tree rules based on the labels induced by spectral clustering algorithm on autism data (PVB form).

3.2.2.4 T0 data with T1 diagnosis as label (excluding cognitive test)

The forth analysis involved the prediction of T1 diagnosis using T0 features as in the first subset. However, in this case the feature related to the cognitive test result has been excluded. This decision has been made under the suggestion of the clinicians involved in the analysis since the results of the test might bias the diagnosis in a relevant way. The data set is thus composed of 55 patients and 24 features (as in Table 3.4) and the two classes (DGS form versus all the others) are unbalanced. As shown in Fig. 3.19, even after removing cognitive test as feature, both linear and RBF SVM kept their prediction performances high. The final hyper-parameters identified by a grid-search cross-validation are shown in Table 3.8

In this case one can also go further by trying to identify the most relevant features for the described classification. In order to do this and thus obtaining some feature importance, recursive feature elimination with a linear SVM and random forest were exploited: their area under the ROC curve are also shown in Fig. 3.19. On the other

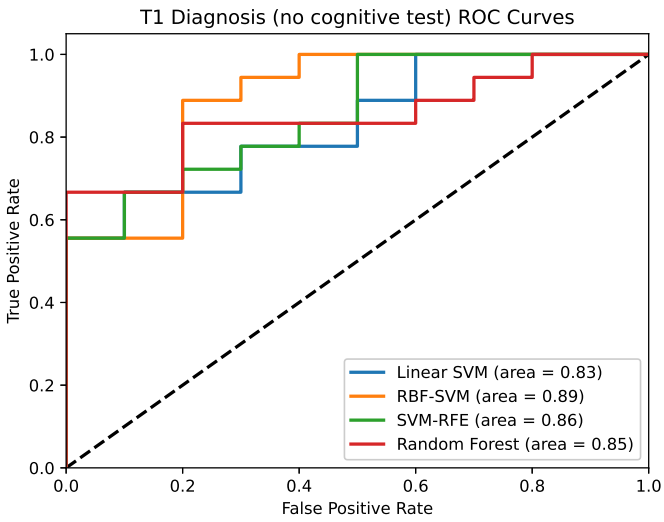


FIGURE 3.19: ROC curve for each of the classification algorithms employed for the prediction of T1 diagnosis with T0 autism data excluding cognitive test.

TABLE 3.8: Supervised algorithms best hyper-parameters for autism data without cognitive test. The table shows the best hyper-parameters identified by a grid search cross-validation for each of the applied models.

Linear SVM	RBF-SVM	SVM-RFE	Random Forest
C: 1	C: 10	C: 1	No. of trees: 400
Balanced wgt.	Gamma: 0.01	No. of features: 10	Split criterion: Entropy
	Balanced wgt.	Balanced wgt.	Max depth: 3
		Step: 1	Max features: Sq. root
			Min samples leaf: 4
			Min samples split: 2

hand, Fig. 3.20 and Fig. 3.21 show the lists of the ten most important features for the two methods employed. As one can see, there are seven out of ten common features and the most important ones are three ADOS tests (ADOS TOT., ADOS LING. COMUN and ADOS INT. SOC.) and ADI A (in different order, according to the classification method).

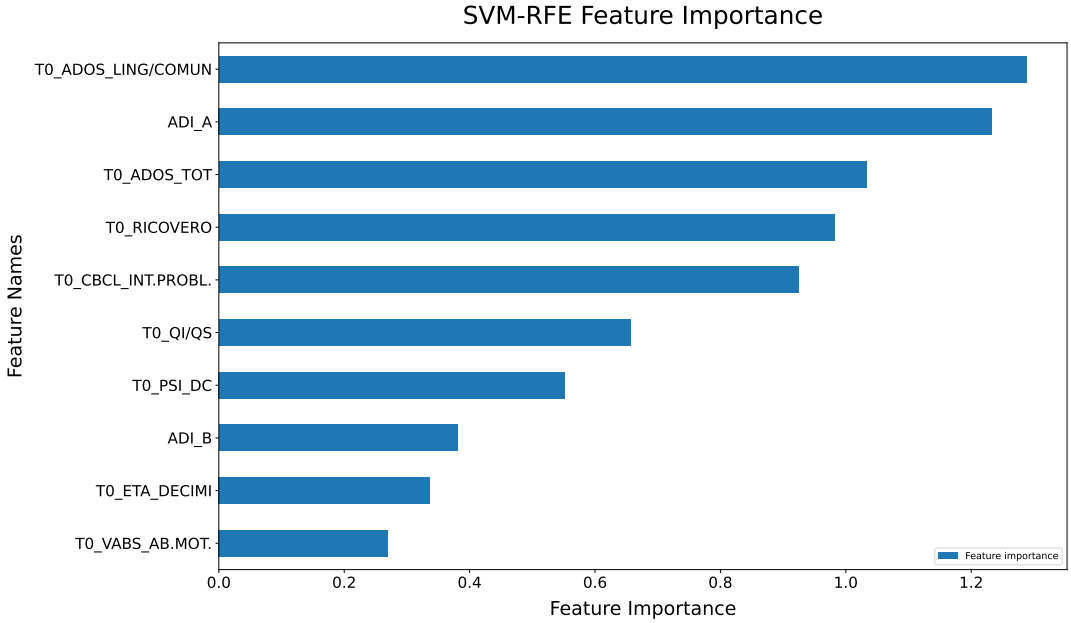


FIGURE 3.20: List of the ten most important features according to RFE with linear SVM in descending order for the prediction of T1 diagnosis with T0 autism data excluding cognitive test.

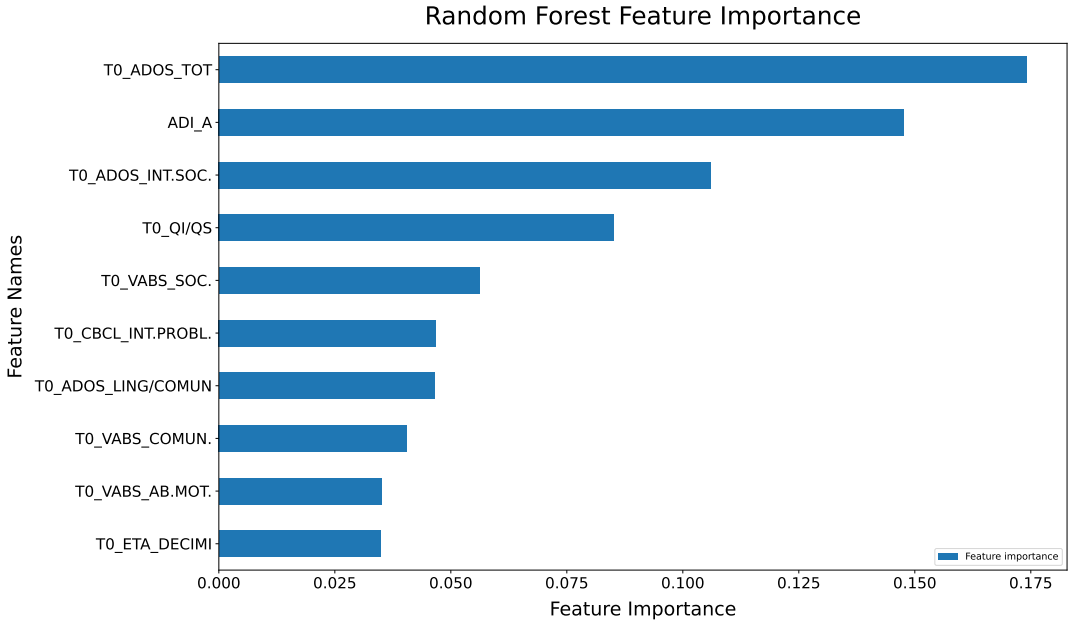


FIGURE 3.21: List of the ten most important features according to random forest in descending order for the prediction of T1 diagnosis with T0 autism data excluding cognitive test.

3.2.2.5 T0 selected features with T1 diagnosis as label

The final analysis involving this data set has been performed once again following the suggestions of the clinicians involved in the project. In order to avoid a bias classification due to some of the features incorporated into the data set, the same analysis was performed by using only few specific features namely: QI/QS, ADOS TOT., VABS COMUN., VABS AB. QUOT., VABS SOC. and VABS AB. MOT. These features represent some test results such as the intelligent quotients, the total score of the Autism Diagnostic Observation Schedule (ADOS TOT.) and the four Vineland Adaptive Behavior Scales related to communication, daily living, socialisation and motor skills respectively. In this case, with only six features and few missing values, the data set is composed of 241 samples. Despite of the increased sample size, the two classes (DGS Autism vs all the others) are still unbalanced and thus the SVM functional was corrected according to the classes size. Fig. 3.22 and Table 3.9 show the performance of linear and RBF SVM employed on the analysis.cIn this case the performance are lower than before pointing to a lower prediction ability of the selected features. The best hyper-parameters identified by a grid-search cross-validation were C=10 for linear SVM and C=10 and gamma=1 for the RBF model. Finally, having only six available features, it was not necessary to perform a proper feature selection as in the previous analysis, thus only the prediction capabilities of the selected features were verified.

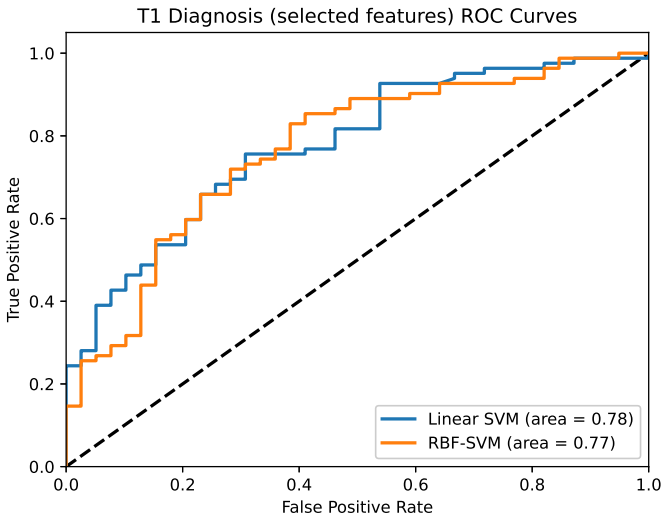


FIGURE 3.22: ROC curve for each of the classification algorithms employed for the prediction of T1 diagnosis with autism selected features at time T0.

TABLE 3.9: Linear and RBF SVM performances on the prediction of T1 diagnosis using autism selected features at time T0. The table shows the values of area under the ROC curves, accuracy and balanced accuracy for each of the applied models.

Methods	Test Accuracy	Test Balanced Accuracy	Test AUC
Linear SVM	0.69	0.69	0.78
RBF-SVM	0.72	0.70	0.77

In conclusion, in all the subsets of data considered, good accuracy values were achieved especially when considering the prediction of T1 diagnosis. At the same time, using a clustering algorithm combined with a classification tree, it was possible to identify the features determining the groups and possibly support the diagnosis process. However, in all the cases the number of available samples was very limited hence increasing the sample size would be useful to improve the reliability of the results.

3.3 Osteogenesis imperfecta

Osteogenesis imperfecta (OI), also known as "brittle bone disease", is a rare genetic disease characterised by many symptoms including some of the features of osteoporosis causing increased bone fragility and other issues such as low bone mass, connective tissue manifestations, and other extraskeletal manifestations. These might include dentinogenesis imperfecta, hyperlaxity of skin and ligaments, blue sclera, wormian bones and impaired hearing [86]. These symptoms are typically used to diagnose the disease, with particular attention to the presence of blue sclera and dentinogenesis imperfecta, nonetheless some limitations are present and other information is needed in order to correctly identify the disorder. In most cases, people affected by osteogenesis imperfecta also have a mutation in one of the genes encoding the α chains of collagen type 1, namely COL1A1 and COL1A2 respectively located on chromosome 17 and 7. Analysis about the mutation of these two genes can thus be performed in order to provide useful information for the diagnosis process. In other cases, OI is easy to detect due to the presence of affected family members [87].

Osteogenesis imperfecta may have different degrees of severity, from lethal perinatal intrauterine fractures to mild forms without fractures. Originally four different types of osteogenesis were defined [86]:

- **Type I**: it is the most common one. It is characterised by mild disease without major bone deformities and blue sclera.
- **Type II**: it is characterised by respiratory failure due to multiple rib fractures and it is lethal in perinatal period.
- **Type III**: patients are very short in stature with severe spine deformities and fractures. Sclera is less blue and more greyish.

- **Type IV**: it is characterised by normal sclera and mild to moderate bone deformities and short stature. It can be seen as an intermediate severity between type I and III.

However, in recent years additional types of OI were introduced with the aim to fill the gaps between the four original categories and trying to separate the original types according to the levels of clinical entries described above. Three more have been added in 2004 [87] and later, with the discovery of OI genetic determinant the categories were extended to fifteen. While the first four categories were established on radiological and clinical information about the disorder, the categories from V to XV were defined on the basis of molecular biomarkers. The original classification has then increased to twenty different type of osteogenesis [88]. Due to the large number of possible classifications and the resulting difficulty in clinical practice, the original four types of osteogenesis are still used in practice, especially because they are the most common ones.

In order to improve and support the diagnosis process, one can ideally rely on the use of objective ways to detect OI, as in the case of many other diseases or disorders. Nonetheless, there is a limited amount of works in literature regarding the use of machine learning or artificial intelligence to help detecting osteogenesis imperfecta. One of the first work published in this matter is about the use of structural models and machine learning in order to predict the mutated sequences associated with osteogenesis imperfecta [89]. By applying a high throughput mutation analysis using molecular dynamics, Authors were able to identify some structural interactions associated with the disease. On the other hand, other works did not make use of clinical and/or genetic information but used images instead. Authors in [90] performed a classification using the images of 40 bone fragments from 28 patients where 13 were affected by OI and 15 were healthy controls. By using machine learning and in particular the SVM classifier, they were able to classify the subjects with an AUC of 0.96. Moreover, Authors in [91] performed an analysis on a total of 306 subjects of which 173 were affected by OI and 133 were the controls with the aim to use facial features to detect OI. By using a PCAlog model on the patients face images the Authors were able to identify specific morphological features (at the level of temples and eyes) which can help to distinguish cases from controls. The patients were affected by OI types I, III and IV and the analysis also confirmed that subjects affected by OI type III present the most severe facial characteristics.

Due to the large number of OI types and the difficulties that may arise from this issue, by analysing a data set of patients affected by different forms of OI one can aim at possibly identifying new alternative ways to devise the diagnosis, reduce the number of available types and thus reduce the issues that may be due to misdiagnosis. This study involved the use of two different OI data sets whose results are discussed in the following sections.

TABLE 3.10: List of features employed for the analysis on the osteogenesis imperfecta data.

Osteogenesis Imperfecta Data Features List		
Feature Name	**Feature Description**	**Feature Type**
Age	Patient age	Continuous
Child/Adult	Patient age (category)	Binary
Gender	Patient gender	Binary
Proband	Patient proband	Binary
Family History	Familiarity with OI	Binary
Height	Patient height	Continuous
Height Percentile	Patient percentile height	Categorical
Sclera	Coloured sclera	Binary
Dentinogenesis Imperfecta	Dentinogenesis imperfecta diagnosis	Binary
Joint Hyperlaxity	Joint hyperlaxity diagnosis	Binary
Spine Deformity	Spinal column deformity diagnosis	Binary
Long Bones Deformity	Long bones deformity diagnosis	Binary
Deafness	Deafness diagnosis	Binary
Cardiac Lesion	Cardiac lesion diagnosis	Binary
Movement Disorder	Movement disorder diagnosis	Categorical
COL Mutation	COL gene mutation	Binary
Gene Involved	Gene involved	Categorical
Mutation Type	Causative mutation type	Binary
Struttura	Mutation site	Categorical

3.3.1 The data set I

This data set is about patients affected by OI disease from the Dr. Luca Sangiorgi's group at CLIBI laboratory of the Istituto Ortopedico Rizzoli (Rizzoli Orthopaedic Institute) in Bologna, Italy. It includes 489 patients of different ages from newborn children to older people. For each patient a total of 22 features is available. The list of available features and their type is reported in Table 3.10. These features include general information about the patients such as their gender, age and family history, as well as clinical information related to the disease such as the coloured sclera, height, number of fractures and genetic information related to the mutated gene/genes and their mutation type. Due to the nature of this information, most of the features are categorical or binary while the continuous ones are only two (age and height). As the other clinical data sets considered before, this bears a major drawback: being the data related to a very rare disease, the number of missing values is very large, especially considering that in some cases the data is related to patients born dead for whom it was not possible to get much information. In other cases, not all the patients agreed to their genes sequencing, thus this kind of information is missing in those cases. For each patient a diagnosis label is present. This label describes one of the five possible type of OI collected in the database (the four original classes plus undifferentiated). As in the former analysis, no control patients are present and thus all the patients are affected by the disease.

Due to the particular clinical interest in this disease, four different subsets have

been created in order to better identify relevant information about the diagnosis process:

- Whole data set.
- All patients with some features excluded.
- 18-50 years old patients only.
- 18-50 years old patients only with some features excluded.

These different analyses have been performed in order to evaluate all the possible cases. In particular, there are some features such as the number of fractures, which bring trivial information about the disease. One way to identify other relevant features is thus to drop these features completely from the analysis. At the same time, since the more clinically interesting cases are related to adults, the other analysis included only patients with age between 18 and 50 years. For each analysis, all the data was normalised in the range [0, 1] and all categorical variables were treated using one-hot encoding. Scikit-learn [70] library has been used to implement the known machine learning algorithms.

3.3.2 Results I

3.3.2.1 Whole data set

In this first analysis all the patients and all the features have been considered. The starting size of the data set is 489 patients and 22 features, however, due to the many missing values, the effective size of the subset is 65 patients and 17 features. Fig 3.23 shows the data set missing values distribution.

As in the previous cases, some variables have been dropped because they had too many missing entries, after that only the patients without missing values in the remaining features have been removed. After performing data pre-processing, an unsupervised analysis combined with a classification tree was performed, as previously described in section 2.3.2. As shown in Fig. 3.24, the only classes present after removing all the missing values are osteogenesis types 1, 3 and 4. In particular there is only one sample corresponding to osteogenesis type 3, while type 4 has fewer samples than type 1. All the samples distribution appears uniform, with no clear distinction: by using k-means clustering one can isolate the two clusters as in Fig. 3.25, while the classification tree shown in Fig. 3.26 shows the rules according to which the two clusters are formed. In this case, since no cluster was clearly evident, the samples are merely grouped according to their age and height, thus no relevant information has been extracted since a trivial classification has been performed with the tree probably overfitting the data.

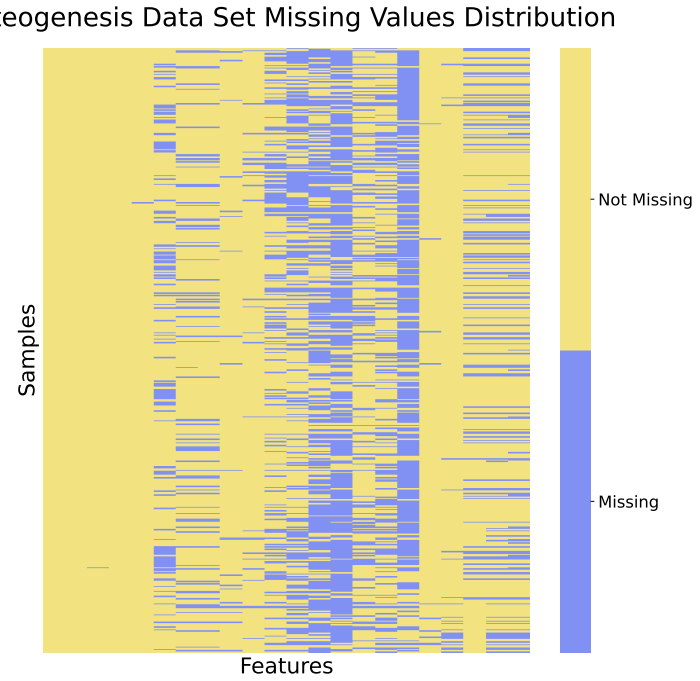


FIGURE 3.23: Missing values distribution in the osteogenesis imperfecta data set.

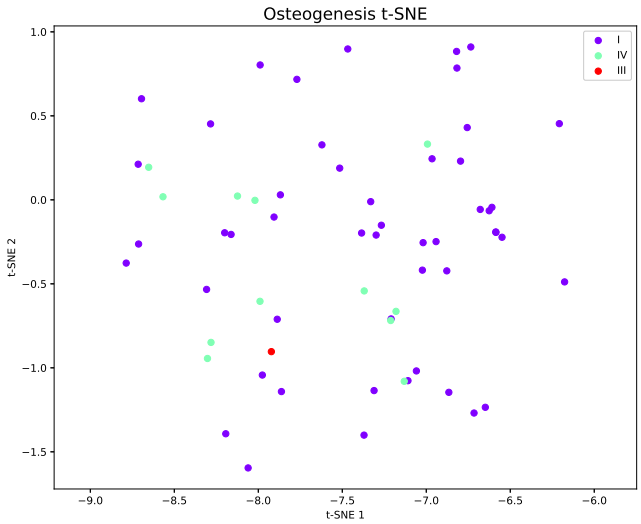


FIGURE 3.24: t-SNE two dimensional projection of osteogenesis data. Each label represents one of the osteogenesis type diagnosis after removing all the missing values.

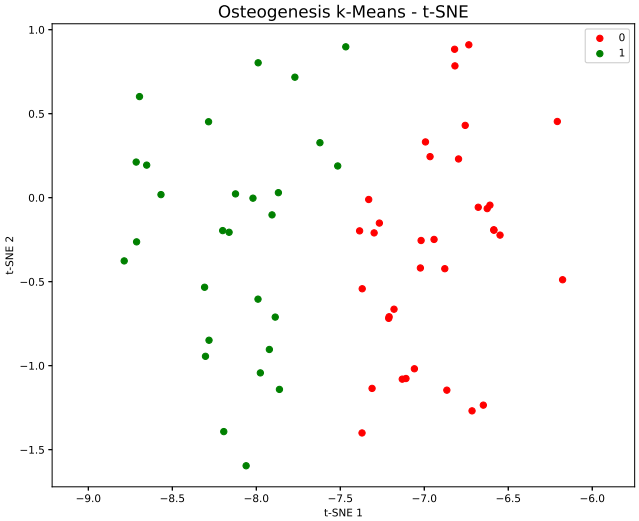


FIGURE 3.25: t-SNE two dimensional projection of clustering on osteogenesis data. Zeros and ones represent the two clusters classes identified by k-means.

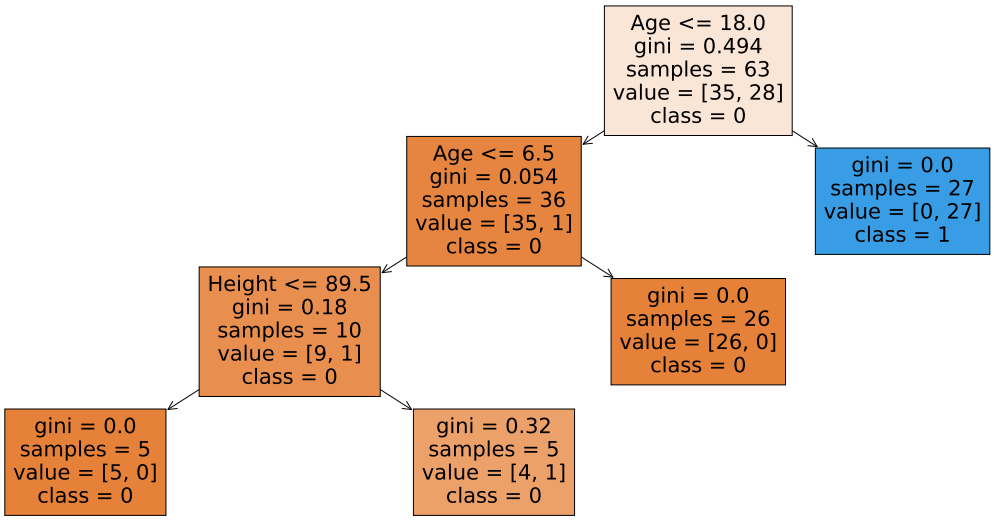


FIGURE 3.26: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data.

3.3.2.2 All patients with some features excluded

The second analysis involved the use of the same data set as before, however three features have been excluded namely the number of fractures, BMD (bone mineral density) spine and anti-fractures drugs. According to clinicians, these features might bring misleading results and trying to cluster patients without them might show more interesting results. However, one has to consider that two of these features (BMD spine and anti-fractures drugs) would have been dropped anyway due to the large number of missing values. The subset was originally composed of 489 patients and 21 features, but due to the large number of missing values it only includes 64 samples and 16 features, i.e. the data set is the same as in the previous section apart from the number of fractures that has been removed. Since the data was not much different, the results are also pretty similar to those shown in the previous section. As one can see from Fig. 3.27 the data distribution did not change very much and k-means clustering identified two clusters (see Fig. 3.28) defined by the the age of the patients, i.e. if they are over or under 18 years old, as described by the classification tree reported in Fig. 3.29.

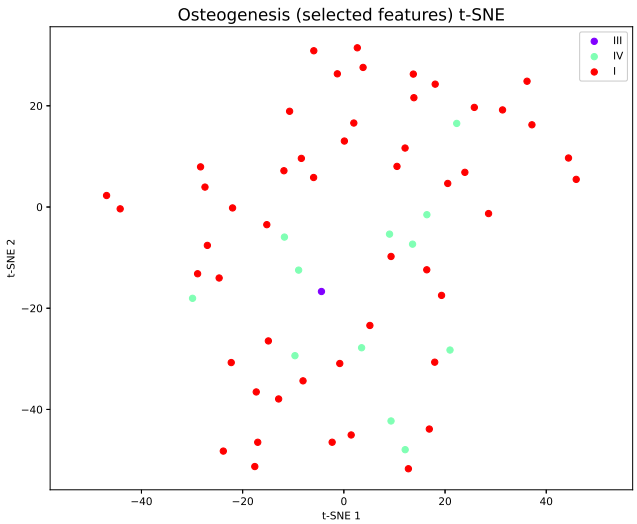


FIGURE 3.27: t-SNE two dimensional projection of osteogenesis data with selected features. Each label represents one of the osteogenesis type diagnosis after removing all the missing values.

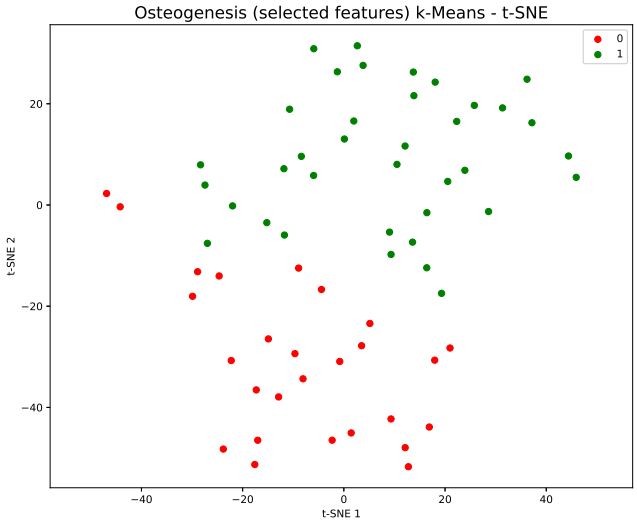


FIGURE 3.28: t-SNE two dimensional projection of clustering on osteogenesis data with selected features. Zeros and ones represent the two clusters classes identified by k-means.

