

UNIVERSITAT DE BARCELONA

FUNDAMENTALS OF DATA SCIENCE MASTER'S THESIS

---

# Proteomics Analysis of Septic and Cardiogenic Shock

---

*Author:*  
Carme ZAMBRANA

*Supervisor:*  
Dr. Vicent RIBAS RIPOLL

*A thesis submitted in partial fulfillment of the requirements  
for the degree of MSc in Fundamentals of Data Science  
in the*

Facultat de Matemàtiques i Informàtica

September