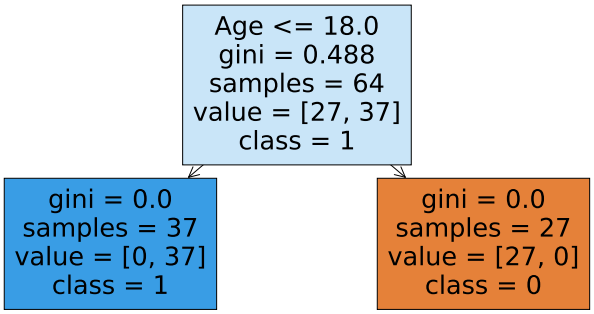


FIGURE 3.29: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data with selected features.

3.3.2.3 18-50 years old patients only

According to clinicians suggestions, it might be interesting to observe osteogenesis data in adults patients only, in order to exclude rare cases such as the born dead children. By limiting the data to patients in age between 18 and 50 years, the subset of data is composed by 199 patients and 24 features. However, exactly as before, the number of missing values is very large and the data set only includes 51 observations and 15 variables, practically. These numbers also confirm that the majority of missing values come from children data. From the previous sections it has been shown that age is a confounding factor since it trivially causes the formation of clusters. In this case, adults only are considered thus a classification according to the age of majority should be avoided. Fig. 3.30 shows the two-dimensional samples distribution: here the samples seem to be more separated than before. Spectral clustering shown in Fig. 3.31 confirms this result by identifying two different groups of patients. As before, a classification tree has been employed to select the features describing this pattern. As shown in Fig. 3.32, age has disappeared from the selected features while genetic information acquired more relevance. The first split is determined by the mutation type (whether it is qualitative or quantitative) while the other splits are made according to the involved gene (COL1A1 or COL1A2) and the family history with osteogenesis. Although these new rules might not be useful from a clinical point of view, they show that genetic information is somehow relevant in identifying osteogenesis subgroups.

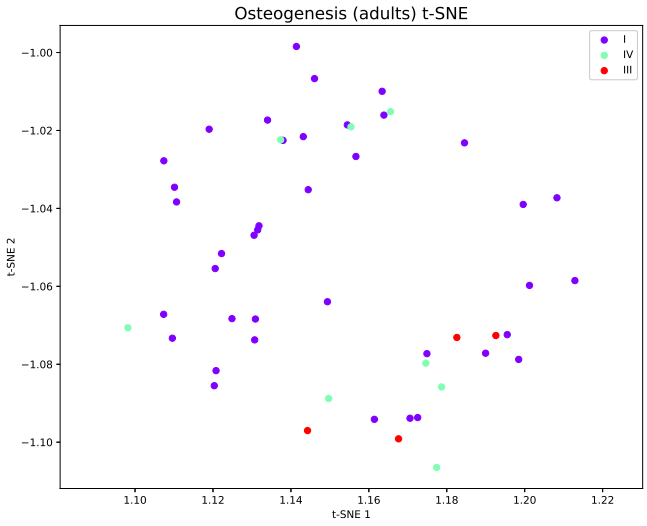


FIGURE 3.30: t-SNE two dimensional projection of osteogenesis adults data. Each label represents one of the osteogenesis type diagnosis after removing all the missing values.

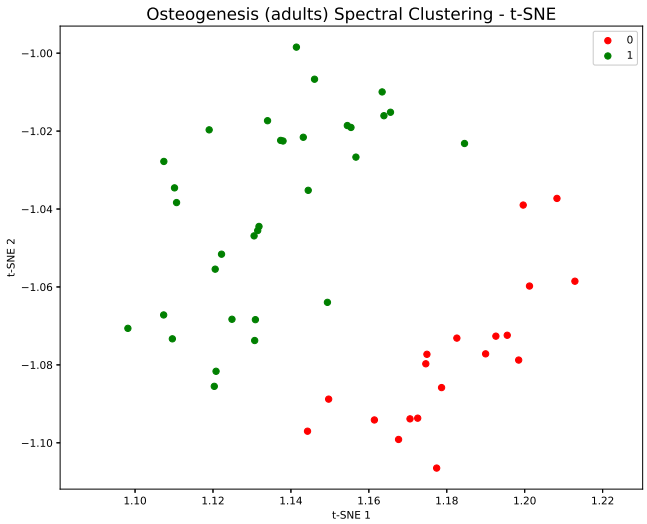


FIGURE 3.31: t-SNE two dimensional projection of clustering on osteogenesis adults data. Zeros and ones represent the two clusters classes identified by k-means.

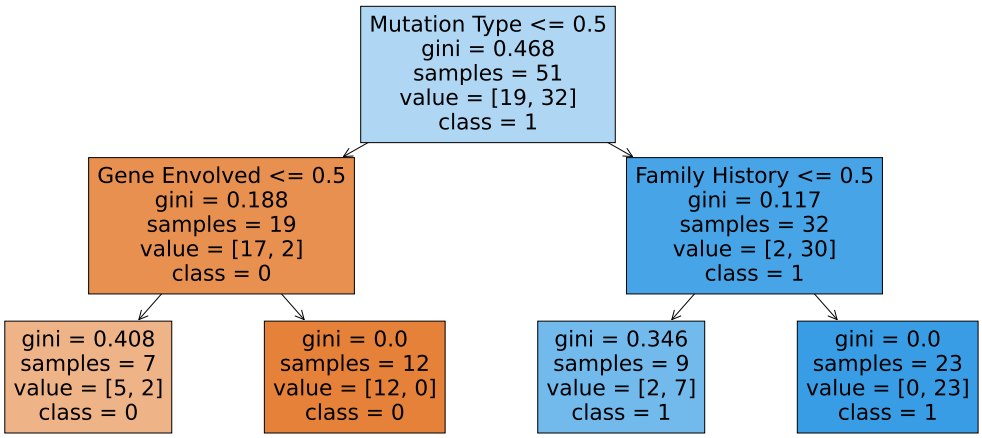


FIGURE 3.32: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis adults data.

3.3.2.4 18-50 years old patients only with some features excluded

The final analysis involving the osteogenesis data has been performed by using the adults data as in section 3.3.2.3 while keeping the same features as in section 3.3.2.2. As before, a major change from the previous analysis was not expected since the only difference is the absence of the number of fractures. As one can see from Fig. 3.33, the data distribution is mostly identical to the one shown in section 3.3.2.3 and two clusters have been identified once again by using spectral clustering (see Fig. 3.34).

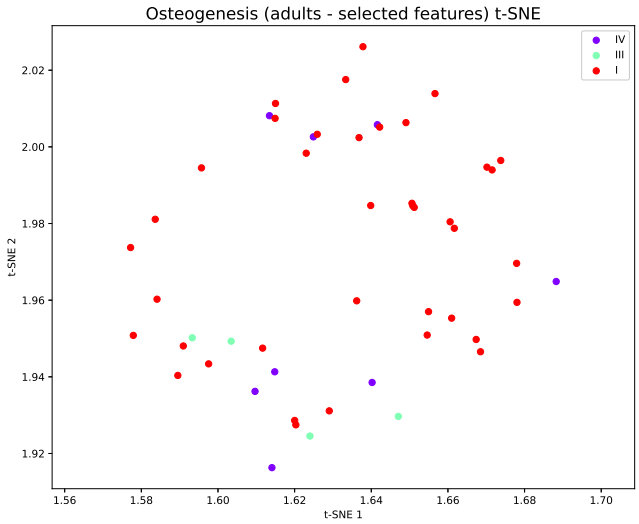


FIGURE 3.33: t-SNE two dimensional projection of osteogenesis adults data with selected features. Each label represents one of the osteogenesis type diagnosis after removing all the missing values.

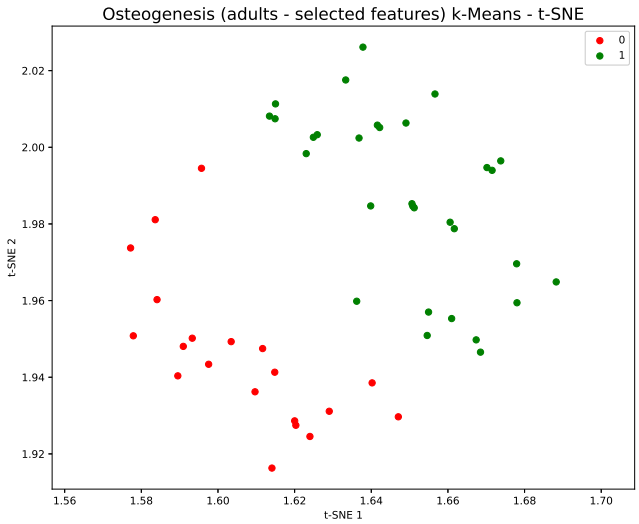


FIGURE 3.34: t-SNE two dimensional projection of clustering on osteogenesis adults data with selected features. Zeros and ones represent the two clusters classes identified by k-means.

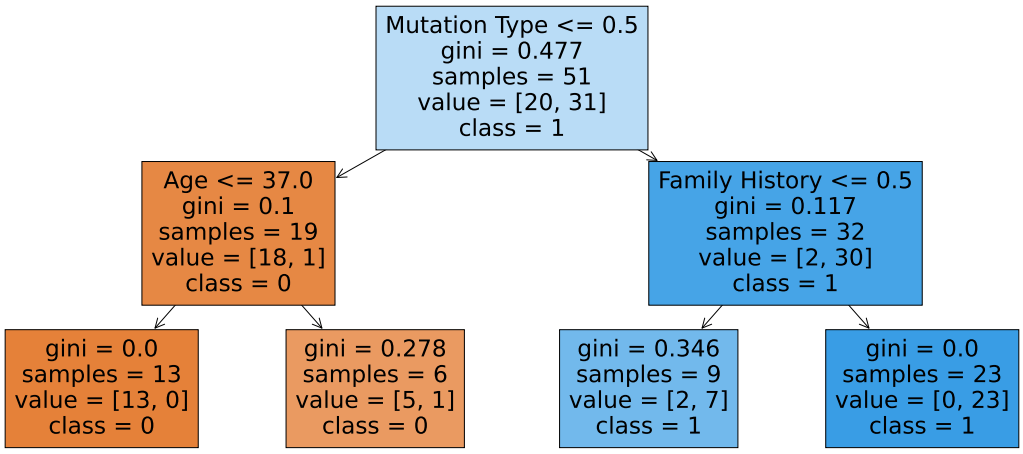


FIGURE 3.35: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis adults data with selected features.

Finally, the classification tree reported in Fig. 3.35 shows the features classifying the two groups: mutation type and family history are still relevant but in this case age substituted the type of gene involved in the mutation.

3.3.3 The data set II

As described in sections 3.3.1 and 3.3.2, the main disadvantage of the osteogenesis project was the large number of missing values which highly limited the possible analyses. After few years from the first version of the data set, the Istituto Ortopedico Rizzoli was able to collect more data and acquire more clinical and genetic information about the patients as well as an increased sample size. Therefore, this second version of the data set now includes more patients and more features for a total of 686 samples and 69 features. The features list, their description and their type are shown in Table 3.11. The patients are classified according to six possible labels: numbers from 1 to 5 (included) represent one of the five possible types of OI. On the other hand, class 0 represents all those patients affected by OI whose specific type has not been specified. Moreover, due to its peculiar characteristic (perinatal death), type II of OI has been excluded from this analysis. As before, four different subsets have been created:

- Selection 1: Whole data set.
- Selection 2: Patients with positive COL1 gene mutation.
- Selection 3: Adults.
- Selection 4: Children.

TABLE 3.11: List of features employed for the analysis on the osteogenesis imperfecta data (second version).

Osteogenesis Imperfecta Data II Features List		
Feature Name	**Feature Description**	**Feature Type**
Familiarity	Familiarity with OI	Categorical
Gender	Patient gender	Binary
Age	Age at time of visit	Continuous
Height Z	Height Z score	Continuous
Weight Z	Weight Z score	Continuous
BMI Z	BMI Z score	Continuous
Def lim 1	Head deformity or limitations	Ordinal
Def lim 2	Arms deformity or limitations	Ordinal
Def lim 3	Legs deformity or limitations	Ordinal
Def lim 4	Trunk deformity or limitations	Ordinal
Def lim 5	Pelvis deformity or limitations	Ordinal
Skin	Skin abnormality diagnosis	Binary
Joint hyperlaxity	Joint hyperlaxity diagnosis	Binary
Spine Deformity	Spinal column deformity	Categorical
Bone density	Bone density abnormality diagnosis	Binary
Deafness	Deafness diagnosis	Binary
Valv	Valvular heart disease diagnosis	Binary
Facial	Facial dysmorphic feature diagnosis	Binary
Sclera	Sclera colour	Categorical
Wormian	Wormian bones diagnosis	Binary
Dent	Dentinogenesis imperfecta	Binary
Vert	Vertebral collapse diagnosis	Binary
Fractures	No. of fractures	Ordinal
Mutation Gene CL	Causative mutation	Categorical
Mutation Type	Causative mutation type	Categorical
Mutation Sede	Mutation site	Categorical
Mutation effect	Protein mutation effect	Categorical

3.3.4 Results II

For each subset the aim was to identify clusters of patients in order to define possible alternative classifications. Under the clinicians suggestion, mutation type, mutation sede and mutation effect features have been used only in the second subset (COL1 positive).

3.3.4.1 Selection 1 (whole data set)

The first subset includes 329 patients and 24 out of the 27 features shown in Table 3.11. Fig. 3.36 shows the missing values distribution for this data set: one can observe that most of the missing values are concentrated in specific features including weight, height and BMI. Following the clinicians advice, being these features considered highly important in distinguishing the different OI types, these have been kept by dropping the patients instead. Of the remaining samples, only two are affected by OI type V, as shown in Fig. 3.37.

As in the first version of the data set, the samples seem to be mixed together with no clear separation. Therefore, the same unsupervised strategy has been adopted as before: k-means was used to group patients and then a classification tree was employed to identify which features determine the clusters. Fig. 3.38 shows the three clusters identified by k-means. In this case, however, the classification tree is not very useful since it is very complex and tends to overfit the data. As shown in Fig. 3.39, the main variable determining the first split is spine deformity, followed by two genetics features (familiarity with OI and mutation gene CL).

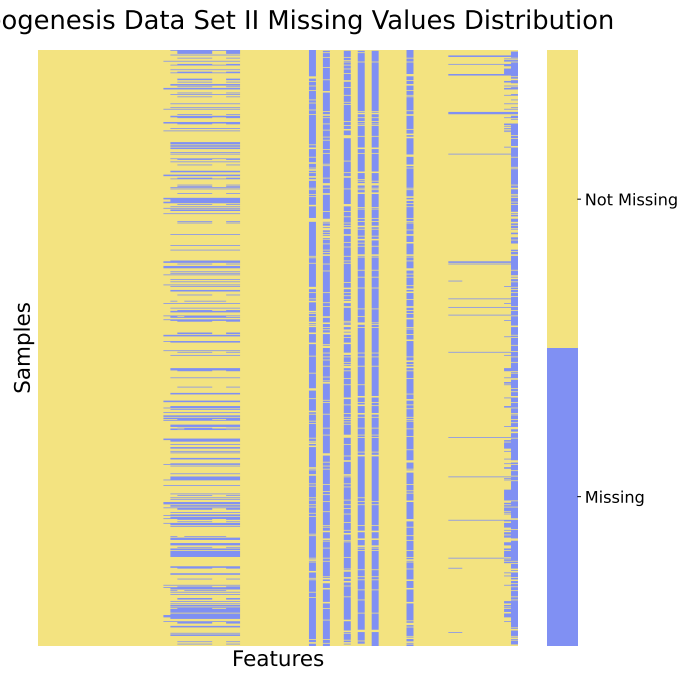


FIGURE 3.36: Missing values distribution in the osteogenesis imperfecta data set (second version).

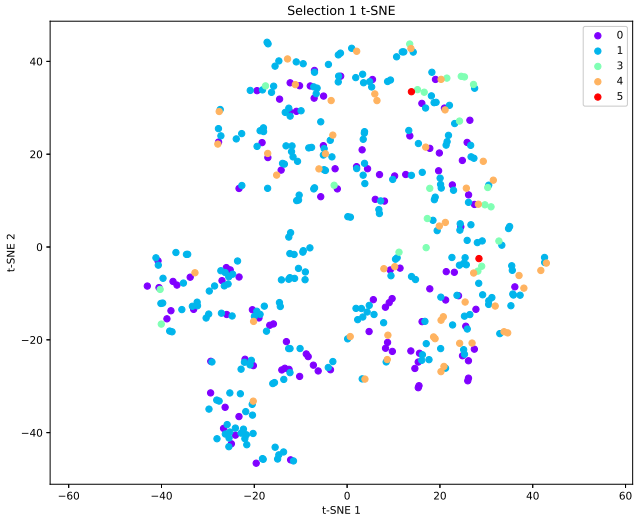


FIGURE 3.37: t-SNE two dimensional projection of osteogenesis data II (selection 1). Each number represents one of the possible types of OI.

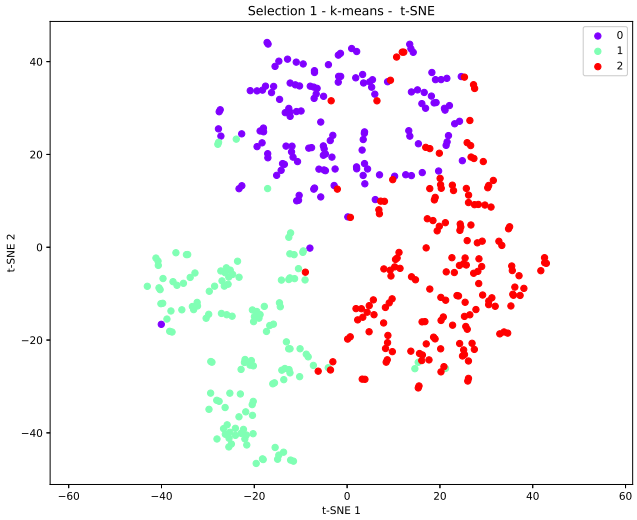


FIGURE 3.38: t-SNE two dimensional projection of clustering on osteogenesis data II (selection 1). Zero, one and two represent the three clusters classes identified by k-means.

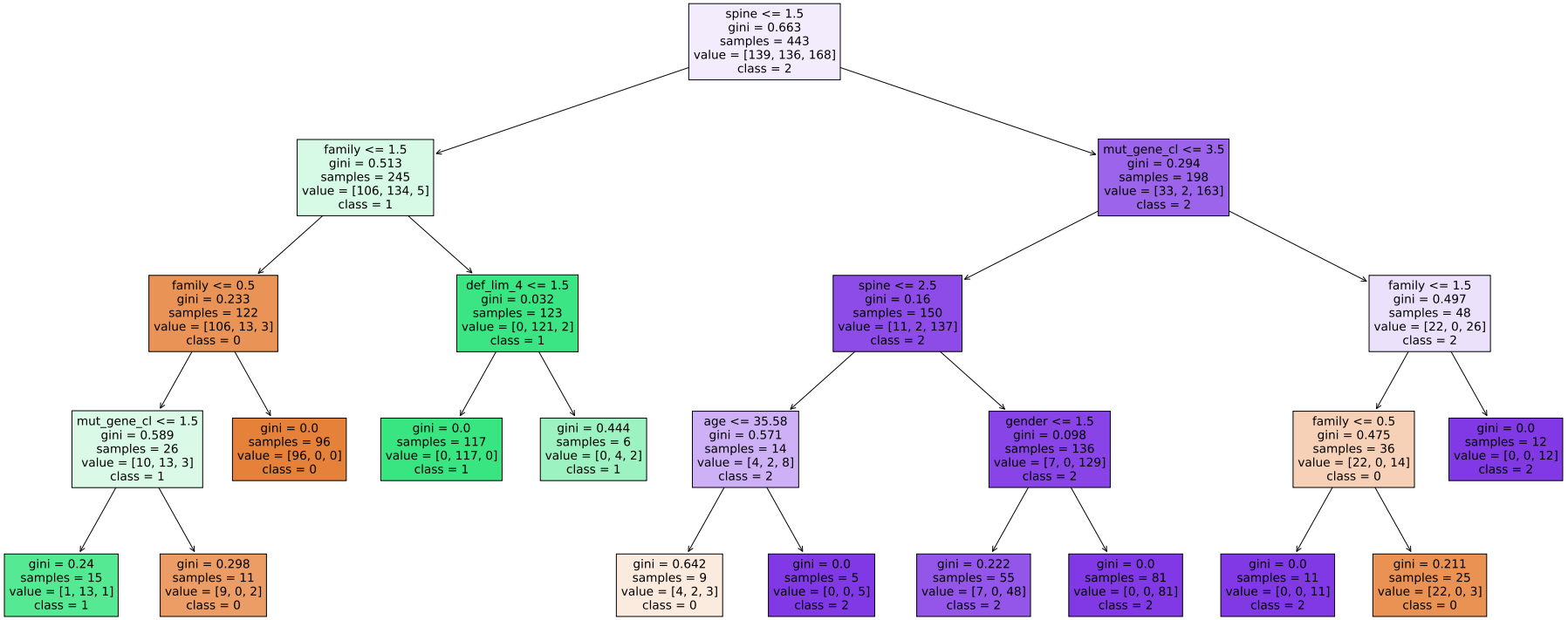


FIGURE 3.39: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data II (selection 1).

3.3.4.2 Selection 2 (COL1 positive patients)

This second subset includes only patients with positive COL1 gene mutation. In this case the samples size is reduced to 326 while all the features shown in in Table 3.11 were considered. Here, the samples are still mixed together but two clusters clearly emerge from the t-SNE two-dimensional projection (see Fig. 3.40). The two clusters identified by k-means (Fig. 3.41) are explained by the categorical variable mutation effect. In this case the zeros cluster is defined by the patients with a causative qualitative mutation effect on the protein (mut effect ≤ 1.5), while ones are those with a quantitative (mut effect > 1.5) mutation effect (see Fig. 3.42).

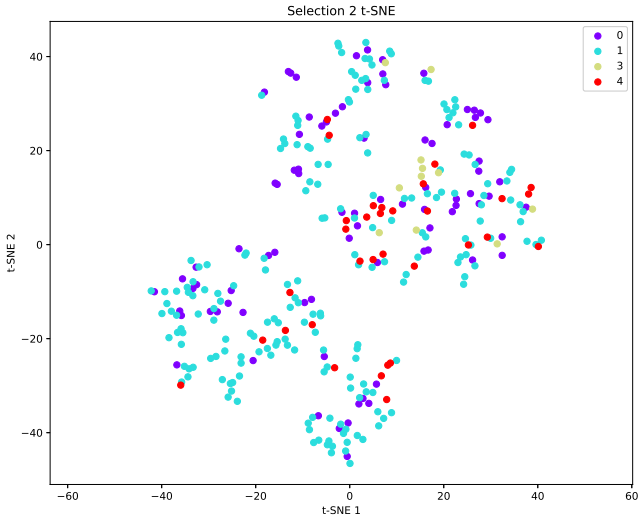


FIGURE 3.40: t-SNE two dimensional projection of osteogenesis data II (selection 2). Each number represents one of the OI types.

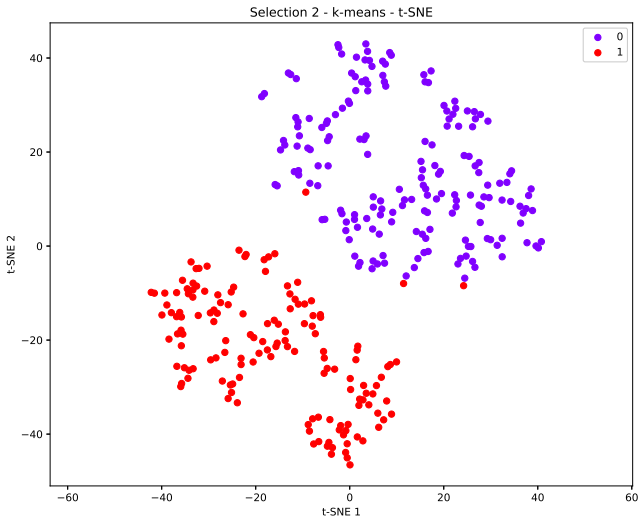


FIGURE 3.41: t-SNE two dimensional projection of clustering on osteogenesis data II (selection 2). Zeros and ones represent the two clusters classes identified by k-means.

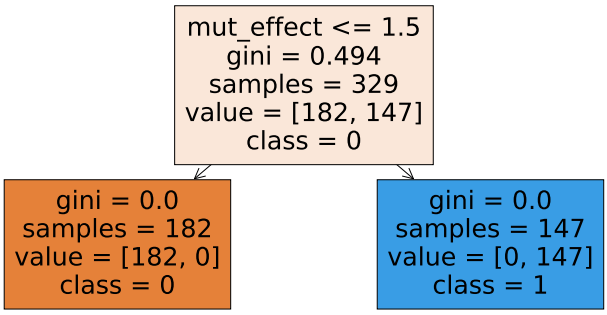


FIGURE 3.42: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data II (selection 2).

3.3.4.3 Selection 3 (adults)

Given possible differences in diagnosis aspects between children and adults, one can analyse these two categories separately. When considering adults, only patients older than 18 have been included and the number of samples was 210 with 24 features. In this case, from t-SNE projection it is possible to clearly visualise three different clusters (Fig. 3.43). As before, using k-means one is able to group patients (Fig. 3.44) according to the rule described by the classification tree (Fig. 3.45). Here, cluster zero includes patients with trunk limitations or deformities (def_lim_4 > 1.5), cluster 1 includes patients with no trunk deformities or limitations (def_lim_4 $\leq$ 1.5) and with a familiarity with OI (family > 1.5), while cluster two is composed of patients with no trunk deformities or limitations (def_lim_4 $\leq$ 1.5) and without familiarity with OI (family $\leq$ 1.5).

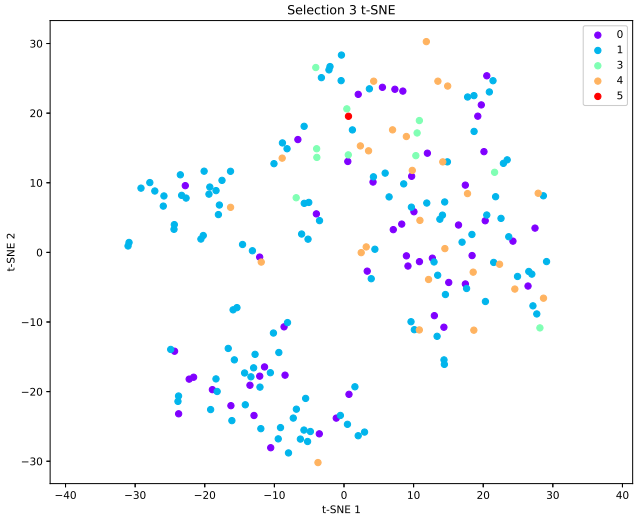


FIGURE 3.43: t-SNE two dimensional projection of osteogenesis data II (selection 3). Each number represents one of the possible types of OI.

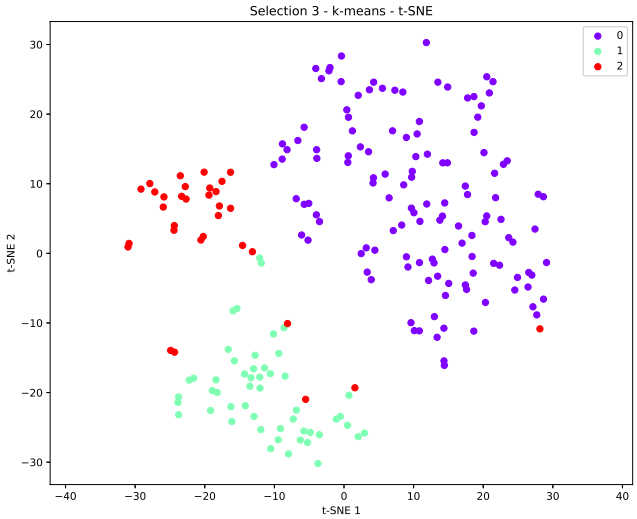


FIGURE 3.44: t-SNE two dimensional projection of clustering on osteogenesis data II (selection 3). Zero, one and two represent the three clusters classes identified by k-means.

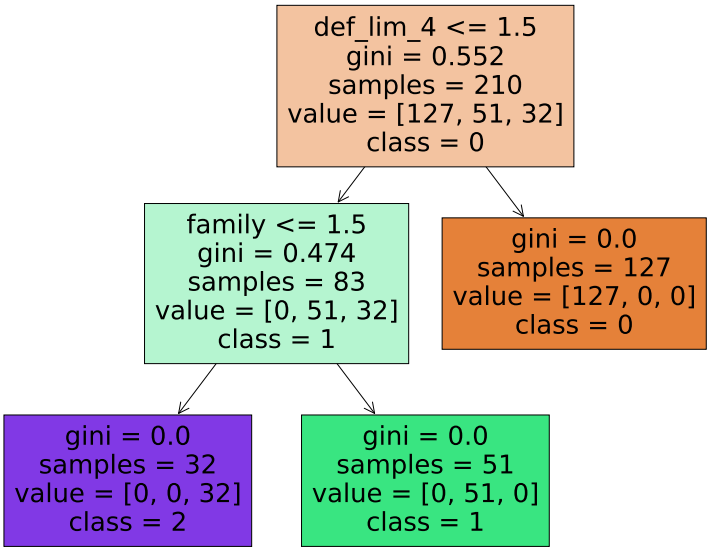


FIGURE 3.45: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data II (selection 3).

3.3.4.4 Selection 4 (children)

The last analysed subset includes children only (age less than 18 years) with a sample size of 233 and 24 features. In this case the t-SNE plot (Fig. 3.46) does not show a clear separation as in the previous cases. By using k-means and a classification tree it was possible to identify two clusters (Fig. 3.48): the classification is less simple than the other cases but the two most important features are the familiarity with OI and mutation gene CL (Fig. 3.48).

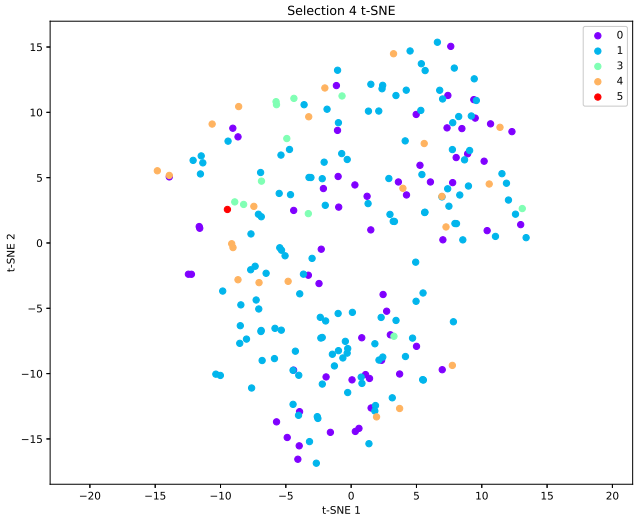


FIGURE 3.46: t-SNE two dimensional projection of osteogenesis data II (selection 4). Each number represents one of the possible types of OI.

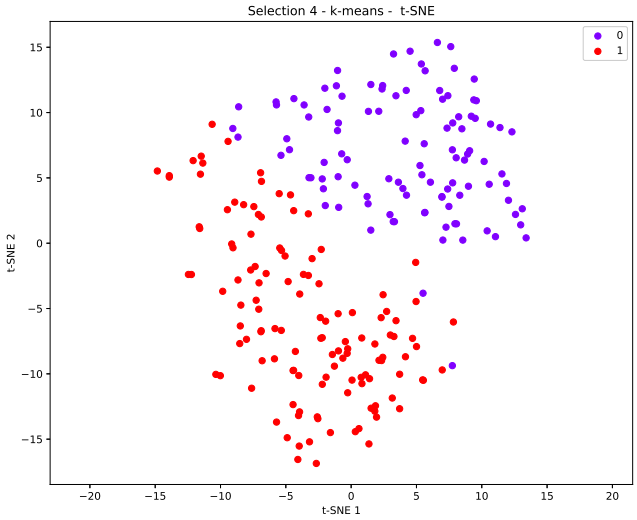


FIGURE 3.47: t-SNE two dimensional projection of clustering on osteogenesis data II (selection 4). Zeros and ones represent the two clusters classes identified by k-means.

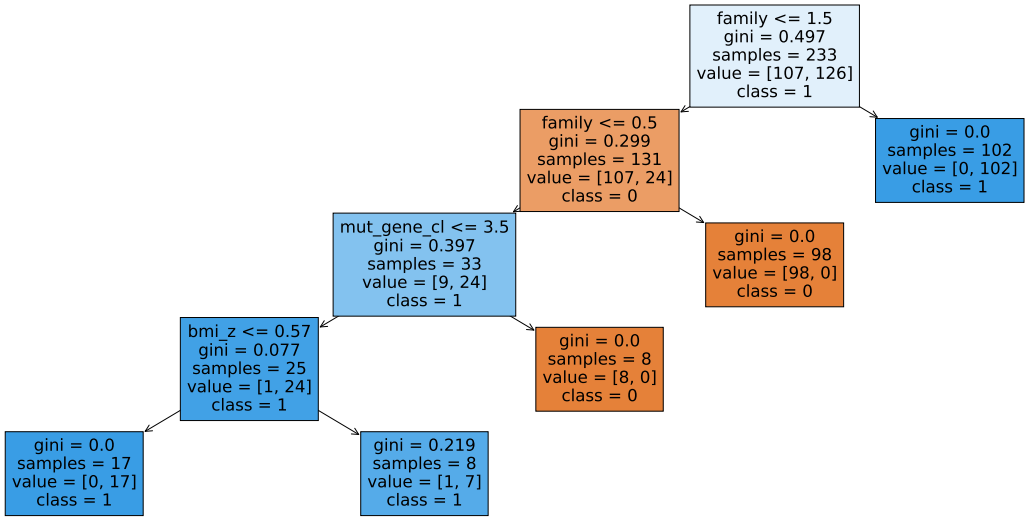


FIGURE 3.48: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis data II (selection 4).

3.3.4.5 Extended data set and supervised analysis

What emerges from the above shown results is that all the clusters were determined by binary or categorical variables. In fact, the data set is mainly composed of this kind of features and only few of them are continuous or have ordinal values. Categorical features also have to be treated with techniques such as one-hot encoding which further increases the dimensionality of the data set introducing other binary columns. The large number of categorical variables highly affects the unsupervised analysis since they tend to create trivial clusters among the samples: these clusters do not necessarily help to identify alternative OI classification but they simply group patients according to a specific characteristic (e.g. those with OI familiarity and those without it). By removing some of the categorical features, such as the genetic ones which seem to be the most relevant in creating clusters, the results do not change very much since the unsupervised methods simply identify the strongest binary variable (among the remaining ones) which can group the samples. To solve this issue, one can edit the original data set (where possible) by extending some of the binary features considered before. The idea comes from the need of physicians to simplify the data and convert them to binary/categorical while some of this information was originally expressed as continuous or ordinal values achieving a total of 45 variables (the sample size remained the same as in the previous sections). However, introducing new features did not solve the issue: the remaining categorical features dominate over the others and they were still selected when performing the clustering process. Fig. 3.49 and 3.50 show the t-SNE two-dimensional projection of the selection 1 data with the extended features and the related two clusters. As one can see in Fig. 3.51, the categorical features related to familiarity and mutation gene CL keep to be the first identified by the classification tree.

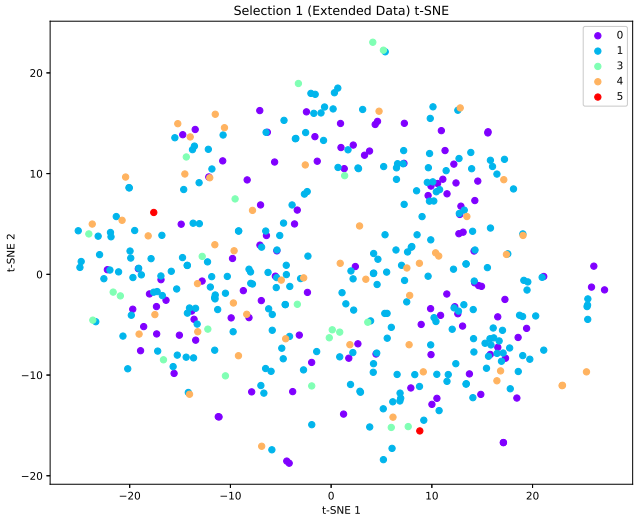


FIGURE 3.49: t-SNE two dimensional projection of osteogenesis extended data II (selection 1). Each number represents one of the possible types of OI.

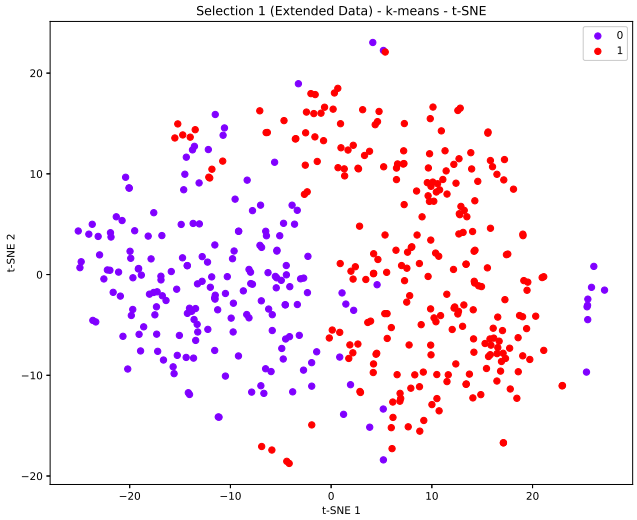


FIGURE 3.50: t-SNE two dimensional projection of clustering on osteogenesis extended data II (selection 1). Zeros and ones represent the two clusters classes identified by k-means.

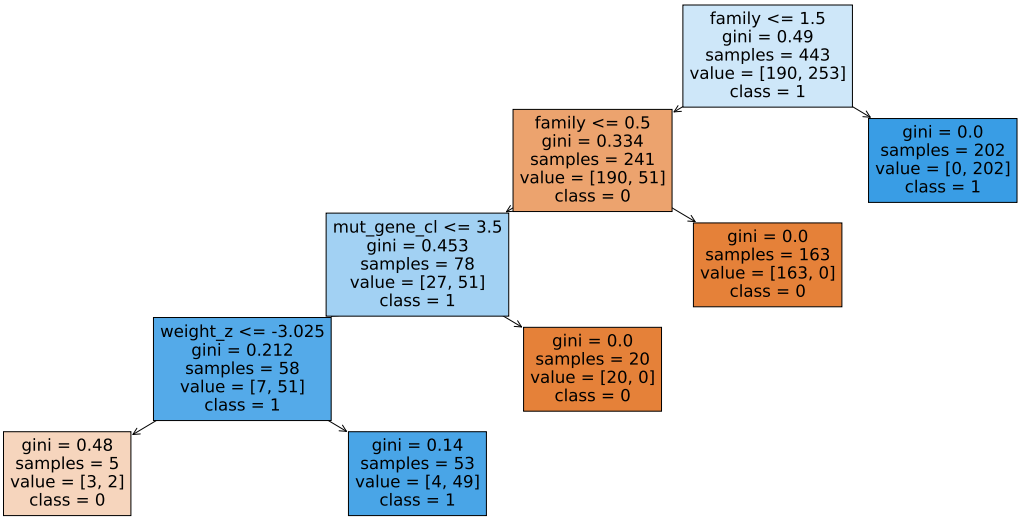


FIGURE 3.51: Classification tree rules based on the labels induced by k-means algorithm on osteogenesis extended data II (selection 1).

Therefore, a supervised approach was used to analyse the data: in this case the features are selected independently of their kind (binary, categorical, continuous or ordinal). The only thing that matters is whether they can explain the label and if they are categorical or continuous is not as much relevant as in the unsupervised case. By doing so, one can also verify the consistency of the original labels (OI types) with the algorithms prediction. In a first instance, an attempt to classify all the patients according to the five available labels (OI types I, III, IV and V plus the class 0) was made, but the multi-label classification did not give satisfying results, even removing the class 0 which introduces some noise in the classification process. Then, a binary approach with all the possible classes combinations has been adopted. OI type V had to be excluded since it only included 2 samples. Discrimination between class 0 and other OI types did not reach significant results. The same cannot be said about OI type I versus type III and IV. The analysis was performed by splitting the data set into training (75%) and test set (25%) with a total of 280 samples for the I vs III case and 307 for the I vs IV case. For each model employed the best hyper-parameters have been selected by using a grid-search approach with a 3-times repeated 2-fold cross-validation. In this case, 2 folds have been chosen in order to have the minority classes (III and IV) equally distributed for each training and validation fold. Tables 3.12 and 3.13 show the values of accuracy, balanced accuracy and AUC for the algorithms employed for the analysis namely elastic net, linear SVM and random forest while their best hyper-parameters are shown in Tables 3.14 and 3.15. Both values of accuracy and balanced accuracy have been shown since the classes are highly unbalanced: type III and IV samples are much less than type I with 22, 49 and 258 samples respectively. As one can see, the algorithms got better performance when classifying types I and III rather than types I and IV. On the other hand, it was not possible to successfully discriminate type III from IV: these two OI types are more similar to each other than type I.

TABLE 3.12: Accuracy, balanced accuracy and AUC values for OI types I and III classification.

Methods	OI Type I Versus Type III		
	Accuracy	Balanced Accuracy	AUC
Logistic Elastic Net	0.90	0.76	0.91
Linear SVM	0.80	0.89	0.92
Random Forest	0.96	0.70	0.98

TABLE 3.13: Accuracy, balanced accuracy and AUC values for OI types I and IV classification.

Methods	OI Type I Versus Type IV		
	Accuracy	Balanced Accuracy	AUC
Logistic Elastic Net	0.90	0.73	0.74
Linear SVM	0.75	0.68	0.68
Random Forest	0.90	0.67	0.87

TABLE 3.14: Supervised algorithms best hyper-parameters for OI extended data II - type I vs III classification. The table shows the best hyper-parameters identified by a grid search cross-validation for each of the applied models.

OI Type I Versus Type III		
Logistic Elastic Net	**Linear SVM**	**Random Forest**
C: 199.53	C: 0.10	No. of trees: 300.00
L1 ratio: 0.10	Balanced weights	Split criterion: Entropy
		Max depth: 10
		Max features: Square root

TABLE 3.15: Supervised algorithms best hyper-parameters for OI extended data II - type I vs IV classification. The table shows the best hyper-parameters identified by a grid search cross-validation for each of the applied models.

OI Type I Versus Type IV		
Logistic Elastic Net	**Linear SVM**	**Random Forest**
C: 25.12	C: 0.10	No. of trees: 300.00
L1 ratio: 0.90	Balanced weights	Split criterion: Entropy
		Max depth: 10
		Max features: Square root

In addition to this, feature selection was also performed for each of the employed methods. As shown in Tables 3.16 and 3.17, some of the categorical variables appear among the most important ones. Nonetheless, in this case, their presence is not trivial since it specifically implies that they are useful for the class prediction. Moreover, other features (which were not considered by the clustering approach) seem to be relevant for the classification and now appear in the first positions. When it comes to the consensus of feature selection, one can see that each technique selected different features. However, some of these appear in all the three lists, such as height Z, weight Z, familiarity and mutation gene CL in the I versus III case, while height Z only is selected by all the methods in the I versus IV case. These results show that even in this case where the number of features is not very high, it is hard to identify a stable set of features.

TABLE 3.16: List of the ten most important features in OI types I and III classification.

OI Type I Versus Type III		
Logistic Elastic Net	**Linear SVM**	**Random Forest**
Height Z	Sclera	Height Z
BMI Z	Height Z	Weight Z
Dent Oth	Family	Age
Def lim dett 33	Dent Oth	BMI Z
Weight Z	Vert	Def lim dett 42
Mut Gene CL	Weight Z	Family
Facial Oth	Def lim dett 41	Sclera
Def lim dett 11	Facial maxil	Fractures
Def lim dett 32	Deafness neuro	Def lim dett 41
Family	Mut gene CL	Mut gene CL

TABLE 3.17: List of the ten most important features in OI types I and IV classification.

OI Type I Versus Type IV		
Logistic Elastic Net	**Linear SVM**	**Random Forest**
Def lim dett 31	Facial front	Height Z
Height Z	Def lim dett 42	Age
Fractures	Fractures	BMI Z
Def lim dett 22	Dent di	Weight Z
Def lim dett 32	Facial triang	Fractures
Age	Height Z	Def lim dett 42
Valv aort	Deafness condu	Facial front
Def lim dett 51	Def lim dett 41	Dent di
Valv polm	Weight Z	Facial triang
Def lim dett 23	Def lim dett 34	Bone dens

3.4 Oxford Street

In this section, various issues faced when analysing the Oxford Street II multiomics data set are presented and discussed. In particular, the analysis was not limited to the simple disease prediction but also to the investigation of the feature selection, its stability, possible confounders, the biomarkers identification as well as the disease progression.

3.4.1 The Oxford Street II data set

The Oxford Street II data set was originally included in the EXPOsOMICS project [92] with the aim to evaluate the effects of exposure to air pollution on subjects affected by different diseases. The exposome can be defined as the totality of exposure from both internal and external sources over a complete lifetime. These exposures can include radiation, chemical and biological agents, other exposures in general as well as the impact of socio-economic position and social relations they can have on health [93]. The data underlying the results presented in this section are available on request from the International Agency for Research on Cancer (IARC) and their use and availability is regulated by the Exposomics Steering Board and the IARC Ethical Committee. A total of 59 volunteers were recruited for the experiment: 19 of these were affected by ischemic heart disease (IHD) with normal lung function, 20 were affected by chronic obstructive pulmonary disease (COPD) with no previous IHD events, and 20 were in healthy status without history of COPD or IHD. The subjects with high level of exposure to traffic-related air pollution (TRAP) were not included in the experiment as well as smokers and people who quit smoking in the previous year. All the patients affected by IHD and COPD were selected from existing databases of outpatient respiratory and cardiology clinics at the Royal Brompton and Harefield NHS Foundation Trust [94]. In order to evaluate the exposure levels, each participant had to walk within two distinct locations in London characterised by different levels of air pollution: Hyde Park and Oxford Street. Moreover, for each subject, a total of six blood samples were collected two hours before, two hours after and 24 hours after each walk in both locations. In addition to this, other information related to various omics such as adductomics, miRNA, transciptomics and metabolomics has been collected. The data set also included general information about the patients namely body mass index (BMI), age, sex, diet, distance walked, blood pressure, medication use and TRAP measurements. The four omics used for the analysis will be now briefly described.

Adductomics Adductomics is the study of DNA adducts. DNA adducts are compounds that can form bonds with DNA. Although enzymes can repair the adducts, some of these are not repaired and they can cause mutations during the cell division [95]. DNA adducts have acquired importance in the exposome field, since their

study allows to identify exposures to dangerous chemicals in the environment possibly helping mitigate the exposure itself and reduce the risk of disease and cancer [96].

miRNA MicroRNAs (miRNA) are short non-coding RNA molecules (about 22 nucleotides) with two main functions: RNA silencing and gene expression regulation at post-transcriptional level [97], thus miRNA are important for gene expression maintenance and stability. Recent studies have shown that miRNA profiles can be changed by exposure to air pollutants. Therefore, miRNA study can help understanding gene expression mechanisms altered by ambient pollutants [98].

Transcriptomics The transcriptome can be defined as the set of gene transcripts transcribed in a specific cell type, organism or tissue, including both coding RNA (translated to proteins) and non-coding RNA which influenced gene expression. Changes in gene transcripts have been object of study in order to get information about biological mechanisms and pathways and can thus be used to get better insights on diagnostic and therapeutic targets [99].

Metabolomics Metabolomics is the study of processes involving metabolites, the products of metabolism. The aim of metabolomics is to identify and quantify within cells, biofluids, tissues or organisms. The study of metabolites is useful to understand the effects of air pollution leading to health outcomes [100]. Metabolomics is also applied for biomarker discovery and to identify and understand the mechanisms leading to different physiological conditions and diseases. There are two main methods to extract metabolites information: untargeted and targeted mass spectrometry. The first one measures the metabolites present in a sample without a priori information on the metabolome. The second one has higher sensitivity but, being based on a priori information, it can only be used to study specific metabolites and metabolic pathways [101].

For this analysis the original scope of the data set was changed. Instead of evaluating the effects of air pollution on the patients, the available data were used to predict the patients status of health. In particular one wants to evaluate whether omics features can be used to classify subjects according to their diagnosis and simultaneously identify which of these features can be used to understand the disease and thus support the diagnosis process. In the following sections all the steps of the procedures and all the issues faced when dealing with these data will be discussed.

3.4.2 Preliminary analysis results

The first part of the analysis involved the use of omics data only to predict the patients illness status. In particular, having to deal with three possible statuses (COPD, IHD and healthy) each disease was approached as a binary classification problem: healthy versus IHD and healthy versus COPD, considering both supervised and unsupervised techniques in both cases. Considering a total of 59 patients, and six blood samples for each, the two data sets should be composed of 240 samples in the case of COPD and healthy subjects and 234 for IHD and healthy ones. However not all the omics data were available for each patient. Moreover few of them also dropped the experiments and they may lack some information. Due to this, the data set size is smaller according to the case and the considered omic. After using t-SNE as dimensionality reduction algorithm to have better insight on the data distribution, for each case a supervised classification was performed using four different algorithms namely lasso, elastic net, random forest and RFE with linear SVM. These algorithms have been exploited in order to make diagnosis prediction as well as identifying the most relevant features. For each analysis the data set was split in two parts: 80% was used as training set while the remaining 20% was used as test set. A 2-fold cross-validation was also performed on the training set for hyper-parameters tuning. Since more than one sample was available for each subject, both training-test splitting and cross-validation have been performed making sure that each sample belonging to the same patient was in the same set/fold. Scikit-learn [70] library has been used to implement the known machine learning algorithms.

3.4.2.1 Healthy versus IHD

In this first analysis the aim was to classify patients according to the binary labels IHD and healthy using omics data only, first separately (i.e. using one omic per time) and then using all the omics together. The results for each analysis are shown in the following paragraphs.

Adductomics In this first scenario adductomics data only have been used. The number of available samples is 198 while the number of adducts is 32. As one can see from a first two-dimensional projection using t-SNE in Fig. 3.52, the two classes do not seem well separated. This result is also confirmed by a supervised analysis performed using different classification algorithms. As shown in Fig. 3.53, most of the algorithms did not perform well on the test set with lasso and elastic net performing worse than random choice. The models best hyper-parameters identified via 2-fold grouped cross-validation are reported in Table 3.18.

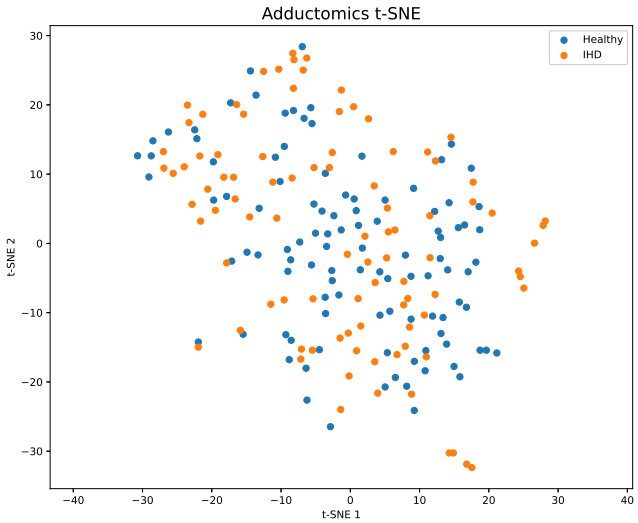


FIGURE 3.52: t-SNE two dimensional projection of IHD and healthy samples using adductomics data. Each label represents one of the two diagnoses.

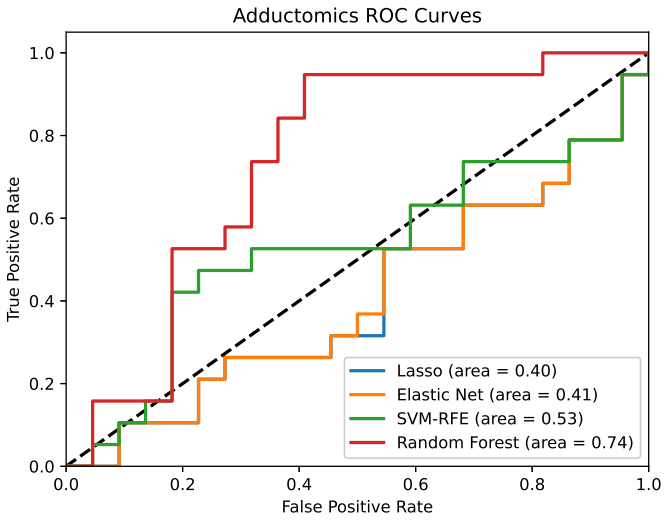


FIGURE 3.53: ROC curve for each of the classification algorithms employed for the prediction of IHD using adductomics data.

TABLE 3.18: Supervised algorithms best hyper-parameters for the prediction of IHD using adductomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0001	C: 0.0001	C: 100	No. of trees: 100
	L1 ratio: 0.1	Selected features: 10	Split criterion: Entropy
		Step: 1	Max depth: 5
			Max features: Square root
			Min samples leaf: 4
			Min samples split: 20

miRNA The second analysis involved the use of miRNA data only. In this case the number of samples was 191 while the number of miRNAs was 365. As shown in Fig. 3.54, the data is still not well separated, however both IHD and healthy samples seem grouped together. Considering that this is just a two-dimensional projection, it might be that the data is better separated in the original space. Fig. 3.55 shows the ROC curve of each classification algorithm employed. In this case, all the algorithms achieved an AUC over 75% showing good prediction performance. The related best hyper-parameters identified by 2-fold cross-validation are shown in Table 3.20. Given the improved classification performance, as a further step, the most important features were also considered. Table 3.20 shows the five most important features for each classification technique. As one can see, most of the selected features is different and the common ones are very few and no one has been selected for all the four models.

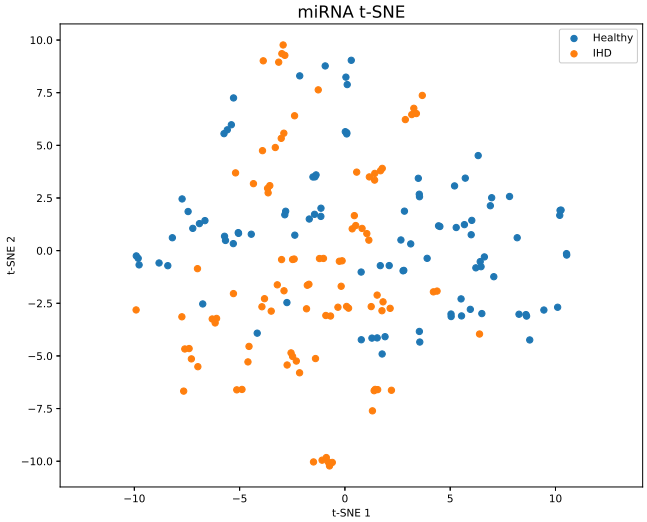


FIGURE 3.54: t-SNE two dimensional projection of IHD and healthy samples using miRNA data. Each label represents one of the two diagnoses.

TABLE 3.19: Supervised algorithms best hyper-parameters for the prediction of IHD using miRNA data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0199	C: 0.3981	C: 0.00794	No. of trees: 100
	L1 ratio: 0.1	Selected features: 50	Split criterion: Entropy
		Step: 1	Max depth: 3
			Max features: Square root
			Min samples leaf: 20
			Min samples split: 2

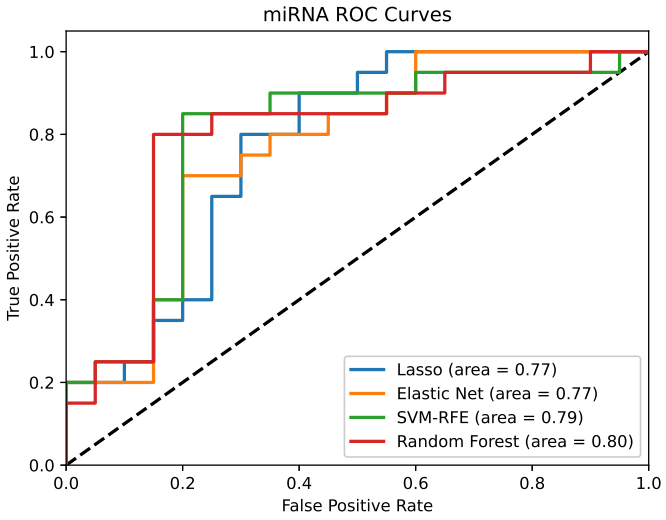


FIGURE 3.55: ROC curve for each of the classification algorithms employed for the prediction of IHD using miRNA data.

TABLE 3.20: List of the five most important features selected by classification algorithms on miRNA data.

Lasso	Elastic Net	SVM-RFE	Random Forest
hsa-miR-4672	hsa-miR-5787	hsa-miR-4787-5p	hsa-miR-296-5p
hsa-miR-374b-5p	hsa-miR-4788	hsa-miR-5787	hsa-miR-423-3p
hsa-miR-6510-5p	hsa-miR-1914-3p	hsa-miR-6069	hsa-miR-2861
hsa-miR-16-2-3p	hsa-miR-16-2-3p	hsa-miR-328	hsa-miR-625-5p
hsa-miR-186-5p	hsa-miR-335-5p	hsa-miR-16-2-3p	hsa-miR-200c-3p

Transcriptomics The third analysis made use of transcriptomics data. In this case the number of features is much higher (30923) while the number of samples is 191. The t-SNE projection shows an interesting result: as shown in Fig. 3.56 few clusters (different from the diagnosis type) can be identified. By using k-means and the elbow criterion one is able to select 5 as the number of optimal clusters (see Fig. 3.57). A classification tree has been used to explain the formed groups of patients and the features describing the clusters are shown in Fig. 3.58. The first split is determined by the value of A_33_P3312182 being less or greater than 6.766, followed by four other RNA transcripts and their related thresholds, namely A_23_P118289, A_23_P259357, A_33_P3216945 and A_23_P399078. When it comes to the supervised analysis, three out of four algorithms scored good performance results, with the exception of random forest as shown in Fig. 3.59. The models best hyper-parameters are shown in Table 3.21. As before, the most important features did not achieve high consensus between models (see Table 3.22).

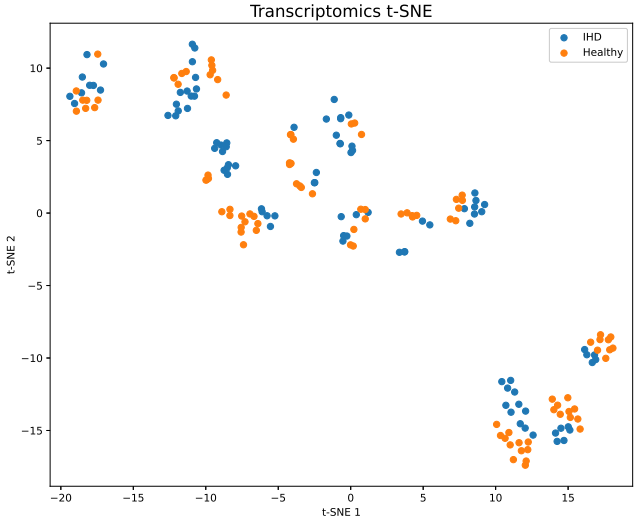


FIGURE 3.56: t-SNE two dimensional projection of IHD and healthy samples using transcriptomics data. Each label represents one of the two diagnoses.

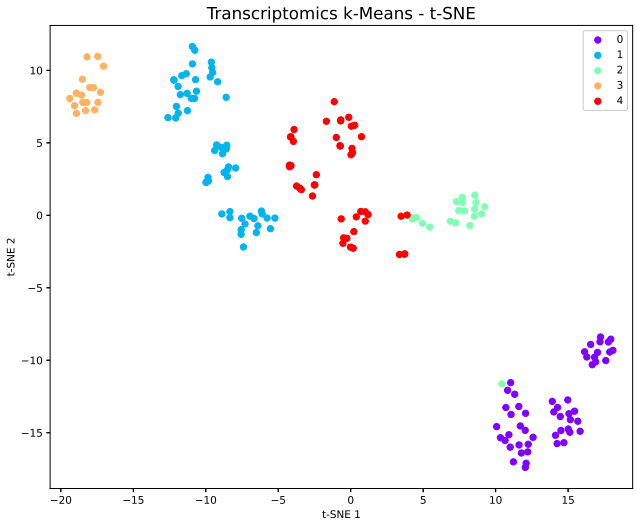


FIGURE 3.57: t-SNE two dimensional projection of clustering on IHD and healthy samples with transcriptomics data. The numbers from zero to four represent the five clusters classes identified by k-means.

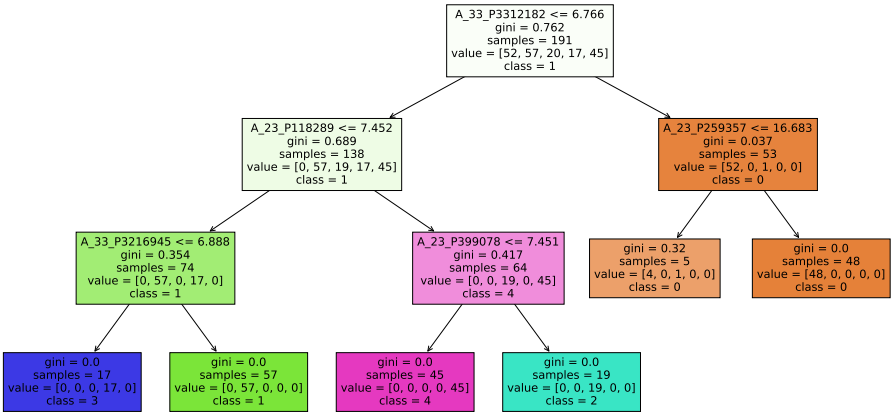


FIGURE 3.58: Classification tree rules based on the labels induced by k-means algorithm on IHD and healthy samples using transcriptomics data.

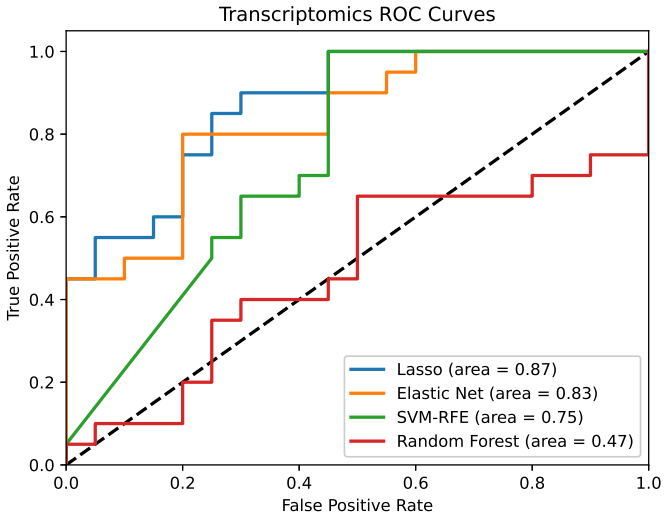


FIGURE 3.59: ROC curve for each of the classification algorithms employed for the prediction of IHD using Transcriptomics data.

TABLE 3.21: Supervised algorithms best hyper-parameters for the prediction of IHD using transcriptomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.2511	C: 0.001	C: 0.0063	No. of trees: 3000
	L1 ratio: 0.10	Selected features: 20	Split criterion: Entropy
		Step: 300	Max depth: 5
			Max features: Square root
			Min samples leaf: 1
			Min samples split: 10

TABLE 3.22: List of the five most important features selected by classification algorithms on IHD and healthy samples with transcriptomics data.

Lasso	Elastic Net	SVM-RFE	Random Forest
A_23_P324384	A_33_P3363260	A_21_P0006538	A_33_P3410296
A_33_P3379396	A_33_P3659876	A_19_P00329511	A_21_P0012911
A_33_P3398143	A_21_P0008515	A_33_P3280521	A_23_P210176
A_33_P3353921	A_32_P4403	A_24_P55148	A_33_P3367102
A_24_P55148	A_32_P149251	A_33_P3216532	A_21_P0000069

Metabolomics The forth analysis regarded the use of metabolomics data to predict IHD status. The data set was composed of 212 patients and 10067 features. As shown in Fig. 3.60 the two classes are almost perfectly separated. This result is also confirmed by the performances of the classification algorithms employed as shown in Fig. 3.61, while the related best hyper-parameters identified by 2-fold cross-validation are reported in Table 3.23. Despite the almost perfect AUC for all the four models, the most important features still lack of consensus between models (see Table 3.24), pointing to the need to find a way to select features in a robust way.

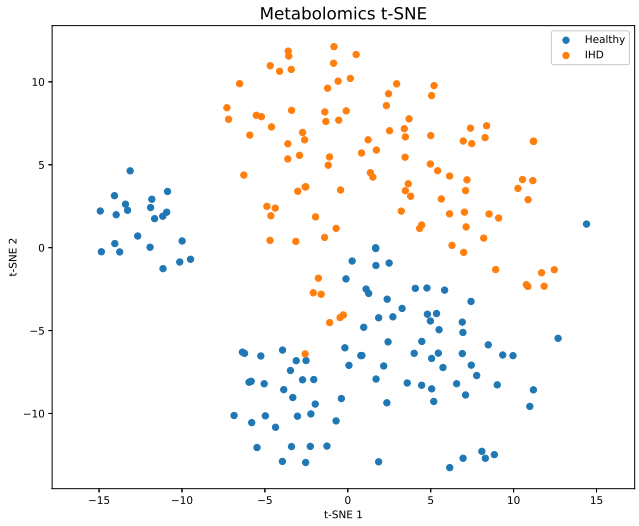


FIGURE 3.60: t-SNE two dimensional projection of IHD and healthy samples using metabolomics data. Each label represents one of the two diagnoses.

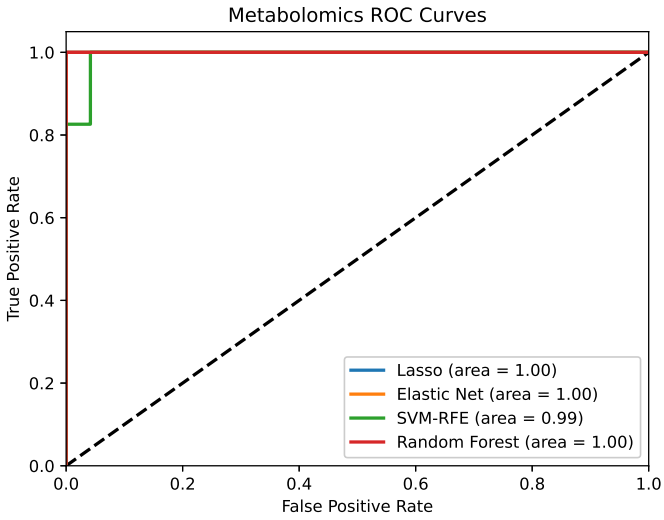


FIGURE 3.61: ROC curve for each of the classification algorithms employed for the prediction of IHD using metabolomics data.

TABLE 3.23: Supervised algorithms best hyper-parameters for the prediction of IHD using metabolomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0079	C: 0.001	C: 0.02	No. of trees: 1000
	L1 ratio: 0.3	Selected features: 10	Split criterion: Entropy
		Step: 100	Max depth: 3
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 10

TABLE 3.24: List of the five most important features selected by classification algorithms on IHD and healthy samples with metabolomics data.

Lasso	Elastic Net	SVM-RFE	Random Forest
410.1254@6.34219	410.1254@6.34219	410.1254@6.34219	278.0679@2.31897
623.4388@7.92532	342.0079@5.42850	627.2748@6.90690	577.2746@7.01407
375.3123@7.18853	623.4388@7.92532	486.3169@6.68188	403.3652@7.72745
208.1079@5.18358	431.3044@5.88208	803.045@8.78652	952.6821@7.67908
803.045@8.78652	629.8494@4.68212	770.4687@7.00565	311.1232@2.33333

Multiomics The final analysis involved the use of all the four omics together (167 samples and 41387 features). As one can see from Fig. 3.62 all the models scored very good as in metabolomics case (the models best hyper-parameters are shown in Table 3.25). However, as shown in Table 3.26 almost all the most important features come from metabolomics data, confirming it is the best omic to predict IHD status. Nevertheless, the consensus between the selected features, is still very low. Moreover, one should also consider that the number of metabolites and transcripts is much higher than the one related to adducts and miRNAs, thus under a mere probability point of view these two omics have lower chances of being selected.

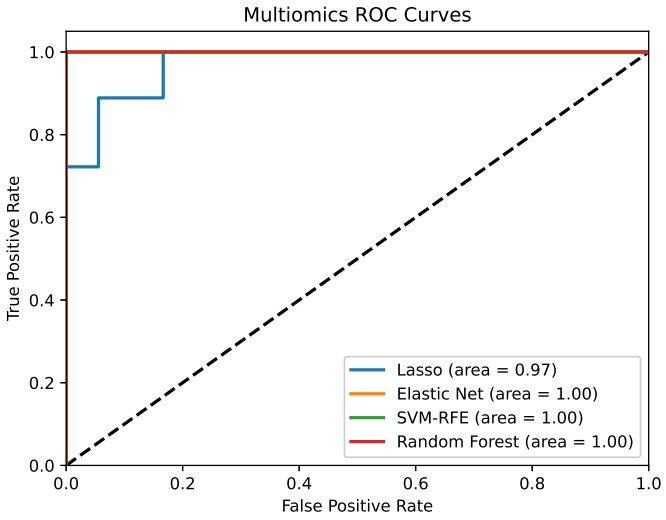


FIGURE 3.62: ROC curve for each of the classification algorithms employed for the prediction of IHD using all the omics data.

TABLE 3.25: Supervised algorithms best hyper-parameters for the prediction of IHD using multiomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0501	C: 0.063	C: 0.01585	No. of trees: 2000
	L1 ratio: 0.10	Selected features: 50	Split criterion: Entropy
		Step: 1000	Max depth: 3
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 2

TABLE 3.26: List of the five most important features selected by classification algorithms on IHD and healthy samples using all the omics data.

Lasso	Elastic Net	SVM-RFE	Random Forest
627.2748@6.90690	A_23_P113701	978.4897@8.41506	1153.6049@7.02360
236.1392@5.43261	666.2666@7.92155	A_23_P113701	A_23_P66402
236.1772@6.19898	403.7799@9.26897	283.1176@5.38265	A_33_P3284004
A_33_P3398143	A_33_P3229918	208.1079@5.18358	221.148@5.92175
368.2579@7.85856	833.9491@6.91451	818.5165@8.69470	A_23_P27381

3.4.2.2 Healthy versus COPD

The second analysis regarded the prediction of the other disease included in the data set, namely COPD. As in the previous section the analysis was made using omics data only, first separately and then using all the omics together. In the following paragraphs the results for each omics are shown.

Adductomics In this first case, adductomics data only was used to predict COPD diagnosis. The data set was composed of 213 samples and 32 adducts. Fig. 3.63 shows a two-dimensional projection of the data. As in the IHD case the two groups of patients appear mixed together. This result is also confirmed by the supervised analysis: the use of adductomics did not allow to reach reasonably accurate results as shown in Fig. 3.64, while the models best hyper-parameters are reported in Table 3.27.

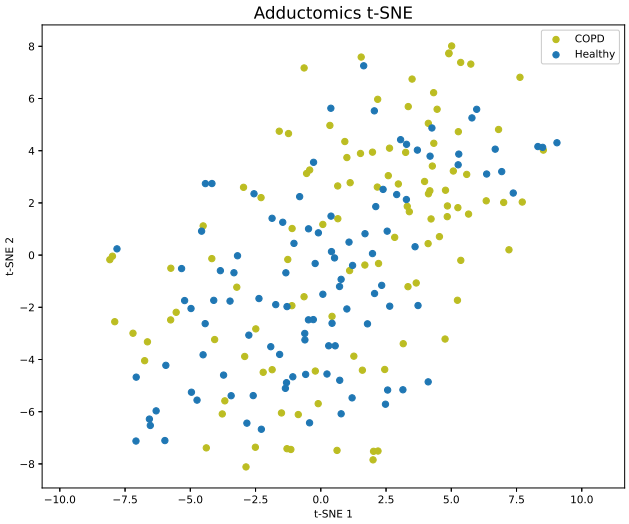


FIGURE 3.63: t-SNE two dimensional projection of COPD and healthy samples using adductomics data. Each label represents one of the two diagnoses.

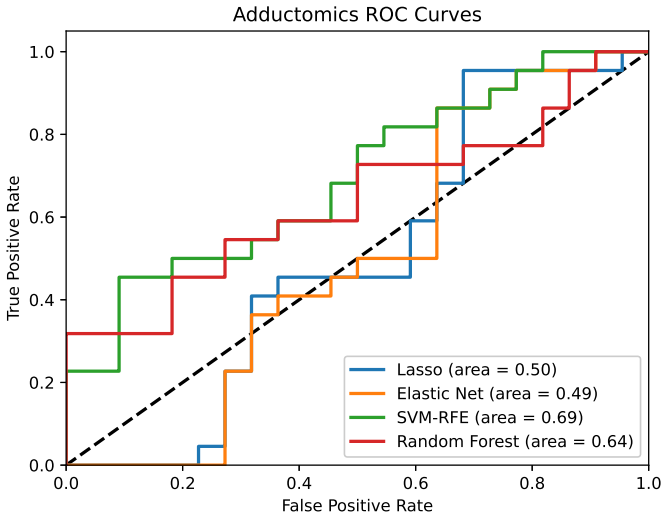


FIGURE 3.64: ROC curve for each of the classification algorithms employed for the prediction of COPD using adductomics data.

TABLE 3.27: Supervised algorithms best hyper-parameters for the prediction of COPD using adductomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.008	C: 0.013	C: 0.316	No. of trees: 1600
	L1 ratio: 0.3	Selected features: 5	Split criterion: Entropy
		Step: 1	Max depth: 3
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 2

miRNA In this case the data set was composed of 198 samples and 365 features. As for adductomics data, the use of miRNA features was not enough to achieve good prediction performance on the COPD and healthy classification (see Table 3.28 for the best hyper-parameters and Fig. 3.65 for the ROC curves). At the same time, as shown in Fig. 3.66, the two-dimensional projection made by t-SNE did not show any presence of alternative clear groups.

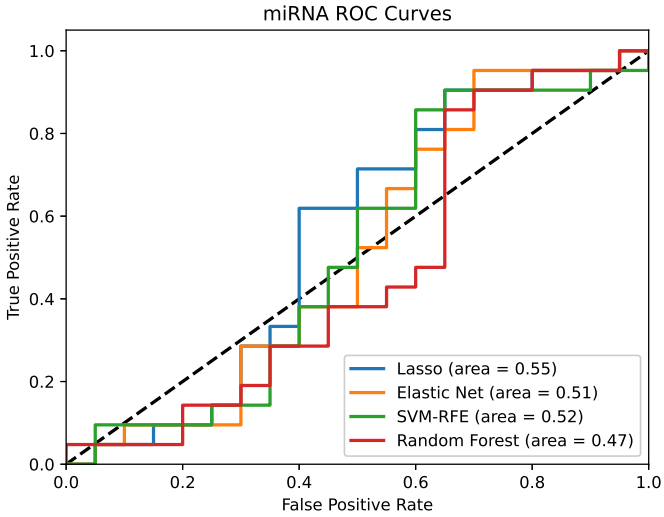


FIGURE 3.65: ROC curve for each of the classification algorithms employed for the prediction of COPD using miRNA data.

TABLE 3.28: Supervised algorithms best hyper-parameters for the prediction of COPD using miRNA data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0001	C: 0.0001	C: 100	No. of trees: 100
	L1 ratio: 0.10	Selected features: 10	Split criterion: Entropy
		Step: 1	Max depth: 5
			Max features: Square root
			Min samples leaf: 4
			Min samples split: 20

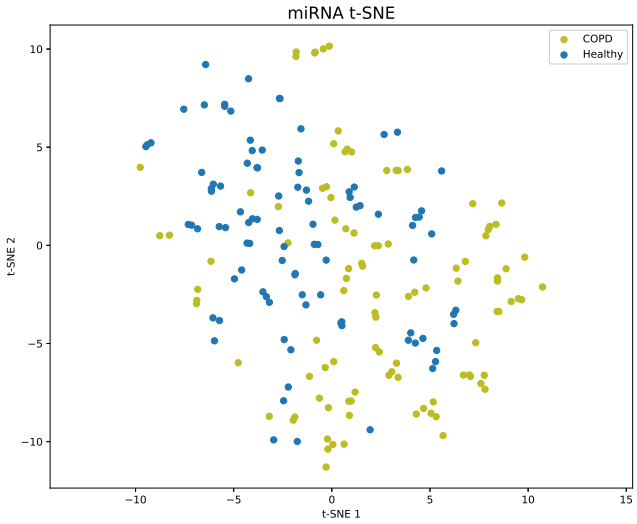


FIGURE 3.66: t-SNE two dimensional projection of COPD and healthy samples using miRNA data. Each label represents one of the two diagnoses.

Transcriptomics The third analysis involved the use of transcriptomics data: here the data set was composed of 200 samples and 30923 features. As in the IHD case the dimensionality is huge and made the classification more difficult as shown by the poor results reported in Fig. 3.67. The models best hyper-parameters are shown in Table 3.29. However, the unsupervised analysis pointed to the presence of six different groups of patients identified by k-means and the elbow method (see Fig. 3.68 and 3.69). Although some samples belonging to different clusters seem to overlap, one should take into consideration that the plot merely represents a 2-dimensional approximation induced by t-SNE. The classification tree reported in fig. 3.70 shows the features describing this new labelling. In this case the tree is more complex than in the IHD case since the number of clusters is bigger. The first split is defined by A_33_P3261953 with a threshold value of 8.803, followed by seven other RNA transcripts including A_23_P93792, A_23_P203994, A_33_P3322909, A_33_P3419785, A_33_P3297277, A_21_P0014703 and A_23_P162874 with their corresponding threshold values.

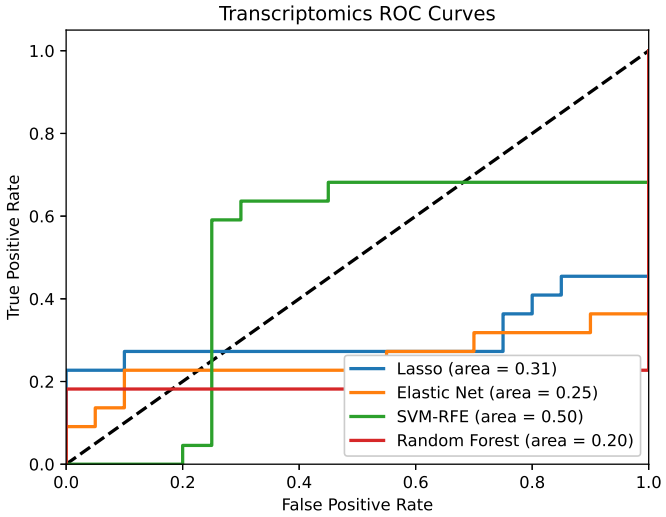


FIGURE 3.67: ROC curve for each of the classification algorithms employed for the prediction of COPD using transcriptomics data.

TABLE 3.29: Supervised algorithms best hyper-parameters for the prediction of COPD using transcriptomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.2511	C: 0.3981	C: 0.0398	No. of trees: 1600.00
	L1 ratio: 0.10	Selected features: 50	Split criterion: Entropy
		Step: 300	Max depth: 3
			Max features: Square root
			Min samples leaf: 20
			Min samples split: 2

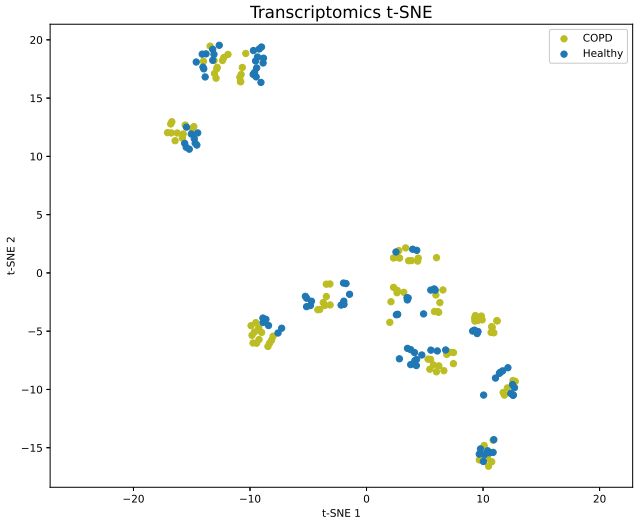


FIGURE 3.68: t-SNE two dimensional projection of COPD and healthy samples using transcriptomics data. Each label represents one of the two diagnoses.

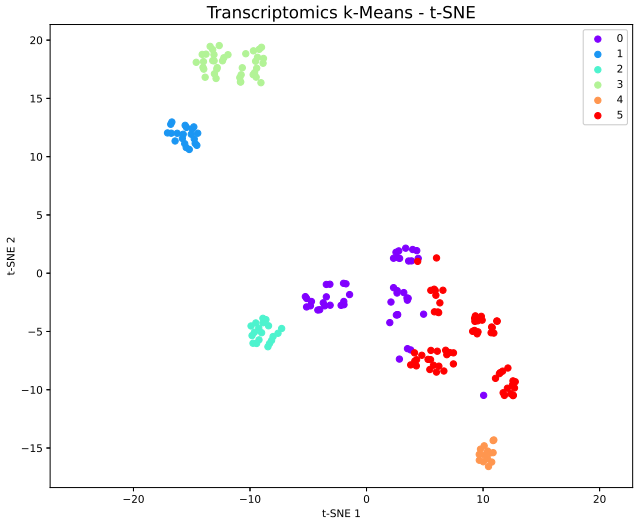


FIGURE 3.69: t-SNE two dimensional projection of clustering on COPD and healthy samples with transcriptomics data. Numbers from zeros to five represent the six clusters classes identified by k-means.

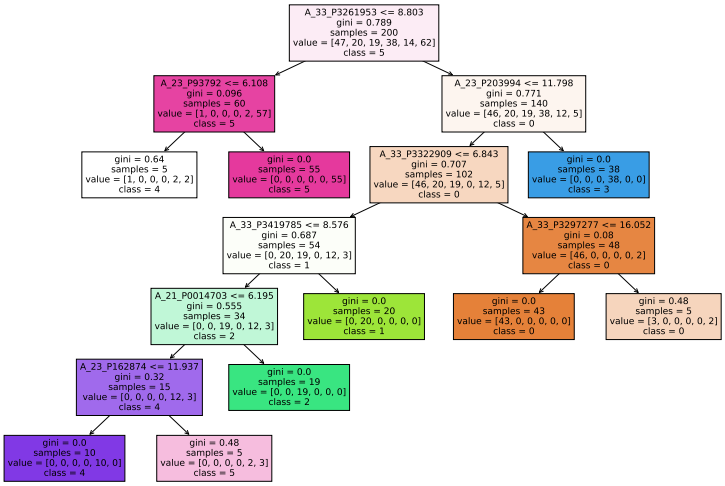


FIGURE 3.70: Classification tree rules based on the labels induced by k-means algorithm on COPD and healthy samples using transcriptomics data.

Metabolomics The forth analysis made use of metabolomics data to predict the COPD diagnosis. The data set size was of 224 samples and 10067 metabolites. In this case the samples do not seem as well separated as in the IHD case (see Fig. 3.71). This result is also confirmed by the poor models prediction performances as shown in Fig. 3.72. Here the results are better than in transcriptomics case but they are much lower than those achieved by the same omic for the IHD prediction. The models best hyper-parameters are shown in Table 3.30.

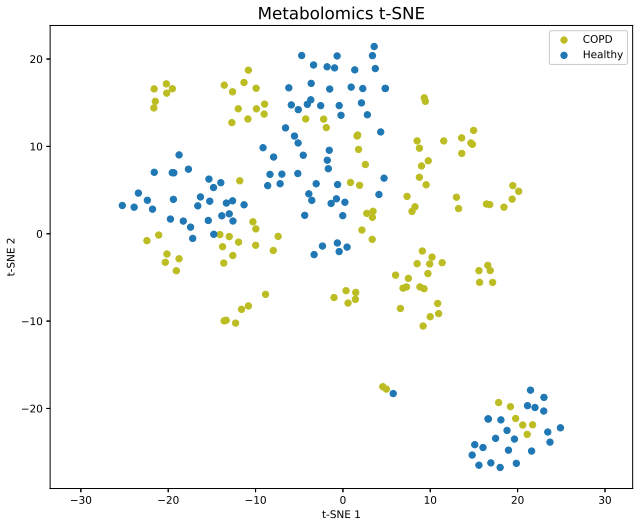


FIGURE 3.71: t-SNE two dimensional projection of COPD and healthy samples using metabolomics data. Each label represents one of the two diagnoses.

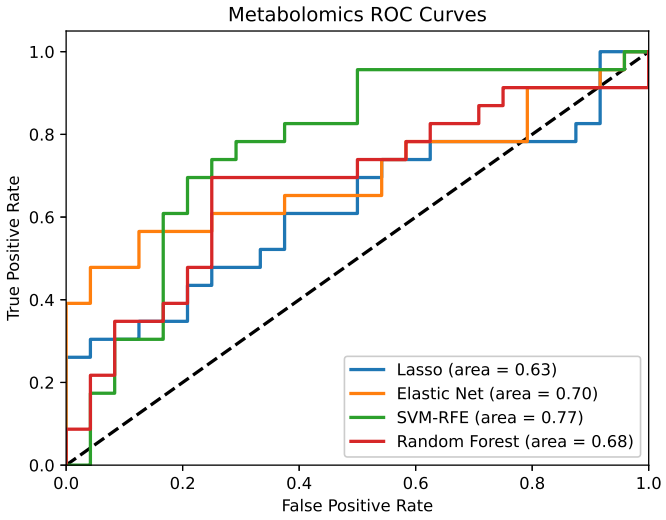


FIGURE 3.72: ROC curve for each of the classification algorithms employed for the prediction of COPD using metabolomics data.

TABLE 3.30: Supervised algorithms best hyper-parameters for the prediction of COPD using metabolomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.005	C: 0.0794	C: 0.7943	No. of trees: 2000
	L1 ratio: 0.1	Selected features: 10	Split criterion: Entropy
		Step: 100	Max depth: 5
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 2

Multiomics Finally, the use of all the four omics was made for the COPD prediction. The data set size is of 174 samples and 41387 features. Unfortunately in this case the prediction performances are very low probably due to the large amount of noise introduced by low-power predictive omics. Fig. 3.73 shows the ROC curves for the applied models: three out of four models (namely lasso, elastic net and SVM) slightly outperform random choice while random forest's AUC does not reach the threshold of 0.5. Table 3.31 shows the best hyper-parameters identified by a 2-fold grouped cross-validation for each model.

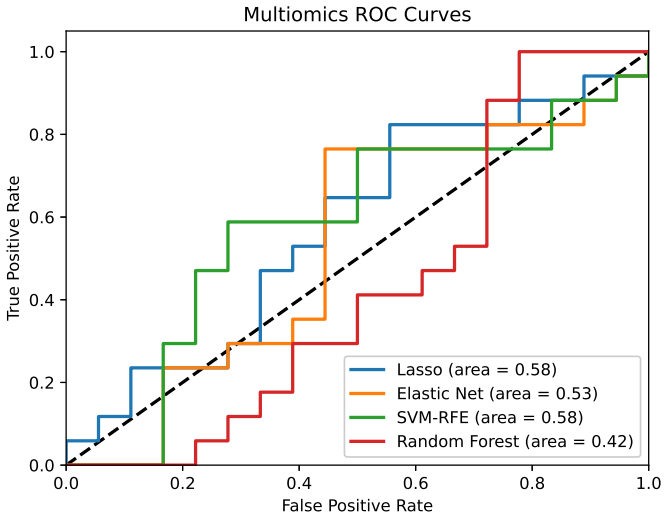


FIGURE 3.73: ROC curve for each of the classification algorithms employed for the prediction of COPD using all the omics data.

TABLE 3.31: Supervised algorithms best hyper-parameters for the prediction of COPD using multiomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0501	C: 0.0398	C: 0.0079	No. of trees: 3000
	L1 ratio: 0.1	Selected features: 50	Split criterion: Entropy
		Step: 500	Max depth: 5
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 5

3.4.2.3 Location prediction

Given the original purpose of the data set, one can also verify whether the available omics could be used to predict the location where the patients had to walk for the experiment. The idea was to use classification algorithms to predict the site where the patients samples have been collected (Hyde Park and Oxford Street) and possibly identify features related to the different levels of air pollution in the two areas. In this case, one is not interested on the diagnosis status of the patients, thus all the participants have been included in the analysis regardless of their health condition. The data set can now count on an increased size of 255 samples while the feature dimensionality is still very high (41387). Unfortunately the analysis did not show satisfying results: the ROC curves related to the use of all the omics data are shown in Fig. 3.74 while those related to other omics have not been included due to their

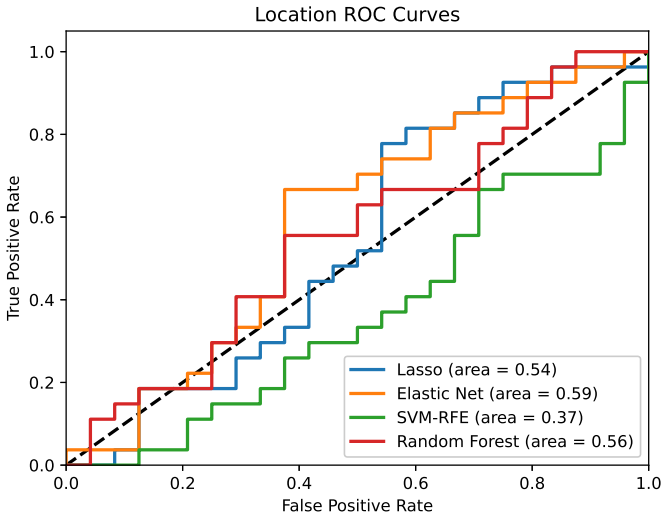


FIGURE 3.74: ROC curve for each of the classification algorithms employed for the prediction of the site (Hyde park and Oxford Street) using multiomics data.

TABLE 3.32: Supervised algorithms best hyper-parameters for the location prediction using multiomics data. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.1	C: 0.2512	C: 0.0501	No. of trees: 7500
	L1 ratio: 0.3	Selected features: 5	Split criterion: Entropy
		Step: 300	Max depth: 5
			Max features: Square root
			Min samples leaf: 2
			Min samples split: 10

poor performance. The models best hyper-parameters identified by cross-validation are shown in Table 3.32.

3.4.3 Confounders

As shown in the previous sections, omics data only were used to make a prediction for both the IHD and COPD cases. However only the first one achieved significantly positive results. Due to this, the following analyses will be focused on the healthy/IHD case. An important issue emerged from the preliminary analysis was the possible presence of confounders in the data set. A confounder can be seen as a variable influencing both dependent and independent variables which may cause the creation of spurious correlations [102]. In this case the possible confounders in the data set are the general information related to the patients, namely age, sex and BMI which might create a bias in the results. This information was not considered in the preliminary analysis, nonetheless the analysis was repeated including these features as well. The aim was to avoid the selection of features correlated with

the above mentioned confounders and possibly identifying variables only correlated with the diagnosis. Moreover, given the previous results, the analysis was focused on two specific omics: miRNA and metabolomics. The first one has been chosen for its easier biological interpretation with the disease while the second one has been selected for its high prediction capability. In other words, the following analysis will be related to the separate use of miRNA and metabolomics data to classify healthy and IHD subjects.

miRNA In this case the same analysis as in 3.4.2.1 was proposed, the only difference was the inclusion of the three confounders variables (age, sex and BMI). As shown in Fig. 3.75, the AUC values improved thanks to the confounders effect. At the same time, the most important features identified in the feature selection process also changed. Table 3.33 shows the models best hyper-parameters identified via 2-fold grouped cross-validation while Table 3.34 shows the five most important features for each classification method. As one can see, the confounders, and in particular BMI, appear in the first positions for each technique. On the other hand, there is no other feature which has been selected by all the methods simultaneously. Lasso and elastic net has exactly the same features, just in different order. This result is not surprising since elastic net mixed penalty should select the same features as lasso plus other correlated to them.

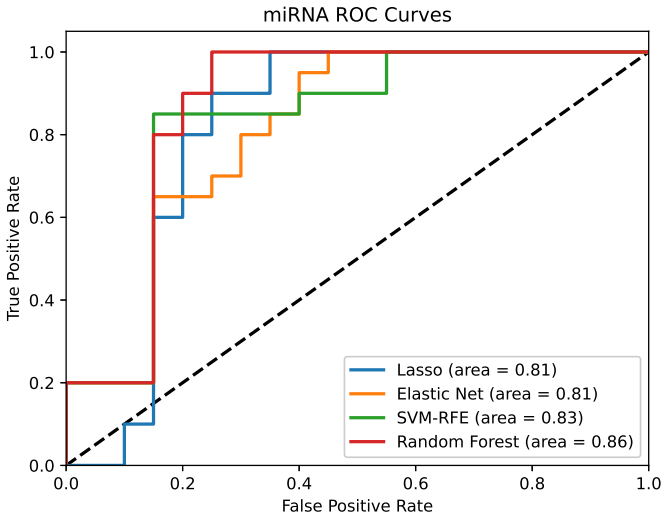


FIGURE 3.75: ROC curve for each of the classification algorithms employed for the prediction of IHD using miRNA data and the confounders.

TABLE 3.33: Supervised algorithms best hyper-parameters for the prediction of IHD using miRNA data and the confounders. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.0159	C: 0.0794	C: 0.0501	No. of trees: 100
	L1 ratio: 0.3	Selected features: 5	Split criterion: Entropy
		Step: 300	Max depth: 5
			Max features: Square root
			Min samples leaf: 1
			Min samples split: 2

TABLE 3.34: List of the five most important features selected by classification algorithms on IHD and healthy samples using miRNA data and the confounders.

Lasso	Elastic Net	SVM-RFE	Random Forest
BMI	BMI	Sex	BMI
hsa-miR-16-2-3p	Sex	BMI	hsa-miR-3676-3p
Sex	hsa-miR-4672	hsa-miR-4672	hsa-miR-5001-5p
hsa-miR-219-5p	hsa-miR-16-2-3p	hsa-miR-186-5p	hsa-miR-142-5p
hsa-miR-4672	hsa-miR-219-5p	hsa-miR-3676-3p	hsa-miR-4672

Metabolomics The same analysis has been proposed for metabolomics data including the confounders. Here the improvement in terms of performance (as shown in Fig. 3.76) are less evident since AUC values where already very high in the beginning. Table 3.35 shows the best hyper-parameters identified by a 2-fold grouped cross-validation for each of the employed classifier. The same results obtained with miRNA data in the previous paragraph can be observed in this case: confounders features (specifically BMI) are now appearing among the most important features for each model as reported in Table 3.36.

TABLE 3.35: Supervised algorithms best hyper-parameters for the prediction of IHD using metabolomics data and the confounders. The table shows the best hyper-parameters identified by a grid search cross-validation for each model.

Lasso	Elastic Net	SVM-RFE	Random Forest
C: 0.004	C: 0.002	C: 0.0501	No. of trees: 5000
	L1 ratio: 0.3	Selected features: 5	Split criterion: Entropy
		Step: 300	Max depth: 3
			Max features: Square root
			Min samples leaf: 10
			Min samples split: 2

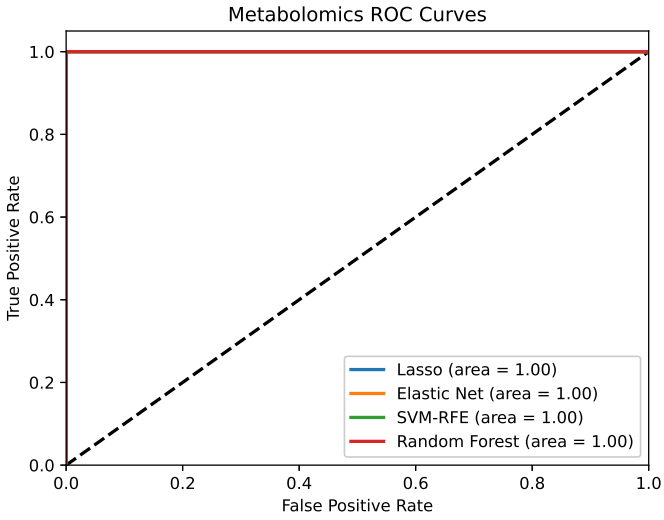


FIGURE 3.76: ROC curve for each of the classification algorithms employed for the prediction of IHD using metabolomics data and the confounders.

TABLE 3.36: List of the five most important features selected by classification algorithms on IHD and healthy samples using metabolomics data and the confounders.

Lasso	Elastic Net	SVM-RFE	Random Forest
BMI	BMI	322.2181@7.6841	1153.6049@7.02360
404.7344@8.4525	404.7344@8.4525	BMI	BMI
417.7407@8.6282	417.7407@8.6282	286.0556@3.8696	267.101@2.2428
556.2439@5.8565	486.3169@6.6818	657.3491@6.3065	404.7344@8.4525
486.3169@6.6818	556.2439@5.8565	243.0956@6.0145	354.2454@7.6842

As one can see from the results of both miRNA and metabolomics, the confounder variables (and in particular sex and BMI) appear in most of the cases. The methods performances did not change very much from the two cases, probably because the features previously identified were correlated with the confounders not present in the analysis and acted as a "substitute" for them. When it comes to the analysis including the confounders, sex appearance is due to the fact that the majority of IHD samples in this data set are male, while BMI appears since the majority of IHD samples typically have an higher BMI value than healthy ones. This is a classical example of spurious correlation since being male or having a higher BMI does not necessarily imply that a person suffers of IHD. Another issue emerging from these results is the lack of consensus among the features selected by the classification algorithms for both omics data. This aspect is particularly crucial since a potential biomarker needs to be validated and should be robust to be useful in clinical scenarios. The lack of consensus is also a signal of instability among the feature selection methods (this issue will be addressed in section 3.4.8).

3.4.4 Principal path feature selection

Here, a new and different feature selection approach has been proposed. It stems from the consideration that a disease often progresses in a continuous way, whereas a two-class labeling is a largely reductionist approach. In particular, by using the principal path (PP) algorithm [22] one can study the evolution of a patient from the healthy to the sickness status. At the same time one wants to perform feature selection by choosing variables which are both predictive and progressive with respect to the patient's status. The PP algorithm allows to create a path connecting healthy patients with those affected by IHD. By doing so, it captures the data distribution constrained by a starting and an ending point by creating an evolutionary and smooth path between the two points. Moreover, this method can also be used to perform feature selection by considering the most correlated variables with the way points progression along the path. In practice, each way point represents an intermediary patient/step across the disease evolution with its own coordinates and features. Therefore, the corresponding features values should increase or decrease following the progression of the path. Thus, it is possible to calculate the correlation between the features values and the path progression (i.e. the waypoints progression along the path): the most correlated features should be those better describing the disease evolution [103].

Fig. 3.77 and 3.78 show a t-SNE two dimensional representation of the path connecting healthy and IHD samples using miRNA and metabolomics data respectively. In both cases the confounder features have also been included to take into account of their effect. Each path is composed of 50 way points and for both cases the starting and ending points correspond to the centroids of the two clusters (healthy and IHD). The idea is that the centroid identified by k-means can be seen as a summary of the two patients group since it is not possible to define a "healthier" or a "sicker" patients as extreme cases. The purple path is a trivial path (shown for comparison purpose only) which merely connects the starting and the ending point without capturing the data distribution. Both t-SNE representations might differ from those previously presented: this is due to the presence of other points (the PP and trivial path) in the space which may alter the original samples projection. In any case, it is visible how IHD and healthy samples are much better separated with metabolomics data than with miRNA confirming metabolomics higher prediction power.

It should be considered though, that a single principal path might not be enough to capture this kind of evolution in a reliable way. In order to improve the robustness of this analysis, different paths were created by perturbing the starting and the ending points and, for each feature, the average of the correlation values scored in every path was measured. Fig. 3.79 shows an example of perturbed paths on metabolomics data, where four of them have been created (a higher number of paths can and has been used but only four of them were shown in order to make the plot more understandable). Once the most correlated features have been identified, one also wants to make sure that those features are not only progressive with the disease but also

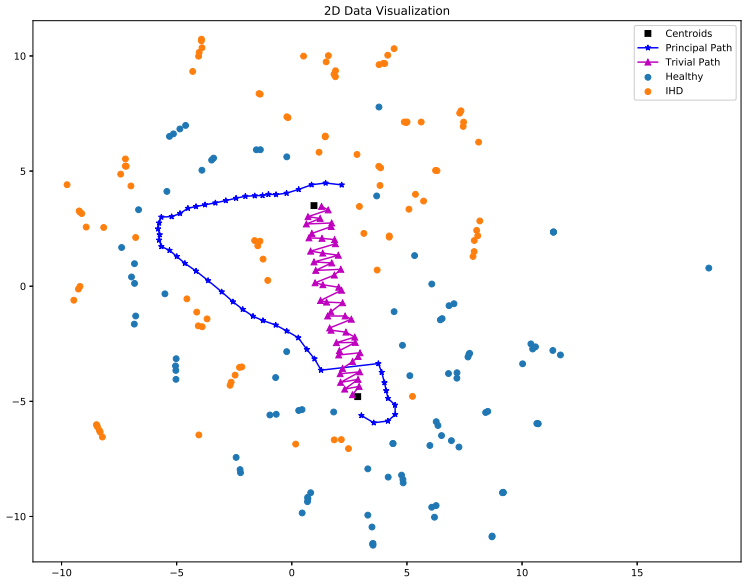


FIGURE 3.77: t-SNE two dimensional representation of the principal path on miRNA data. The purple line indicates the trivial path merely connecting the starting and the ending point without capturing the data distribution.

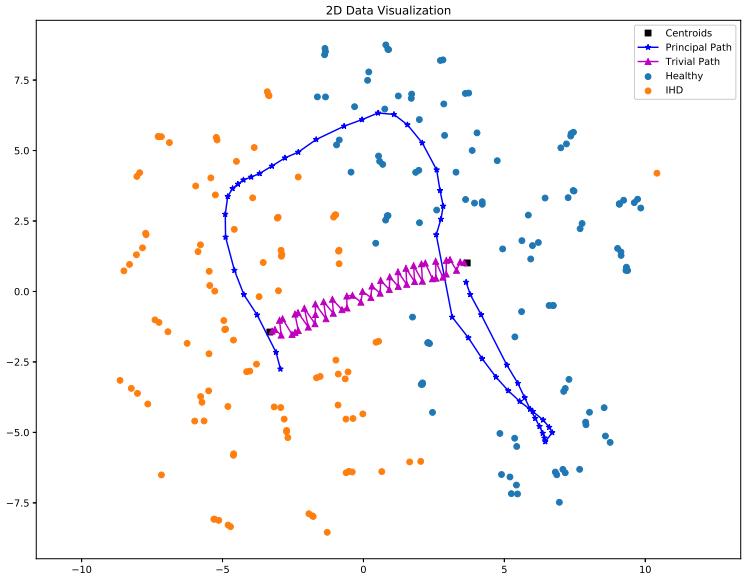


FIGURE 3.78: t-SNE two dimensional representation of the principal path on metabolomics data. The purple line indicates the trivial path merely connecting the starting and the ending point without capturing the data distribution.

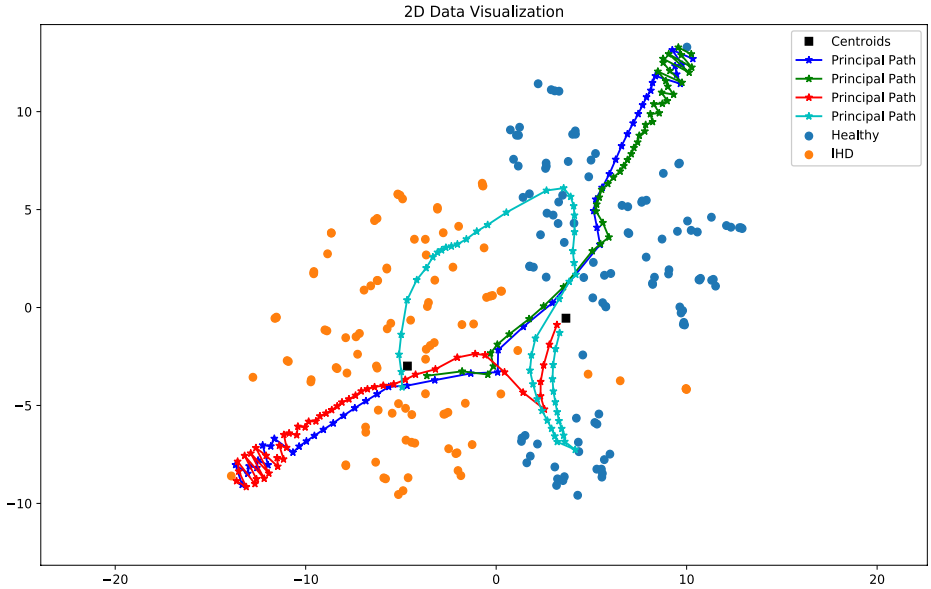


FIGURE 3.79: t-SNE two dimensional representation of four perturbed principal paths on metabolomics data created by considering the two groups centroids and their closest samples as starting and ending points.

powerful in terms of prediction accuracy. To do so, one can make use of a classification tree to evaluate the prediction performance of the most correlated features (one at a time). Classification trees also allow to identify the threshold of the feature value to make the prediction more interpretable and helpful under a practical point of view. Therefore, the selected features are those variables that best classify healthy and sickness status and that are progressive with the disease and thus putatively can smoothly represent a transition between healthy and IHD status.

Tables 3.37 and 3.38 show the most predictive features selected among those with an average correlation value with the path progression higher than 0.80. The tables also include the corresponding correlation, classification tree accuracy, AUC and threshold values. For this analysis the two groups centroids and the 5 closest points to those centroids were considered, for a total of 6 starting and ending points. Then all the possible combinations of starting and ending points were considered, creating a total of 36 perturbed paths. From these results one can also see that, despite having included the confounders in the analysis, the PP did not select any of them for the metabolomics case, where the prediction power is higher, while sex was included only in the miRNA case.

3.4.5 Identifying biomarkers

A further step of this analysis was trying to understand the biological meaning of the results obtained until now. The first goal was to identify the metabolites and miRNAs selected by the PP and by other classification algorithms. As one can see,

TABLE 3.37: List of the most correlated features with the principal path on IHD and healthy samples with miRNA data.

miRNAs	Accuracy	AUC	Correlation	IHD Threshold
hsa-miR-3162-3p	0.90	0.89	-0.86	≤ 3.48
hsa-miR-4433-5p	0.81	0.80	-0.84	≤ 3.72
Sex	0.73	0.74	-0.84	Male
hsa-miR-3135b	0.73	0.73	-0.80	≤ 4.73
hsa-miR-374b-5p	0.70	0.70	-0.88	> 9.78
hsa-miR-181a-5p	0.70	0.69	-0.86	≤ 9.50
hsa-miR-374c-5p	0.69	0.70	0.89	> 4.95
hsa-miR-374a-5p	0.69	0.69	0.85	> 10.16

TABLE 3.38: List of the five most correlated features with the principal path on IHD and healthy samples with metabolomics data.

Metabolites	Accuracy	AUC	Correlation	IHD Threshold
785.0254@8.695904	1.00	1.00	-0.80	≤ 161610.00
419.7567@9.274288	0.93	0.93	-0.89	≤ 204501.50
470.721@8.721389	0.93	0.93	-0.81	≤ 376664.00
405.7404@8.695825	0.91	0.91	-0.82	≤ 806722.50
364.2574@7.3848987	0.90	0.90	-0.80	≤ 187550.00

while miRNA names correspond to their exact ID, the same cannot be said about the metabolites since an untargeted metabolomics was performed when the Oxford Street data was collected. The metabolites names are composed of two parts separated by an "@". The first one is the mass-to-charge ratio (m/z) measured my the mass spectrometry while the second one is the retention time. Although such information is available, it is not enough to identify the metabolites ID, consequently it was not possible to verify the biological meaning of the metabolites.

On the other hand, it was possible to identify genes targeted by the selected miRNA to verify whether they are actually involved with IHD or more generally with cardio-vascular diseases (CVD). From some research conducted consulting online databases such as miRBase [104] it was possible to obtain more information about the miRNAs identified by the PP. Both hsa-miR-181a-5p and hsa-miR-374a-5p target ATM. Hsa-miR-181a-5p also targets BCL2 while hsa-miR-374a-5p targets WNT5A.

A first association between ataxia telangiectasia mutated kinase (ATM) and CVD was found in [105], where, according to the Authors, an ATM deficiency affects heart function, infarct thickness, fibrosis, apoptosis and expressions of their relative proteins. Following this work, many researches have been conducted to study the impact this protein has on CVD such as the role it plays in cardiac remodelling [106]. When it comes to BCL2, different studies confirm that it is mainly associated with hypertension: its level is higher in patients affected by the disease [107].

Hsa-miR-3162-3p is the most predictive miRNA identified using the previously exposed method. It is related with CTNNB1 gene which activates WNT/β-catenin

[108]. β-catenin has many implications in health and disease by controlling different processes such as intercellular adhesion, signal transmission, tumour formation, apoptosis and necrosis [109]. The activation of this protein also causes progressive cardiac disfunction and heart failure while its inhibition might be used for therapeutic intervention of hypertensive heart disease [110]. Both miR-181a and miR-3162-3p are involved in asthma as well. In particular, the second one was observed to be higher in asthma patients [109]. WNT5A is targeted by hsa-miR-374a-5p and is also part of the β-catenin pathway.

These findings agree upon the way points correlation sign: a decrease of miR-3162-3p causes an upregulation of CTNNB1 and then an increase of β-catenin which is observed to be higher in patients affected by cardio-vascular disease. The same logic can be applied to hsa-miR-374a-5p and ATM whose path correlation sign is positive: hsa-miR-374a-5p is higher in IHD patients, which is a possible signal of an ATM downregulation causing the disease. On the other hand, the correlation sign does not match the one related to miR-181a-5p, but this is possibly due to miR-181a-5p targeting a multitude of different genes whose global effect might be different from the single targeting of ATM. In fact, as described in [111], it has been observed that a decrease expression of miRNA-181a-5p and miRNA-181a-3p is present in patients affected by artery disease such as atherosclerosis.

In order to further verify the PP results one can decide to compare, in both terms of correlation and prediction accuracy, the most important features selected by elastic net and the PP approach without using the known confounding factors (age, sex and BMI). Until now, elastic net has been used by splitting the data set into training and test set plus cross-validation for hyper-parameters tuning. The PP, on the other hand, does not need this kind of approach and has been applied on the whole data set. Therefore, the data set was split into training and test set for both techniques. More specifically the training set was composed of 131 samples while the remaining 60 samples were used as test set. The test set was created by including only patients for whom all the 6 collected blood samples were available (as described in 3.4.1): by doing so one is able to verify whether the results are consistent even changing the time and location of the samples collection and thus not dependent on the environmental exposure. The training set was then used to launch the PP algorithm and to perform cross-validation for elastic net training and model selection. Afterwards, a classification tree (trained and tested on the same training and test set respectively) has been implemented for each of the first five more important features identified by elastic net and PP. By doing so one is able to use an unbiased method to verify the prediction accuracy of each miRNA taken singularly and verify which were the most predictive. Moreover, since the training/test splitting may create feature selection instability (this issue will be better discussed in 3.4.8), the splitting process was repeated for five times and for each of these the PP and elastic net feature selection processes were repeated. By doing so one was also able to identify the most stable features identified by the two methods. Table 3.39, shows the five most frequent

features identified by each method, their related frequency and average accuracy and standard deviation obtained from the classification tree. As one can see, the features identified by the PP are more stable, accurate and have lower standard deviation than those identified by elastic net. However, once the features have been identified, one can also verify the correlation of these features with the confounders excluded from the analysis: Fig. 3.80 and 3.81 show the correlation matrix of the features identified by the two approaches and the confounding factors. As one can see the features identified by the PP are more correlated with BMI and more correlated among themselves.

TABLE 3.39: List of the five most frequent features identified via the principal path and elastic net on IHD and healthy samples with miRNA data.

Methods	Avg. Acc.	Avg. St. Dev.	Features	Frequency
Elastic Net	0.41	0.03	miR-1227-5p	0.80
	0.74	0.03	miR-3676-3p	0.60
	0.57	0.07	miR-16-2-3p	0.60
	0.68	0.13	miR-186-5p	0.40
	0.52	0.12	miR-4788	0.40
Principal Path	0.80	0.08	miR-3162-3p	1.00
	0.68	0.07	miR-4433-5p	1.00
	0.64	0.02	miR-374c-5p	1.00
	0.61	0.04	miR-374a-5p	0.40
	0.79	0.08	miR-3135b	0.40

EN	Age	BMI	miR-1227-5p	miR-3676-3p	miR-16-2-3p	miR-186-5p	miR-4788
Age	1.00	0.04	-0.11	0.11	-0.24	-0.19	-0.05
BMI	0.04	1.00	0.06	-0.48	0.11	-0.40	0.47
miR-1227-5p	-0.11	0.06	1.00	0.16	-0.26	-0.09	-0.10
miR-3676-3p	0.11	-0.48	0.16	1.00	-0.30	0.04	-0.31
miR-16-2-3p	-0.24	0.11	-0.26	-0.30	1.00	0.13	0.03
miR-186-5p	-0.19	-0.40	-0.09	0.04	0.13	1.00	-0.16
miR-4788	-0.05	0.47	-0.10	-0.31	0.03	-0.16	1.00

FIGURE 3.80: Pearson correlation matrix between the five most frequent features (miRNA) identified by elastic net and the confounding factors (age and BMI).

PP	Age	BMI	miR-3162-3p	miR-4433-5p	miR-374c-5p	miR-374a-5p	miR-3135b
Age	1.00	0.04	0.07	0.13	0.01	-0.07	-0.05
BMI	0.04	1.00	-0.42	-0.33	0.26	0.38	-0.22
miR-3162-3p	0.07	-0.42	1.00	0.82	-0.52	-0.51	0.27
miR-4433-5p	0.13	-0.33	0.82	1.00	-0.45	-0.46	0.25
miR-374c-5p	0.01	0.26	-0.52	-0.45	1.00	0.69	-0.39
miR-374a-5p	-0.07	0.38	-0.51	-0.46	0.69	1.00	-0.43
miR-3135b	-0.05	-0.22	0.27	0.25	-0.39	-0.43	1.00

FIGURE 3.81: Pearson correlation matrix between the five most frequent features (miRNA) identified by the principal path and the confounding factors (age and BMI).

In conclusion, although the association between some of the miRNA identified by the PP and heart diseases has been observed in literature, no strong association with IHD specifically has been identified. At the same time, two limits of the PP feature selection method have been investigated. The first one is also common to other feature selection techniques and merely depends on the input data: being BMI an important confounder, most of the identified features are correlated with it and their relevance is less evident when considering the causality with IHD. The second issue comes from the feature selection method itself: identifying the most correlated features might bring useful information to understand the disease, but the features selected in this way are not only correlated with the disease but are also correlated with each other. In other words, the features are not orthogonal and the amount of information carried by each feature is inferior when compared to other feature selection techniques (e.g. elastic net). Due to this, the features selected by the PP may perform better singularly but they perform worse when combined together. At the same time, being the features correlated between themselves and with the confounding factors as well, a model including the confounders and those features will necessarily perform worse than a model including the confounders and other orthogonal features since the amount of information carried is larger. These issues will be discussed in the following sections.

3.4.6 Adjusting for confounders

In order to investigate the first issue observed with the PP feature selection approach, one can correct the data to take into account of the confounders effect. By using an approach similar to the one reported in [112] one wants to exploit the residuals of a

linear regression to "clear" the data from the effects of the confounders. In this work, Authors considered a linear regression between the outcome and the confounders, they then subtract the so predicted outcome to the original value in order to obtain a new residual-outcome value which has been purified from the confounders effect. In this thesis, a similar approach was implemented. Given an outcome variable Y (healthy or IHD), a set of confounders variables C (BMI, sex and age) and the remaining features set X (miRNA data in this case), one wants to study the effect of both variables on Y:

$$Y = \beta_0 + \beta_1 C + \beta_2 X. \tag{3.1}$$

By using an approach similar to [112] it is possible to clear the features set X from the effects of the confounders. To do so, one can use a linear regression model to predict X using C only:

$$\hat{X} = \alpha_0 + \alpha_1 C. \tag{3.2}$$

Once this has been done, the residual Z can be calculated as the value of X which is not explained by C (the part of X which is independent of C):

$$Z = X - \hat{X}. \tag{3.3}$$

Finally, the new value of Z is used as input to the original model to predict the outcome Y considering only the part of X not affected by the confounders:

$$Y = \gamma_0 + \gamma_1 Z. \tag{3.4}$$

Therefore, this approach has been employed to get new values for miRNA which did not take into account of the confounding factors effects.

The new residual matrix has then be used as input to repeat the analysis for both elastic net and PP feature selection. Moreover, since the PP would keep selecting features correlated between themselves, a new approach to mitigate this issue was applied. This approach has been called recursive principal path. Starting from the residual matrix as in (3.3), the first PP is computed on those data. Afterwards, the best feature is identified as before (the most correlated one) and then the residual matrix is re-computed using that feature as new variable to remove its effect from the other features. This procedure is repeated until a desired number of features has been identified:

$$\begin{aligned} RES_1 &= Z = X - \alpha_0 - \alpha_1 C \\ RES_2 &= RES_1 - a_0 - a_1 Best\_Feature\_1 \\ &\vdots \\ RES_k &= RES_{k-1} - b_0 - b_1 Best\_Feature\_{k-1}. \end{aligned} \tag{3.5}$$

As before, this approach was repeated until it was possible to select five features. Fig. 3.82 and 3.83 show the correlation matrices for the five most important features selected by elastic net and PP. As one can see from these new figures, the correlations among features and with the confounders are lower than before. Although some significant correlation is still present, one is able to mitigate the effects of redundant features using the PP feature selection. Nevertheless, these results can also be compared with the features selected by lasso. Differently from lasso, elastic net makes use of a double penalty (lasso and Tikhonov regularisations), thus it may happen that features correlated between themselves are still included. On the other hand, when features are correlated, lasso tends to pick one of them and automatically exclude the others [27] reducing the redundancy in the final selected features. Fig. 3.84 shows the correlation matrix of lasso features and confirms this result: features selected via lasso are generally less correlated than those identified by elastic net or the PP.

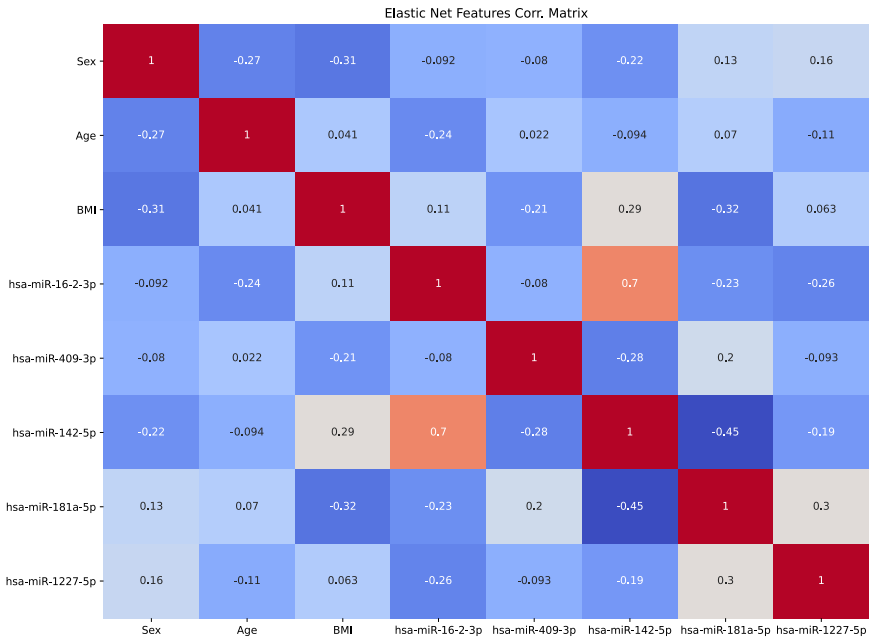


FIGURE 3.82: Correlation matrix between the confounding factors (age, sex and BMI) and the five most frequent features (miRNA) identified by elastic net on residual data.

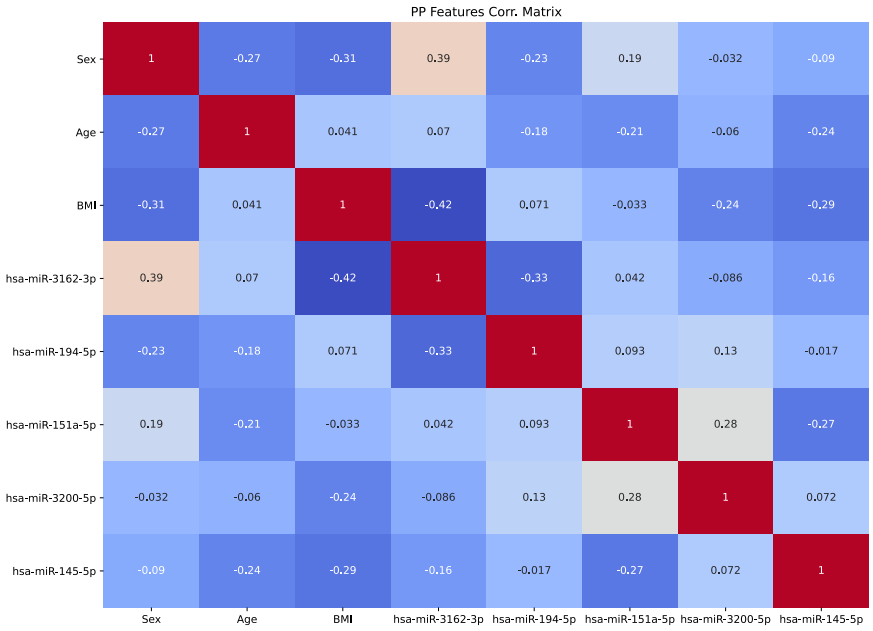


FIGURE 3.83: Correlation matrix between the confounding factors (age, sex and BMI) and the five most frequent features (miRNA) identified by the recursive PP on residual data.

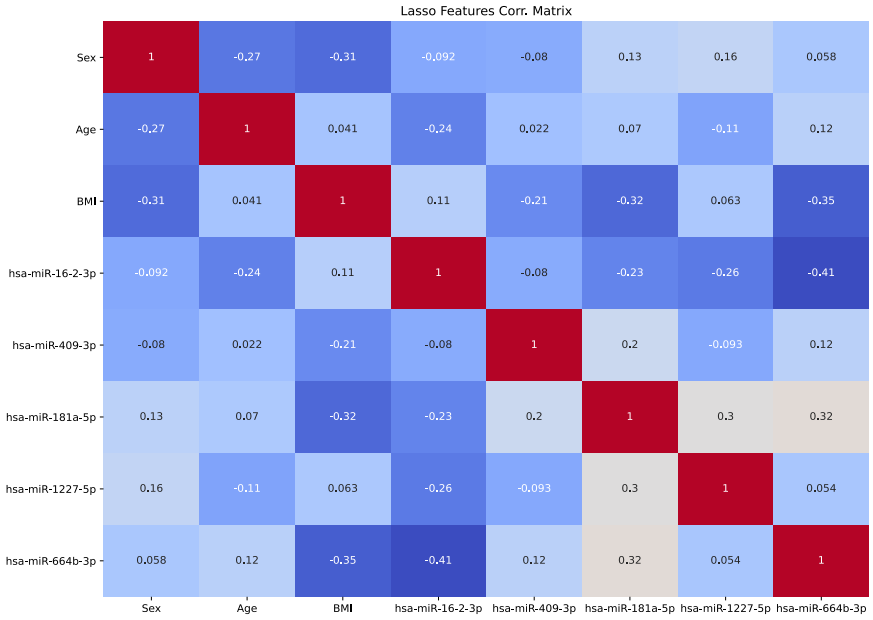


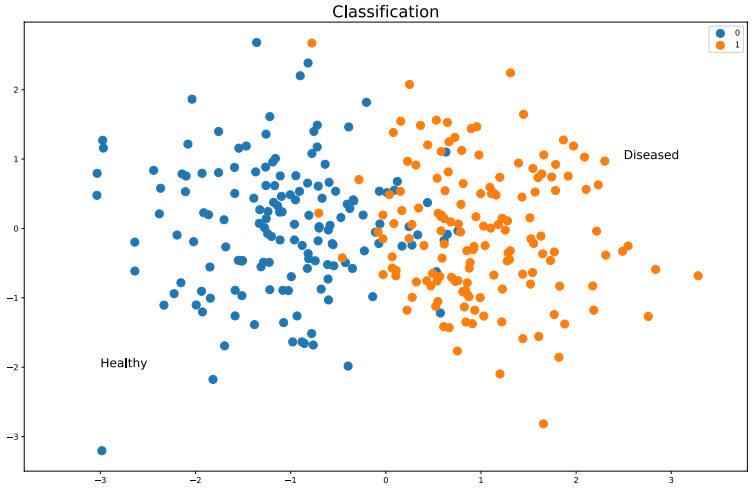
FIGURE 3.84: Correlation matrix between the confounding factors (age, sex and BMI) and the five most frequent features (miRNA) identified by lasso on residual data.

3.4.7 Continuous diagnosis

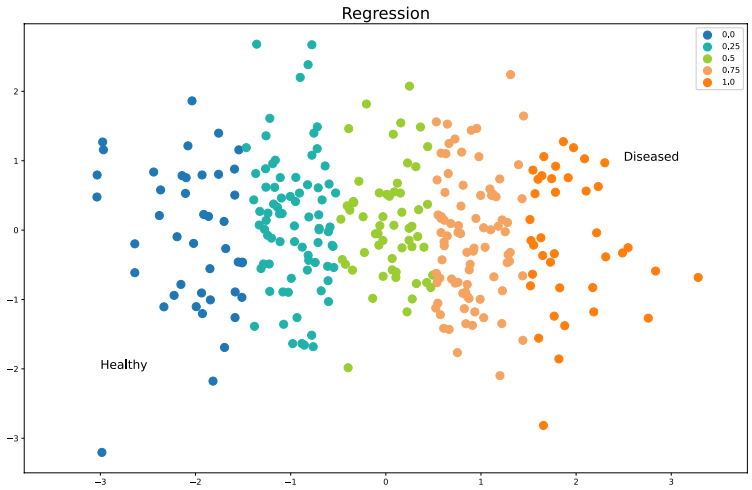
Despite the new approach introduced in the previous section, lasso and elastic net are still preferable feature selection methods for automatically identify predictive and orthogonal features. Therefore, one can merge the two methods: on one side one can exploit lasso/elastic net feature selection capabilities, on the other one it is possible to use the PP to map the transition from healthy to diseased. Until now, the healthy/disease prediction has been considered as a binary problem and thus a classification task. However, the transition from healthy status to diseased is not immediate and one can assume that it follows progressive steps. Given this, one can also assume that a binary discrimination might not be the right approach to investigate the causes of a disease. If one wants to predict not only the diagnosis but also a patient's prognosis, one wants to identify features that are not only predictive but also progressive with the disease. In order to do this, it is possible to use the PP to assign new progressive labels to each data sample. Fig. 3.85 shows an intuitive example of what one would like to observe in this scenario: instead of having a clear separation between the two classes (Fig. 3.85a), a gradual and progressive transition is preferable (Fig. 3.85b).

Since the PP way points describe the transition between a healthy patient to one affected by IHD, each way point can be seen as an intermediate stage of the illness status. By using a classification algorithm such as the k-nearest neighbours, one can assign to each sample a new progressive label related to the illness stage. Therefore, one can convert a classification problem into a regression one and then use regression algorithms to perform the analysis as well as feature selection. The features identified in this way are then, putatively, progressive with the illness status and more robust and stable than those identified with a classification algorithm. These features could be used not only to describe the binary healthy/sick status but also to identify intermediate states of illness. The idea of transforming the diagnosis into a continuous process comes from the ergodic theory in the statistical physics field. The intrinsic time of a disease could last even for years but a longitudinal study (i.e. a study in which the patients are observed for many years) is difficult to implement. Due to this, in this kind of experiments one can "swap" time and space and instead of observing some patients for a time span, one can observe many patients at a specific time: as an alternative to study the time flow you can use a path (or many paths) to capture the progress of the disease.

In order to validate this new approach, one can compare its results with those of a lasso binary classification. Therefore, lasso has been employed for both binary classification and regression. Moreover, to avoid compatibility issues between the two methods, model selection was not performed and the level of λ regularisation parameter was fixed a priori for the two models. The number of PP way points was also reduced from 50 to 10 and then the twelve classes (ten way points plus starting and ending points) were converted in a range between 0 and 1 to bring the problem on the same scale of the binary classification. The performance of the analysis was



(A) Classification with binary labels.



(B) Regression with continuous labels.

FIGURE 3.85: Example representation of the transition from a binary to a continuous task.

evaluated using R^2 and accuracy (by converting the twelve classes to binary again by setting the standard 0.50 probability threshold). The five most important features were used to build a logistic regression model to evaluate their performance in a different way. The original data values were also used instead of the residual ones since lasso feature selection allows to automatically select orthogonal features. In other words, even if lasso selects the confounders as most important features, the other features would not be correlated with them. Table 3.40 shows the results of the analysis. The value of λ was set at 0.01 for both cases and the whole the data set was used (i.e. without train/test splitting) to evaluate R^2 and accuracy values. On the other hand, the data set was split into training and test set to build the logistic regression models. The table reports the accuracy values and their respective confidence interval using all the features selected by lasso, only the most important five features and the five most important features excluding the confounders. As

one can see, lasso generally performs better on the binary case in terms of both accuracy and R^2. Logistic regression scores similar results when using all the features. When considering the first five features, binary lasso performs significantly better: this is because the first five features include sex and BMI, while they do not appear in the PP labels case. In fact, when not considering the confounders, binary lasso still performs better but the spread is lower.

Although the regression lasso performance are lower compared to the binary lasso, one important result emerged from the results. In all the analysis performed until now, the confounders (BMI in particular), played an important role among the selected features. However, as shown in Table 3.40, sex and BMI only appear in the last positions. Therefore, this result was investigated and it has been observed that this effect was due to the perturbation of the PP. In fact, when considering one PP only (the one connecting the healthy and IHD centroids) and use it to relabel the samples the results changed. In Table 3.41 the ten most important features for the three cases are shown. As one can see, when considering one PP only, BMI and sex acquire more importance. The performance also improved and now the PP-labelling has higher R^2 than the two previous cases. Accuracy also improved but it is still lower than the binary lasso case. One can then conclude that using one single path better maps the transition from healthy to IHD samples but it still does not guarantee to get better results for the binary case.

Since it is not possible to establish a priori which method is better, one can verify the biological relevance of the most important selected miRNA. From a database analysis emerged that of the ten features for both methods, four did not have a validated target, including the one in common (miR-4672). The twelve remaining features had seven common targets, namely APP, DICER1, MET, MYC, PTEN, RAB5A and ZEB1. By further verifying the association between the targets lists and cardiovascular diseases, it was possible to identify 85 strong associations for the PP case. Of these, 10 are in common with the lasso case and four are specifically associated with IHD (DUSP1, VEGFA, RAB5A, ZEB1) [113]. On the other hand, binary lasso identify 169 associations of which 5 are associated with IHD (CCND1, KLF4, TNF, RAB5A and ZEB1) [113], [114]. By considering these results only, the features identified by binary lasso seem to have a slight better biological relevance than those identified with the relabeling approach. However these features only consider the association with the disease and none really takes into account of the progression with the disease itself. Unfortunately, there is no way to practically validate this new approach to understand whether these features are actually associated with the evolution of IHD, with one possible solution being the creation of a specific longitudinal data set observing how the disease evolved in patients.

TABLE 3.40: List of the most important features identified via lasso with binary and PP labels using miRNA data.

	Lasso	Lasso PP Labelling
Features	BMI	miR-199b-5p
	miR-16-2-3p	miR-374a-5p
	Sex	miR-628-3p
	miR-409-3p	miR-4672
	miR-130a-3p	miR-335-5p
	miR-340-5p	miR-126-5p
	miR-1255a	miR-101-3p
	miR-1227-5p	miR-3127-5p
	miR-4672	miR-1914-3p
	miR-142-5p	miR-4653-3p
	miR-186-5p	miR-141-3p
	miR-628-5p	miR-301b
	miR-582-5p	Sex
	miR-339-3p	BMI
	miR-497-5p	miR-20a-3p
	Age	
	miR-4306	
	miR-3676-3p	
	miR-22-3p	
	miR-181c-5p	
R^2	0.64	0.59
Accuracy	0.92	0.75
Logistic Acc. (CI)	0.80 (0.73, 0.86)	0.79 (0.72, 0.86)
Logistic Acc. (CI) - Top 5	0.83 (0.73, 0.93)	0.68 (0.61, 0.75)
Logistic Acc. (CI) - Top 5 - No Conf.	0.74 (0.67, 0.81)	0.68 (0.61, 0.75)

TABLE 3.41: List of the ten most important features identified via lasso with binary, PP and one-PP labels using miRNA data.

	Lasso	Lasso PP Labelling	Lasso One-PP Labelling
Features	BMI	miR-199b-5p	miR-3651
	miR-16-2-3p	miR-374a-5p	BMI
	Sex	miR-628-3p	miR-374a-5p
	miR-409-3p	miR-4672	miR-4672
	miR-130a-3p	miR-335-5p	miR-141-3p
	miR-340-5p	miR-126-5p	miR-219-5p
	miR-1255a	miR-101-3p	miR-101-3p
	miR-1227-5p	miR-3127-5p	Sex
	miR-4672	miR-1914-3p	miR-432-5p
	miR-142-5p	miR-4653-3p	miR-144-5p
R^2	0.64	0.59	0.71
Accuracy	0.92	0.75	0.82

3.4.8 Feature selection stability

From all the analyses until now conducted, one important shared issue emerged more or less independently of the machine learning methods used. This issue is the feature selection (in)stability. Independently of the algorithm employed (classification, regression or PP methods), it has been very hard to identify robust features, i.e. features that are constantly selected by the various feature selection methods. The main objective of this analysis was to identify possible bio-markers of IHD and although this goal was partially achieved, the lack of a general consensus among the methods employed is still an issue that should be solved in order to understand if a feature might be a good bio-marker for the disease. Feature selection stability is often overlooked in practice: most machine learning applications have a very large amount of data available making the identification of stable features set less relevant. However, when dealing with clinical and biological scenarios, one may be working in a small sample regime with a limited number of samples. Moreover, biological and omics data sets are also characterised by high dimensionality which further enhances the feature instability issue. When the number of features largely exceeds the number of samples, even a small change in the training set can lead to different results in terms of selected features. Feature selection stability can thus be defined as the sensitivity of selected features to small perturbations in the training set [56]. In the previous analysis the instability of feature selection has never been evaluated in an objective way and it was merely observed as a persistent issue in all the analysis. In section 2.6 different ways to quantify the feature selection stability were discussed while here a new method to improve it will be proposed. Part of the results herein exposed have been published in [115].

Given the advantages of the indices mentioned in section 2.6, they have been used to evaluate the stability of the algorithms employed in the Oxford Street analysis. In particular, (2.55) was used for lasso and elastic net, while (2.56) was used for random forest and SVM with RFE. These results were also compared with those obtained using (2.57) to verify how the features redundancy affects the stability on this data set. In the case of random forest and SVM a threshold of 10 was set to identify a subset of features.

In order to evaluate the stability, the training and test splitting process has been repeated for five times and the feature selection was performed for each of the splits. The number of repetitions was limited to five since it was enough to evaluate the stability for this case, but a higher number could have also been used. The leave-one-out approach has not been considered because it was still possible to perform data splitting. A 5-times repeated 2-fold grouped cross-validation was also performed for model selection. By repeating the train/test split, it was possible to identify the features that can robustly predict the health/IHD status using both miRNA and metabolomics data. For each model, the first ten most stable features were selected as those with the highest frequency and average importance weight.

The trade-off values between prediction accuracy and feature stability for each

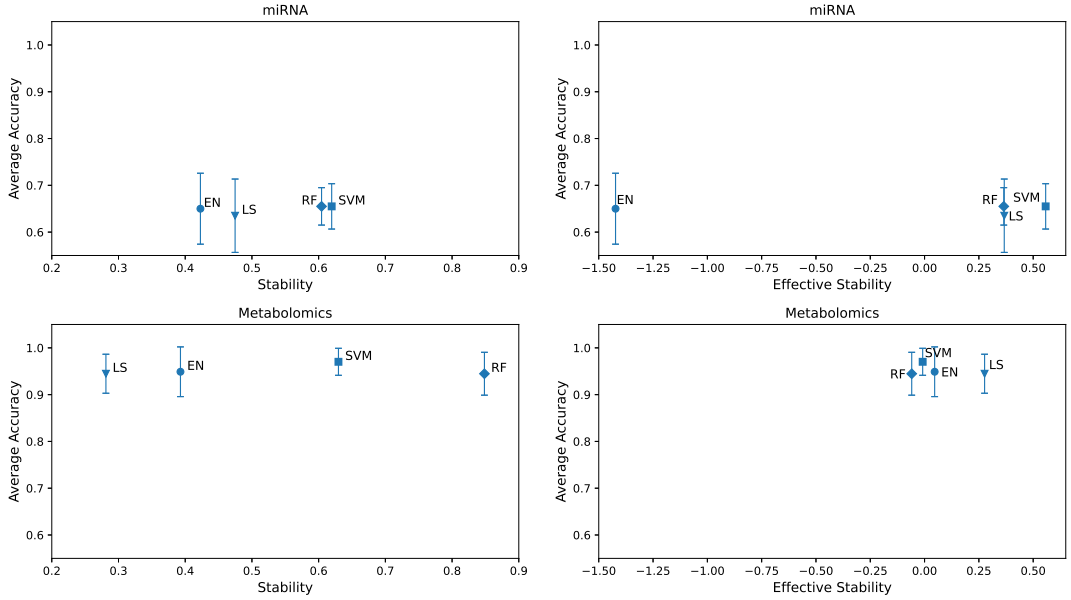


FIGURE 3.86: Accuracy/Stability trade-off between different feature selection methods using miRNA and metabolomics data. The error bars represent one standard deviation for each model's accuracy.

of the four methods is displayed in Fig. 3.86. For each model, the corresponding accuracy error bar (calculated with one standard deviation) is specified. As shown in the figure, stability values vary a lot and only random forest on metabolomics data reaches a high value of stability when considering the first index. However, the dramatic decrease in SVM and RF effective stability might be due to the need to set a threshold for the effective stability index, while the indices reported in the left plots of Fig. 3.86 are calculated considering the features ranking. Spearman's rank correlation was used to consider feature redundancy to calculate (2.57).

The results show low stability for some of the proposed methods, especially when considering the effective stability index values. This is probably because, in the presence of many correlated and highly predictive features, the proposed algorithms fail to consistently select the same features. This issue was already observed in previous sections, where the only common feature identified were the confounding factors (BMI and sex in particular). In order to mitigate feature selection instability a new method which allows to reduce feature redundancy has been introduced. This new approach will be discussed in the following paragraph.

3.4.8.1 Improving the stability of feature selection

In order to improve feature selection stability, an unsupervised feature selection filter to be applied before the supervised feature selection stage has been proposed. Fig. 3.87 shows a schematic summary of the applied procedure. To perform this first stage, the k-medoids algorithm was used. After transposing the input data matrix the feature space is clustersed instead of the sample space, in the same spirit of co-clustering [116] to select representative features. The representative features are the

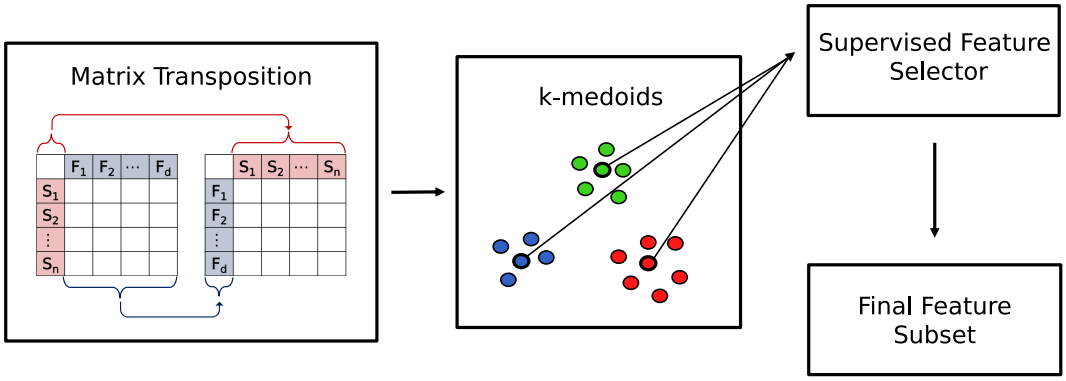


FIGURE 3.87: Description of the unsupervised feature filtering. After having transposed the data matrix, a k-medoids algorithm is run on the features. The feature-medoids only are then used for a subsequent supervised feature selection.

cluster centers which will be used in the next step to perform the supervised feature selection. By doing so, one is able to filter features a priori, independently of the labels in an unbiased way. In other words, the filter should be based only on $p(x)$ instead of being based on $p(y|x)$ distribution. The idea is that, if the features belong to the same cluster, they are similar to each other and only the prototypical and most central feature can be used to perform a classification without loss of information.

A similar process was proposed by [117], where a combination of clustering and k-nearest neighbour has been employed, while [118] exploited k-means to cluster features and a correlation measure to reduce redundancy. In this case, k-medoids has been used since it selects prototypical features [41], whereas k-means [37] would not perform a proper feature selection stage. Moreover, k-medoids is often more resilient to outliers than k-means. Other clustering algorithms such as DBSCAN [119] or mean shift [120] could have been used, however they share the same limitation of k-means of not delivering real samples.

In this case, the elbow criterion [72] has been used to identify the optimal number of clusters: 81 clusters for the miRNA case and 5757 for metabolomics have been identified. By doing so, it was possible to improve the stability of the proposed methods as shown in Fig. 3.88. More specifically, both stability and accuracy improved when using miRNA data, as well as with metabolomics. However, in the second case the improvements are less evident. This is probably due to the large amount of metabolites selected with the elbow criterion. Moreover, the starting level of accuracy in the metabolomics case was already high and the improvements are less evident in this case.

As a further step, it was also possible to evaluate the stability of the selected features between the four models in order to verify how much each algorithm is consistent with the others. The same indices has thus been applied on the ten most important features selected. The results are shown in Table 3.42. As one can see, by using the unsupervised approach, stability between the models highly increased

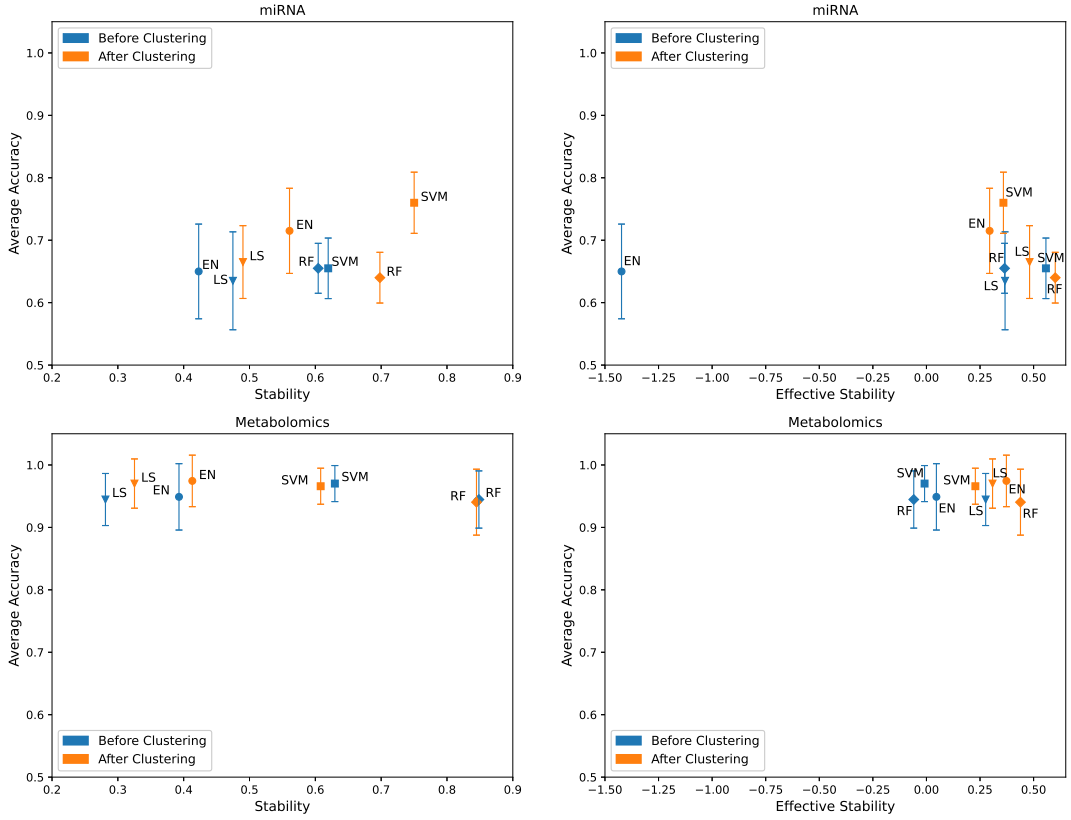


FIGURE 3.88: Accuracy/Stability trade-off between different feature selection methods using miRNA and metabolomics data before and after unsupervised feature selection. The error bars represent one standard deviation for each model's accuracy.

with miRNA data, while it worsened with metabolomics.

Therefore, this issue has been investigated by trying to change the number of clusters used for k-medoids. The analysis was repeated by using 100, 300, 500, 1000, 2000, 3000, 5757 (the value determined by the elbow method), and 10046 (the extreme case considering all the features) clusters. Fig. 3.89 shows the different levels of stability scored according to the different values of k. As shown in the plot, the elbow criterion did not find the optimal number of clusters in terms of stability and it is clear that the effective stability index between the models fell into a local minimum, explaining the massive decrease to -0.42. At the same time, neither of the stability indices monotonically increased with lower values of k, confirming

TABLE 3.42: Stability among different classification methods. The table shows the values of feature selection stability on miRNA and metabolomics data before and after unsupervised feature filtering.

	Before Clustering	**After Clustering**
miRNA $\hat{\Phi}(Z)$	0.45	0.60
miRNA $\hat{\Phi}_C(Z)$	0.11	0.52
Metabolomics $\hat{\Phi}(Z)$	0.47	0.45
Metabolomics $\hat{\Phi}_C(Z)$	-0.26	-0.42

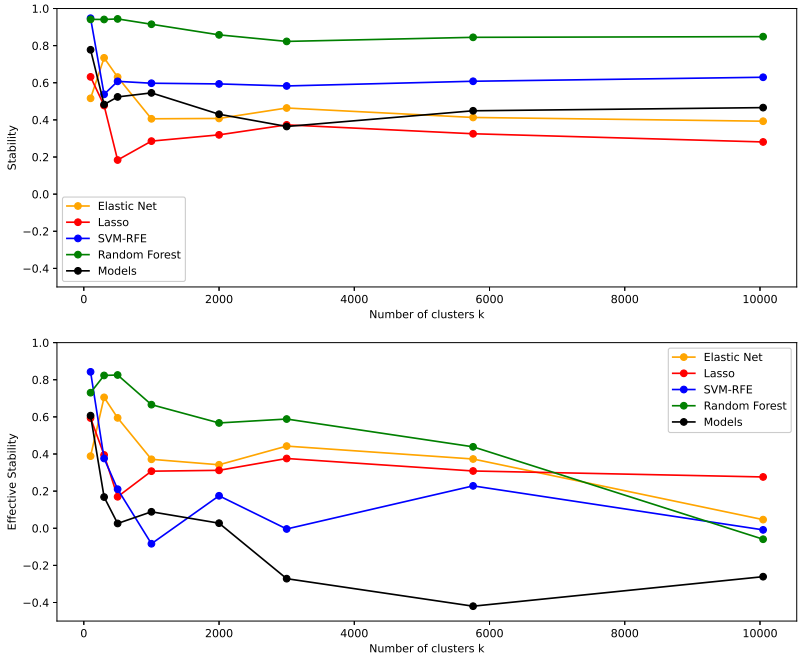


FIGURE 3.89: Variation of stability and effective stability measures according to a different number of clusters on metabolomics data.

that simply decreasing the number of features is not enough to improve the stability value. However, the selection of the optimal value of k is still an open issue since a different number of cluster might be required to achieve a larger stability value.

As a final step, it was also possible to further analyse the results of the PP labelling approach and verify the impact of the unsupervised filter on its feature selection. Tables 3.43 and 3.44 show the values of average test accuracy and R^2 for miRNA and metabolomics respectively by employing the same algorithms used in the classification case. While lasso, elastic net and random forest can be used for both classification and regression tasks, support vector regression has to be exploited as SVM variant for regression. As one can see, the prediction performance improved for most of the models but it is not possible to say which model's version (binary or regression) is the best. Moreover, when it comes to the stability of feature selection, most of the models improved their stability after the unsupervised filter. As shown in Fig. 3.90 and 3.91, in some cases R^2 slightly worsened but the stability was higher; in particular all the stability values increased after the features clustering with metabolomics data. These findings point to the need to define whether it is more important the prediction accuracy or the stability of the selected features in clinical and biological scenarios.

TABLE 3.43: Accuracy and R^2 values before and after unsupervised feature filtering on miRNA data with PP induced labels.

Methods	**miRNA**			
	Avg. Accuracy (before)	Avg. Accuracy (after)	Avg. R^2 (before)	Avg. R^2 (after)
Lasso	0.64	0.68	0.54	0.62
Elastic Net	0.69	0.70	0.63	0.68
SVM-RFE	0.70	0.71	0.65	0.71
Random Forest	0.63	0.63	0.73	0.65

TABLE 3.44: Accuracy and R^2 values before and after unsupervised feature filtering on metabolomics data with PP induced labels.

Methods	**Metabolomics**			
	Avg. Accuracy (before)	Avg. Accuracy (after)	Avg. R^2 (before)	Avg. R^2 (after)
Lasso	0.88	0.91	0.75	0.71
Elastic Net	0.87	0.89	0.76	0.74
SVM-RFE	0.89	0.90	0.70	0.75
Random Forest	0.89	0.89	0.79	0.75

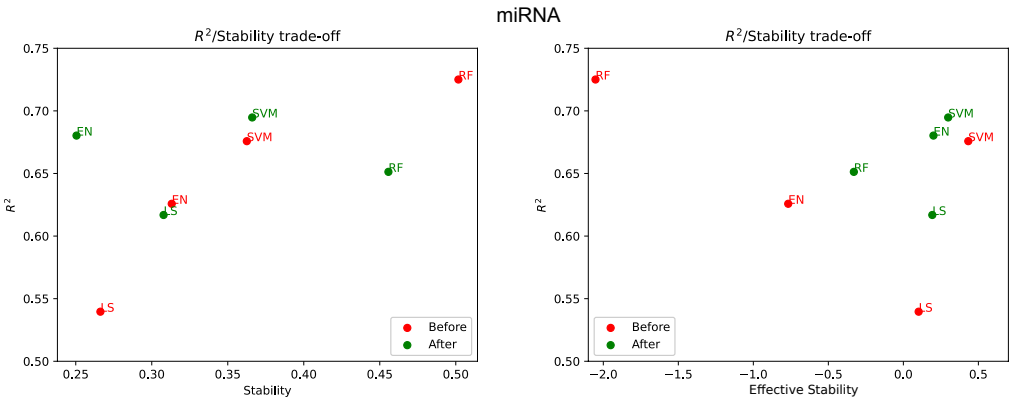


FIGURE 3.90: R^2/Stability trade-off between different feature selection methods using miRNA data before and after unsupervised feature selection.

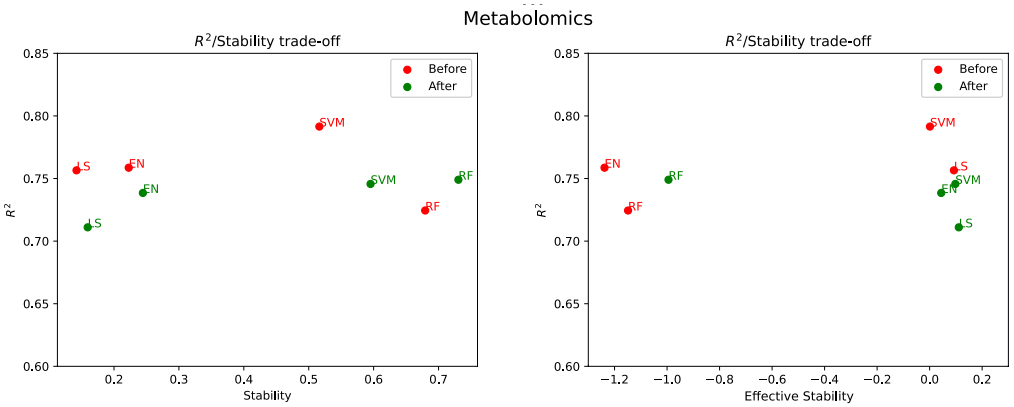


FIGURE 3.91: R^2/Stability trade-off between different feature selection methods using metabolomics data before and after unsupervised feature selection.

Chapter 4

Discussion and conclusions

In this chapter the results exposed so far will be discussed. As already stated, machine learning can have great potential when dealing with clinical, medical and biological scenarios. However, this potential cannot always be exploited by directly using plain general purpose methods. As described in the previous chapter, many issues affecting this kind of analysis have been detected, including:

- Limited sample size.
- High number of missing values.
- Categorical features.
- Correlation among features.
- Confounders and causality.
- Hidden variables.
- Non-binary diagnosis.

All of these aspects need to be taken in consideration if the aim of machine learning for clinical scenarios is to identify the factors causing a disease in order to support the diagnosis and prognosis processes.

Limited sample size and high number of missing values The limited sample size is an issue affecting not only machine learning analysis but also any statistical analysis. Typically, the larger the sample size the more reliable are the results. While in many other fields it is easier to collect a large amount of data, this is not the case of clinical scenarios. In many cases it is challenging to build large data sets, especially when the considered disease is rare and the number of patients is limited in the beginning. If one considers the data sets analysed in this thesis, none of those had a significant sample size and those which reached higher number also had many missing values. This is the case of the dyslexia and autism data set, where the original samples sizes were over 1300 in both cases but after removing the missing values, they dramatically fell to around 400 in the first case and less than 100 in the second

one. This issue comes from the nature of the data set itself: if the data set is built just by including medical records and there are many physicians involved in the process, each doctor might fill some fields and not consider others. At the same time, if data are collected in different time windows, some changing in the way of collecting data may occur. Moreover, another important aspect to consider is the legal one. Being clinical and genetic data considered sensitive, it is not always possible to share the content of the data if the patients did not give their consent. This is for example the case of osteogenesis imperfecta data, where some information is missing due to the absence of patients consent. In fact, the EU General Data Protection Regulation (GDPR) enhances the protection of personal data but also limits the possibility to share data and creates new challenges for scientific research activities [121]. On the other hand, the limited sample size might also be due to the design of the experiment originating the data set. When investigating a specific research question there is often the need to design the experiment from the beginning. Although, the experiment design allows to control as many variables as possible, it is not always possible to recruit many volunteers. In the case of the Oxford Street II data set, for example a total of 60 participants were included in the experiments which is a very limited number under a machine learning point of view. Moreover, some of the participants left the experiment further reducing the sample size. In other cases, some problems with the collection of the data (especially in case of omics data) may occur, generating even more missing values in the data set. Although some imputation techniques exist, it is not always recommended to fill the missing data in this way, especially when the number of missing values is very high. Thus, it would be preferable to adopt more precaution in order to avoid loss of data in the beginning.

Categorical Features The large number of categorical features may affect the data analysis. This issue seems characterise clinical data rather than biological or omics data. Physicians often tend to summarise the patients diagnosis parameters with binary or categorical values which are easier and simpler to understand in practice. However, as shown in the osteogenesis case, this kind of features may lead to unexpected or trivial results. While supervised approaches do not suffer of this problem, clustering and unsupervised methods are more susceptible to the presence of binary features since they tend to create trivial groups which are not necessarily useful from a clinical point of view. On the other hand, when using supervised methods this issue is less evident: if a categorical feature is selected by the algorithm it means it is actually useful for the prediction and thus it should be considered (unless it is a confounder). Clearly, when dealing with biological data such as omics data, the amount of categorical features is very limited considering that this kind of variables are typically expressed as continuous values, thus in these cases, categorical features do not affect the analysis. Therefore, if no label is available, one should be careful in performing unsupervised learning on a data set with binary and categorical variables only. Where possible, categorical features should be treated properly. The most

popular method to deal with them is using one-hot encoding: by doing so each category is transformed into a binary set that can be used by any machine learning algorithm. However, when many categories are present, this method will generate as many new binary columns as the number of the categories increasing not only the dimensionality of the data set but also the number of binary columns. Moreover, increasing the number of binary features, not only increases the dimensionality but can also bring other problems such as the creation of perfectly collinear features which can affect the performance of some algorithms. On the other hand, categorical variables expressed as ordinal values, do not need such approach and can be treated as continuous features, therefore they are preferable in these scenarios.

Correlation among features While categorical features mainly affect clinical data sets, the presence of highly correlated features is often an issue with biological and omics data. In these cases, the number of features is typically very high, especially if compared to the the sample size. Moreover, many features coming from the same biological or molecular pathway, or having similar methylation profile may show high values of correlation or redundancy [122]. This is the case of the omics data in the Oxford Street II data set. As shown in the previous chapter, the redundancy among features not only affected the feature selection process but also the performances of the algorithms. Reducing the number of available features also improves the efficiency of learning tasks [123]. Dealing with highly correlated features is thus imperative and specific approaches to mitigate this issue are needed. This is the reason why an unsupervised clustering method has been proposed as new approach to group features according to their similarity and reducing the data set dimensionality a priori, independently from the labels. Indeed, even feature selection can bring an overfitting component: using a unsupervised feature selection step can reduce overfitting drammatically. The above shown results confirm that mitigating feature redundancy can bring advantages in terms of both accuracy and stability of feature selection.

Confounders and causality Another important issue affecting clinical and medical research is the causality relationship between the features and the outcome of the analysis. One of the aim of clinical machine learning is to identify features that can be used to predict or better understand a disease. In these scenarios it is crucial to identify features or biomarkers that are not only correlated with the outcome but also have a causal effect on it. This need is even more evident when confounding factors are present in the data set such as in the case of the Oxford Street II data set. Here, at least three possible confounders have been identified: age, sex and BMI of the participants. As confounders, they are highly correlated with the outcome and can be used to make the prediction, but they lack a proper casual relationship with the diagnosis. For example, in this specific data set most of the samples affected by IHD are actually males with high values of BMI and older than the healthy ones.

Due to this, the algorithms can simply use these features to make the prediction but they do not take into account that there is no a real casual relationship among the features and the disease. Taking care of the confounding factors is crucial in order to remove at least this misleading effect. One of the approaches that can be used to mitigate this phenomenon, is using the residuals of a simple linear regression model as described in paragraph 3.4.6. Although this approach can help reducing the effects of the confounding variables, it still presents some limits, such as the assumption of a linear relationship between the confounders and the other features. The presence of confounders is often not well managed by the algorithms themselves. As the previously exposed results have shown, most of the methods employed tend to select some relevant feature (such as BMI) and others correlated with it. Nevertheless, the final features should be orthogonal among themselves. This issue can be partially solved by filtering the features in an unsupervised way as described in the previous chapter, but it still does not guarantee that the final features have a causal relationship with the disease. In order to get this information it is necessary to verify whether the features have been experimentally validated and there is a known relationship with a disease. Clearly with this approach it is not possible to discover if a new variable can be a proper biomarker, unless an ad hoc experiment is set up. As alternative, there is a need to develop specific machine learning algorithms which take into account of causality and can identify more relevant features [124].

Hidden variables An interesting aspect that has been observed in this analysis is the possible presence of hidden variables. This is mainly visible when considering clinical data sets: here one can see that it is not always possible to classify patients diagnosis according to the clinical features included in the data set. Considering that the diagnosis process is based on the information collected by the physician, the data reported should be those used to define the label. Therefore, a model trying to predict the diagnosis should score 100% of accuracy for the prediction, at least in principle. As shown in the previous chapter, this is not the case: this implies that there is something not correct in the data or the learning method is confused by some factors. Some problems could arise from mistakes in filling the data, such as typos or missing values: if some values are mistakenly reported, the algorithm may be affected by the wrong values and not classify correctly the patients. In other cases, if the data set is large and has been filled by different clinical structures, such as in the case of the osteogenesis data, many doctors with different experience and observations could have compiled the clinical information, unintentionally introducing biases in the data. This means that the diagnosis process, particularly for bordeline disorders, is possibly ambiguous, at least based on the available collected data. If such subjectivity is introduced (either by inconsistent labeling, or hidden variables), the process of learning gets frustrated and it is not possible for the machine to predict the diagnosis in a correct way. This creates the need to identify these hidden variables and either finding a way to define and introduce them in the data set or to

remove this effect from the beginning. This issue is even more visible in those cases where the disease might be difficult to observe such as in the case of dyslexia or autism disorder, especially when the discrimination is related to identify the specific disorder rather then if the patients is sick or not. In the same way, when considering the osteogenesis cases, the true difficulty stands in defining the specific type of OI rather than the disease itself. The differences among the different types of OI are very subtle and the doctor himself can find some complications into assigning the correct type. This would explain why it was easier to classify certain types of OI (e.g. type I versus type III or IV) while it was not possible to do the same with others (type III versus IV). Even in the cases of good algorithm performances, the accuracy values never reached 100% showing that some information was missing from the data set. The same issue does not concern biological and omics data. In these cases, one does not have clinical information about the patients and the diagnosis only is available. Therefore one does not expect to reach perfect accuracy using omics data only and the relationship between the features and the outcome is more complex. Nevertheless, using this kind of data, it was possible to reach higher values of accuracy when compared to clinical data sets. Obviously, the presence of hidden variables do not affect biological data since the results come from a specific analysis (such as the mass spectrometry for metabolomics) and no subjective opinion from the physician is required. Therefore, when dealing with clinical scenarios it is important to define a priori some standards for the diagnosis assignation in order to prevent the introduction of any bias as much as possible. Omics data is also characterised by biases such as batch effects, but they can be removed ab initio before any machine learning activity [125], [126].

Non-binary diagnosis A final aspect that has been observed and analysed in this thesis, is related to the way the diagnosis is defined. In all the considered data sets, the diagnosis is expressed as a binary value (when only healthy and diseased data are available) or as a categorical value (when multiple diagnoses are present). Defining a disease as a discrete value is surely the simplest and most intuitive way to establish if a person is sick or not since it allows to avoid identifying possibly ambiguous intermediate status. Although its simplicity, there might be some cases where a binary diagnosis is not the best choice. As described in the previous chapter (see section 3.4.7), in some scenarios one wants to know not only if patients are sick but also how much they are sick, i.e. in which stage of the disease they are. This information is useful especially if one wants to study not only the diagnosis of patients but also their prognosis. If an algorithm is able to predict the stage of the disease, it would be easier not only to treat it in a more precise way according to the intensity of the disease, but also getting information about how the disease will evolve and possibly preventing or delaying the process. This aspect is also quite intuitive since in reality the process of getting a disease is often not immediate but it follows a progression, even if this progression is not easily observable and other instruments are needed,

such as the study of a patient's omics. The identification of a continuous diagnosis has been one of the research interests of this thesis. Therefore, a way to convert a binary classification to a process which is continuous has been implemented allowing to face the problem as a regression task. Despite some positive results, this aspect needs further investigations in order to improve the conversion process and fully validate this hypothesis. As alternative, another possibility is to create a priori a data set including patients who are studied for a specific period of time, carefully observing the evolution of a disease, if present. If the patients data would be collected in a continuous way, there would be not the need to convert the diagnosis as it would be expressed as a continuous value from the beginning.

4.1 Conclusions

Clinical data sets are an emerging field of application for machine learning algorithms. Even though these methods proved successful, their systematic application to medical and biological data remains an open issue due to the very nature of the data itself. Although the use of machine learning can provide better insights into medical analysis such as the prediction of pathologies or the identification of their causes, several issues need to be addressed in order to simplify and improve the learning process. In this thesis, different clinical and biological data sets were analysed with the aim to predict a patient's health's state or to stratify patients in subgroups while identifying the features that are most significant for the prediction of the health status. In particular, four different clinical data sets and one omics data set were analysed. For each data set, different issues that should be addressed when performing clinical machine learning have been identified, since they might be a true obstacle in performing the analysis. The first data set considered was related to patients affected by different kind of disorders such as language disorders, dyslexia and other intellectual and learning disabilities. Here, one wanted to predict the patients status of health but performing a multi-class classification did not score good results. At the same time, unsupervised learning failed to identify clearly distinct alternative groups of patients. A similar data set is the one related to patients affected by autism spectrum disorder and other language or cognitive disorders. As in the previous one, healthy samples were not available and performing a classification was tricky especially considering the large number of missing values. However, by converting the problem into binary tasks it was possible to achieve positive results in classifying patients affected by development disorder autism versus all the others. Nonetheless, in both cases, the data sets presented a very large number of missing values, hence increasing the sample size would be useful to improve the reliability of the results. The third and fourth data sets analysed were related to patients affected by osteogenesis imperfecta. The two data sets corresponded to two different versions of the same database including different types of osteogenesis. Here the aim

of the analysis was to identify possible new alternative groups of patients according to their clinical information. Although this goal can be easily achieved by using a clustering algorithm, the large number of categorical features highly affected the analysis. Since only few continuous variables were present, clustering algorithms failed to identify non-trivial groups of patients that could have been useful to understand specific characteristics of patients affected by OI. At the same time, using the original labels to perform a supervised analysis did not bring the expected results: it was possible to positively classify only patients of OI types I and III or I and IV, while other combinations did not achieve significant results. From the analyses of these four clinical data sets few issues affecting the analysis emerged, namely the large number of missing values, the presence of many categorical features and more importantly, the possible presence of not observed hidden variables which determine the diagnosis. All these issues need to be addressed in order to make machine learning a reliable instrument for diagnosis support. If not, the algorithms performance might be poor and the results could be trivial and less useful in practice. When moving to a different kind of data set, such as the multiomics Oxford Street II data set, most of these issues disappeared. Despite omics data do not bear categorical features or hidden variables problems, they brought other difficulties. Being the data set based on a designed experiment, the number of samples was limited in the beginning while the number of features was very high. It is known that high dimensionality brings some trouble when employing machine learning analysis, especially when the number of samples is small. Here it was possible to score very positive results in terms of classification accuracy when trying to classify healthy patients and those affected by ischemic heart disease. The good prediction performance opened the door to other kind of analysis and in particular the identification of the most relevant features. This process has been particularly tricky due to the lack of consensus of the employed machine learning methods. This is due not only to the high dimensionality of the data, but also to the large number of correlated features. While this issue does not affect clinical data sets, in biological and omics data it is very common for the features to be highly correlated since they often belong to the same biological pathway or are related to the same biological processes. The feature selection instability is often overlooked in practice but it is of tremendous importance in clinical and biological scenarios. In these cases the identification of new biomarkers requires that the selected features are reliable. As a first step, one needs to be sure that the algorithms will not choose different features when there are changes in the training data. To solve this issue, bringing from the co-clustering approach, an unsupervised filtering method has been proposed to group features according to their similarity independently from the outcome. By doing so, feature redundancy has been reduced and both stability and prediction accuracy improved for the miRNA and metabolomics cases. Although it was possible to mitigate the feature selection instability, other issues associated with omics data have also been investigated. In particular, the presence of confounders and the study of causality are two important

topics that should be considered when working in these scenarios. In order to get reliable and relevant biomarkers, one needs to consider the possible presence and impact of confounding factors, as it was the case of age, sex and BMI in the considered data set. These features do not explain the diagnosis but their effect tends to affect the algorithms performances: despite their high prediction power, they lack of a causality relationship with the diagnosis, leading to misleading results. While one may take care of confounders by using the residual approach shown in the previous chapter, the study of causality is still an open issue and new ways to make machine learning algorithms consider this aspects are needed [124]. A final matter of interest was the nature itself of the diagnosis. In most clinical and biological data sets the diagnosis is expressed as binary or categorical value and not as a continuous one. Since the diagnosis process is time-dependent, it would be useful to consider the diagnosis as a continuous process allowing not only to study the progression of the disease but also opening the possibility to predict the specific stage of the disease. This would also be useful in creating prognostic tools. Since no continuous data were available, by exploiting the principal path algorithm one can create a path connecting healthy patients to diseased ones. By doing so, one is able to relabel the samples and assign them a progression value leading to the IHD status. Despite the positive results and the potential of this method, further investigations are needed in order to validate this approach, especially when it comes to the validation of the potential biomarkers since the information available in the literature is related to the disease status and not to its intermediate stages.

Considering all these issues and the high potential of machine learning for clinical data, one should approach this research field very carefully since many aspects can highly affect not only the performance but also the meaning of the results. After having deeply investigated the results of these analyses, one can state that in order to facilitate the learning process and getting truly reliable results, clinical data sets should be created ad hoc for this kind of research. In particular, due to the partial unreliability of clinical information and the possible lack of consistency in medical diagnosis, more objective information such as omics data should be preferred. Moreover, in order to reduce the effect of possible confounders, a randomisation of samples is imperative. Hence the creation of this kind of data sets should be set up not too differently from a clinical trial in order to reduce to the minimum all the possible biases introduced by external factors. From this perspective, a smaller but more controlled data set would be preferred to a larger data set where most of the issues that have been described are present. Future research will be focused on the in-depth analysis of these issues to identify new ways to address and mitigate all these problems. Only when all these issues will be solved, one could be certain of the results and trust machine learning algorithms as a true means to support clinical decisions beyond the single cohort and beyond the single, often partial and unsatisfactory, outcome-variable association.

Bibliography

[1] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[3] S. E. Awan, F. Sohel, F. M. Sanfilippo, M. Bennamoun, and G. Dwivedi, "Machine learning in heart failure: Ready for prime time," *Current opinion in cardiology*, vol. 33, no. 2, pp. 190–195, 2018.

[4] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[5] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: Promise and potential," *Health information science and systems*, vol. 2, no. 1, pp. 1–10, 2014.

[6] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, pp. 505–515, 2017.

[7] E. H. Shortliffe, "Chapter 2-design considerations for mycin," in *Computer-Based Medical Consultations: Mycin*, E. H. Shortliffe, Ed., Elsevier, 1976, pp. 63–78, ISBN: 978-0-444-00179-5. DOI: https://doi.org/10.1016/B978-0-444-00179-5.50008-1.

[8] N. Peiffer-Smadja, T. Rawson, R. Ahmad, *et al.*, "Machine learning for clinical decision support in infectious diseases: A narrative review of current applications," *Clinical Microbiology and Infection*, vol. 26, no. 5, pp. 584–595, 2020, ISSN: 1198-743X. DOI: https://doi.org/10.1016/j.cmi.2019.09.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1198743X1930494X.

[9] J. A. Diao, I. S. Kohane, and A. K. Manrai, "Biomedical informatics and machine learning for clinical genomics," *Human molecular genetics*, vol. 27, no. R1, R29–R34, 2018.

[10] K.-H. Yu, C. Zhang, G. J. Berry, *et al.*, "Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features," *Nature communications*, vol. 7, no. 1, pp. 1–10, 2016.

[11] V. Gulshan, L. Peng, M. Coram, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[12] H. Li, P. Boimel, J. Janopaul-Naylor, *et al.*, "Deep convolutional neural networks for imaging data based survival analysis of rectal cancer," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 846–849. DOI: 10.1109/ISBI.2019.8759301.

[13] E. Kalafi, N. Nor, N. Taib, M. Ganggayah, C Town, and S. Dhillon, "Machine learning and deep learning approaches in breast cancer survival prediction using clinical data," *Folia biologica*, vol. 65, no. 5/6, pp. 212–220, 2019.

[14] S. Liu, S. Liu, W. Cai, S. Pujol, R. Kikinis, and D. Feng, "Early diagnosis of alzheimer's disease with deep learning," in *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, IEEE, 2014, pp. 1015–1018.

[15] T. Brosch, R. Tam, A. D. N. Initiative, *et al.*, "Manifold learning of brain mris by deep learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2013, pp. 633–640.

[16] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[17] B. Ristevski and M. Chen, "Big data analytics in medicine and healthcare," *Journal of integrative bioinformatics*, vol. 15, no. 3, 2018.

[18] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "–omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, 2016.

[19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[20] J. Benitez, J. Castro, and I. Requena, "Are artificial neural networks black boxes?" *IEEE Transactions on Neural Networks*, vol. 8, no. 5, pp. 1156–1164, 1997. DOI: 10.1109/72.623216.

[21] C. B. Azodi, J. Tang, and S.-H. Shiu, "Opening the black box: Interpretable machine learning for geneticists," *Trends in genetics*, vol. 36, no. 6, pp. 442–455, 2020.

[22] M. J. Ferrarotti, W. Rocchia, and S. Decherchi, "Finding principal paths in data space," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 8, pp. 2449–2462, 2018.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[24] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[25] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2346178.

[26] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 42, no. 1, pp. 80–86, 2000, ISSN: 00401706. [Online]. Available: http://www.jstor.org/stable/1271436.

[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, ISSN: 13697412, 14679868. [Online]. Available: http://www.jstor.org/stable/3647580.

[28] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007, ISSN: 19326157. [Online]. Available: http://www.jstor.org/stable/4537438.

[29] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] J. Mercer, "Xvi. functions of positive and negative type, and their connection the theory of integral equations," *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, vol. 209, no. 441-458, pp. 415–446, 1909.

[32] H. Drucker, C. J. Burges, L. Kaufman, A. Smola, V. Vapnik, *et al.*, "Support vector regression machines," *Advances in neural information processing systems*, vol. 9, pp. 155–161, 1997.

[33] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[34] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[35] A. Lebedev, E. Westman, G. Van Westen, *et al.*, "Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness," *NeuroImage: Clinical*, vol. 6, pp. 115–125, 2014.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[37] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. DOI: 10.1109/TIT.1982.1056489.

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[39] S. Vassilvitskii and D. Arthur, "K-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2006, pp. 1027–1035.

[40] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, 1953.

[41] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336 –3341, 2009, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2008.01.039. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741740800081X.

[42] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[43] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou, Eds., vol. 17, MIT Press, 2005. [Online]. Available: https://proceedings.neurips.cc/paper/2004/file/40173ea48d9567f1f393b20c855bb40b-Paper.pdf.

[44] M. Bertazzo, D. Gobbo, S. Decherchi, and A. Cavalli, "Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy," *Journal of chemical theory and computation*, vol. 17, no. 8, pp. 5287–5300, 2021.

[45] E. Ragusa, P. Gastaldo, R. Zunino, M. J. Ferrarotti, W. Rocchia, and S. Decherchi, "Cognitive insights into sentic spaces using principal paths," *Cognitive Computation*, vol. 11, no. 5, pp. 656–675, 2019.

[46] E. Gardini, M. J. Ferrarotti, A. Cavalli, and S. Decherchi, "Using principal paths to walk through music and visual art style spaces induced by convolutional neural networks," *Cognitive Computation*, vol. 13, no. 2, pp. 570–582, 2021.

[47] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

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

[49] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[50] H. Kang, "The prevention and handling of the missing data," *Korean journal of anesthesiology*, vol. 64, no. 5, p. 402, 2013.

[51] T. R. Sullivan, A. B. Salter, P. Ryan, and K. J. Lee, "Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data," *American journal of epidemiology*, vol. 182, no. 6, pp. 528–534, 2015.

[52] H. S. Obaid, S. A. Dheyab, and S. S. Sabry, "The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning," in *2019 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference (IEMECON)*, IEEE, 2019, pp. 279–283.

[53] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[54] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[55] D. K. McClish, "Analyzing a portion of the roc curve," *Medical decision making*, vol. 9, no. 3, pp. 190–195, 1989.

[56] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowledge and information systems*, vol. 12, no. 1, pp. 95–116, 2007.

[57] K. Sechidis, K. Papangelou, S. Nogueira, J. Weatherall, and G. Brown, "On the stability of feature selection in the presence of feature correlations," in *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis, and C. Robardet, Eds., Cham: Springer International Publishing, 2020, pp. 327–342, ISBN: 978-3-030-46150-8.

[58] S. Nogueira, K. Sechidis, and G. Brown, "On the stability of feature selection algorithms," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6345–6398, 2017.

[59] J. L. Lustgarten, V. Gopalakrishnan, and S. Visweswaran, "Measuring stability of feature selection in biomedical datasets," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2009, 2009, p. 406.

[60] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[61] S. Nogueira, K. Sechidis, and G. Brown, "On the use of spearman's rho to measure the stability of feature rankings," in *Iberian conference on pattern recognition and image analysis*, Springer, 2017, pp. 381–391.

[62] S. E. Shaywitz, "Dyslexia," *Scientific American*, vol. 275, no. 5, pp. 98–104, 1996.

[63] Z. Cui, Z. Xia, M. Su, H. Shu, and G. Gong, "Disrupted white matter connectivity underlying developmental dyslexia: A machine learning approach," *Human brain mapping*, vol. 37, no. 4, pp. 1443–1458, 2016.

[64] A. Loizou and Y. Laouris, "Developing prognosis tools to identify learning difficulties in children using machine learning technologies," *Cognitive computation*, vol. 3, no. 3, pp. 490–500, 2011.

[65] P Tamboer, H. Vorst, S Ghebreab, and H. Scholte, "Machine learning and dyslexia: Classification of individual structural neuro-imaging scans of students with and without dyslexia," *NeuroImage: Clinical*, vol. 11, pp. 508–514, 2016.

[66] R. Nugent, E. Ayers, and N. Dean, "Conditional subspace clustering of skill mastery: Identifying skills that separate students.," *International Working Group on Educational Data Mining*, 2009.

[67] H. A. Varol, S. Mani, D. L. Compton, L. S. Fuchs, and D. Fuchs, "Early prediction of reading disability using machine learning," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2009, 2009, p. 667.

[68] R. U. Khan, J. L. A. Cheng, and O. Y. Bee, "Machine learning and dyslexia: Diagnostic and classification system (dcs) for kids with learning disabilities," *International Journal of Engineering & Technology*, vol. 7, no. 3.18, pp. 97–100, 2018.

[69] R. F. Ferdinand, "Validity of the cbcl/ysr dsm-iv scales anxiety problems and affective problems," *Journal of anxiety disorders*, vol. 22, no. 1, pp. 126–134, 2008.

[70] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[71] M. Rahman, O. L. Usman, R. C. Muniyandi, S. Sahran, S. Mohamed, R. A. Razak, *et al.*, "A review of machine learning methods of feature selection and classification for autism spectrum disorder," *Brain sciences*, vol. 10, no. 12, p. 949, 2020.

[72] F. Thabtah and D. Peebles, "Early autism screening: A comprehensive review," *International journal of environmental research and public health*, vol. 16, no. 18, p. 3502, 2019.

[73] M. Xu, V. Calhoun, R. Jiang, W. Yan, and J. Sui, "Brain imaging-based machine learning in autism spectrum disorder: Methods and applications," *Journal of Neuroscience Methods*, p. 109 271, 2021.

[74] F. Thabtah, N. Abdelhamid, and D. Peebles, "A machine learning autism classification based on logistic regression analysis," *Health information science and systems*, vol. 7, no. 1, pp. 1–11, 2019.

[75] L. E. Achenie, A. Scarpa, R. S. Factor, T. Wang, D. L. Robins, and D. S. McCrickard, "A machine learning strategy for autism screening in toddlers," *Journal of developmental and behavioral pediatrics: JDBP*, vol. 40, no. 5, p. 369, 2019.

[76] C. Küpper, S. Stroth, N. Wolff, *et al.*, "Identifying predictive features of autism spectrum disorders in a clinical sample of adolescents and adults using machine learning," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[77] S. H. Lee, M. J. Maenner, and C. M. Heilig, "A comparison of machine learning algorithms for the surveillance of autism spectrum disorder," *PloS one*, vol. 14, no. 9, e0222907, 2019.

[78] F. Thabtah and D. Peebles, "A new machine learning model based on induction of rules for autism detection," *Health informatics journal*, vol. 26, no. 1, pp. 264–286, 2020.

[79] M. Gök, "A novel machine learning model to predict autism spectrum disorders risk gene," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6711–6717, 2019.

[80] Y. Lin, S. Afshar, A. M. Rajadhyaksha, J. B. Potash, and S. Han, "A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates," *Frontiers in genetics*, vol. 11, 2020.

[81] T. Eslami, F. Almuqhim, J. S. Raiker, and F. Saeed, "Machine learning methods for diagnosing autism spectrum disorder and attention- deficit/ hyperactivity disorder using functional and structural mri: A survey," *Frontiers in Neuroinformatics*, vol. 14, p. 62, 2021, ISSN: 1662-5196. DOI: 10.3389/fninf.2020.575999. [Online]. Available: https://www.frontiersin.org/article/10.3389/fninf.2020.575999.

[82] S. J. Moon, J. Hwang, R. Kana, J. Torous, and J. W. Kim, "Accuracy of machine learning algorithms for the diagnosis of autism spectrum disorder: Systematic review and meta-analysis of brain magnetic resonance imaging studies," *JMIR mental health*, vol. 6, no. 12, e14108, 2019.

[83] A. Le Couteur, G. Haden, D. Hammal, and H. McConachie, "Diagnosing autism spectrum disorders in pre-school children using two standardised assessment instruments: The adi-r and the ados," *Journal of autism and developmental disorders*, vol. 38, no. 2, pp. 362–372, 2008.

[84] Y. Gong, Y. Du, H. Li, X. Zhang, Y. An, and B.-L. Wu, "Parenting stress and affective symptoms in parents of autistic children," *Science China Life Sciences*, vol. 58, no. 10, pp. 1036–1043, 2015.

[85] A. Perry and D. C. Factor, "Psychometric validity and clinical usefulness of the vineland adaptive behavior scales and the aamd adaptive behavior scale for an autistic sample," *Journal of autism and developmental disorders*, vol. 19, no. 1, pp. 41–55, 1989.

[86] D. Sillence, A. Senn, and D. Danks, "Genetic heterogeneity in osteogenesis imperfecta.," *Journal of medical genetics*, vol. 16, no. 2, pp. 101–116, 1979.

[87] F. Rauch and F. H. Glorieux, "Osteogenesis imperfecta," *The Lancet*, vol. 363, no. 9418, pp. 1377–1385, 2004.

[88] M. Chetty, I. A. Roomaney, and P. Beighton, "The evolution of the nosology of osteogenesis imperfecta," *Clinical Genetics*, vol. 99, no. 1, pp. 42–52, 2021.

[89] S. D. Mooney and T. E. Klein, "Structural models of osteogenesis imperfecta-associated variants in the col1a1 gene," *Molecular & Cellular Proteomics*, vol. 1, no. 11, pp. 868–875, 2002.

[90] A. Z. Abidin, J. Jameson, R. Molthen, and A. Wismüller, "Classification of micro-ct images using 3d characterization of bone canal patterns in human osteogenesis imperfecta," in *Medical Imaging 2017: Computer-Aided Diagnosis*, International Society for Optics and Photonics, vol. 10134, 2017, p. 1 013 413.

[91] M. Rousseau, J. Vargas, F. Rauch, *et al.*, "Facial morphology analysis in osteogenesis imperfecta types i, iii, and iv using computer vision," *Orthodontics & Craniofacial Research*, 2021.

[92] P. Vineis, M. Chadeau-Hyam, H. Gmuender, *et al.*, "The exposome in practice: Design of the exposomics project," *International Journal of Hygiene and Environmental Health*, vol. 220, no. 2, Part A, pp. 142 –151, 2017, Special Issue: Human Biomonitoring 2016, ISSN: 1438-4639. DOI: https://doi.org/10.1016/j.ijheh.2016.08.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1438463916301304.

[93] P. Vineis, O. Robinson, M. Chadeau-Hyam, A. Dehghan, I. Mudway, and S. Dagnino, "What is new in the exposome?" *Environment International*, vol. 143, p. 105 887, 2020, ISSN: 0160-4120. DOI: https://doi.org/10.1016/j.envint.2020.105887. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160412020318420.

[94] E. Ponzi, P. Vineis, K. F. Chung, and M. Blangiardo, "Accounting for measurement error to assess the effect of air pollution on omic signals," *PLOS ONE*, vol. 15, no. 1, pp. 1–16, Jan. 2020. DOI: 10.1371/journal.pone.0226102. [Online]. Available: https://doi.org/10.1371/journal.pone.0226102.

[95] B. Hwa Yun, J. Guo, M. Bellamri, and R. J. Turesky, "Dna adducts: Formation, biological effects, and new biospecimens for mass spectrometric measurements in humans," *Mass spectrometry reviews*, vol. 39, no. 1-2, pp. 55–82, 2020.

[96] S. Balbo, R. J. Turesky, and P. W. Villalta, "Dna adductomics," *Chemical research in toxicology*, vol. 27, no. 3, pp. 356–366, 2014.

[97] M. M. J. Mens, S. C. E. Maas, J. Klap, *et al.*, "Multi-omics analysis reveals micrornas associated with cardiometabolic traits," *Frontiers in Genetics*, vol. 11, p. 110, 2020, ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00110. [Online]. Available: https://www.frontiersin.org/article/10.3389/fgene.2020.00110.

[98] F. R. Mancini, J. E. Laine, S. Tarallo, *et al.*, "Microrna expression profiles and personal monitoring of exposure to particulate matter," *Environmental Pollution*, vol. 263, p. 114 392, 2020, ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2020.114392. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0269749119360373.

[99] D. C. Chambers, A. M. Carew, S. W. Lukowski, and J. E. Powell, "Transcriptomics and single-cell rna-sequencing," *Respirology*, vol. 24, no. 1, pp. 29–36, 2019.

[100] K. van Veldhoven, A. Kiss, P. Keski-Rahkonen, *et al.*, "Impact of short-term traffic-related air pollution on the metabolome – results from two metabolome-wide experimental studies," *Environment International*, vol. 123, pp. 124–131, 2019, ISSN: 0160-4120. DOI: https://doi.org/10.1016/j.envint.2018.11.034. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160412018314703.

[101] C. H. Johnson, J. Ivanisevic, and G. Siuzdak, "Metabolomics: Beyond biomarkers and towards mechanisms," *Nature reviews Molecular cell biology*, vol. 17, no. 7, pp. 451–459, 2016.

[102] S. Greenland, J. Pearl, and J. M. Robins, "Confounding and Collapsibility in Causal Inference," *Statistical Science*, vol. 14, no. 1, pp. 29 –46, 1999. DOI: 10.1214/ss/1009211805. [Online]. Available: https://doi.org/10.1214/ss/1009211805.

[103] E. Gardini, F. M. Giorgi, S. Decherchi, and A. Cavalli, "Spathial: An r package for the evolutionary analysis of biological data," *Bioinformatics*, vol. 36, no. 17, pp. 4664–4667, 2020.

[104] S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright, "miRBase: tools for microRNA genomics," *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D154–D158, Nov. 2007, ISSN: 0305-1048. DOI: 10.1093/nar/gkm952. [Online]. Available: https://doi.org/10.1093/nar/gkm952.

[105] C. R. Foster, L. L. Daniel, C. R. Daniels, S. Dalal, M. Singh, and K. Singh, "Deficiency of Ataxia Telangiectasia Mutated Kinase Modulates Cardiac Remodeling Following Myocardial Infarction: Involvement in Fibrosis and Apoptosis," en, *PLoS ONE*, vol. 8, no. 12, A.-P. Gadeau, Ed., e83513, Dec. 2013, ISSN:

1932-6203. DOI: 10.1371/journal.pone.0083513. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0083513.

[106] K. Singh, "Ataxia-Telangiectasia mutated kinase: Role in myocardial remodeling," en, *Journal of Rare Diseases Research & Treatment*, vol. 2, no. 1, pp. 32–37, Jan. 2017, ISSN: 25729411. DOI: 10.29245/2572-9411/2017/1.1077. [Online]. Available: http://www.rarediseasesjournal.com/articles/ataxiatelangiectasia-mutated-kinase-role-in-myocardial-remodeling.html.

[107] M Buemi, "Reduced bcl-2 concentrations in hypertensive patients after lisinopril or nifedipine administration," en, *American Journal of Hypertension*, vol. 12, no. 1, pp. 73–75, Jan. 1999, ISSN: 08957061. DOI: 10.1016/S0895-7061(98)00217-9. [Online]. Available: https://academic.oup.com/ajh/article-lookup/doi/10.1016/S0895-7061(98)00217-9.

[108] A. Nakagawa, A. T. Naito, T. Sumida, *et al.*, "Activation of endothelial β-catenin signaling induces heart failure," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[109] C. Fang, W. Lu, C. Li, *et al.*, "Mir-3162-3p is a novel microrna that exacerbates asthma by regulating β-catenin," *PLOS ONE*, vol. 11, no. 3, pp. 1–16, Mar. 2016. DOI: 10.1371/journal.pone.0149257. [Online]. Available: https://doi.org/10.1371/journal.pone.0149257.

[110] Y. Zhao, C. Wang, C. Wang, *et al.*, "An essential role for wnt/β-catenin signaling in mediating hypertensive heart disease," *Scientific reports*, vol. 8, no. 1, pp. 1–14, 2018.

[111] Y. Su, J. Yuan, F. Zhang, *et al.*, "Microrna-181a-5p and microrna-181a-3p cooperatively restrict vascular inflammation and atherosclerosis," *Cell death & disease*, vol. 10, no. 5, pp. 1–15, 2019.

[112] S. Demissie and L. A. Cupples, "Bias due to two-stage residual-outcome regression analysis in genetic association studies," *Genetic epidemiology*, vol. 35, no. 7, pp. 592–596, 2011.

[113] S. J. Schomisch, D. G. Murdock, N. Hedayati, J. L. Carino, E. J. Lesnefsky, and B. L. Cmolik, "Cardioplegia prevents ischemia-induced transcriptional alterations of cytoprotective genes in rat hearts: A dna microarray study," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 130, no. 4, 1151.e1–1151.e13, 2005, ISSN: 0022-5223. DOI: https://doi.org/10.1016/j.jtcvs.2005.06.027. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022522305010743.

[114] H. M. Siragy, C. Xue, and R. L. Webb, "Beneficial effects of combined benazepril-amlodipine on cardiac nitric oxide, cgmp, and tnf-αproduction after cardiac ischemia," *Journal of cardiovascular pharmacology*, vol. 47, no. 5, pp. 636–642, 2006.

[115] L. Pestarino, G. Fiorito, S. Polidoro, P. Vineis, A. Cavalli, and S. Decherchi, "On the stability of feature selection in multiomics data," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–7. DOI: 10.1109/IJCNN52387.2021.9533806.

[116] Y. Cheng and G. M. Church, "Biclustering of expression data.," in *Ismb*, vol. 8, 2000, pp. 93–103.

[117] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002. DOI: 10.1109/34.990133.

[118] S. Chormunge and S. Jena, "Correlation based feature selection with clustering for high dimensional data," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 3, pp. 542–549, 2018.

[119] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, 1996, pp. 226–231.

[120] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Transactions on information theory*, vol. 21, no. 1, pp. 32–40, 1975.

[121] D. Peloquin, M. DiMaio, B. Bierer, and M. Barnes, "Disruptive and avoidable: Gdpr challenges to secondary research uses of data," *European Journal of Human Genetics*, vol. 28, no. 6, pp. 697–705, 2020.

[122] L. Toloşi and T. Lengauer, "Classification with correlated features: Unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, pp. 1986–1994, 2011.

[123] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

[124] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[125] B. J. Mertens, "Transformation, normalization, and batch effect in the analysis of mass spectrometry data for omics studies," *Statistical analysis of proteomics, metabolomics, and lipidomics data using mass spectrometry*, pp. 1–21, 2017.

[126] J. Čuklina, P. G. Pedrioli, and R. Aebersold, "Review of batch effects prevention, diagnostics, and correction approaches," in *Mass spectrometry data analysis in proteomics*, Springer, 2020, pp. 373–387.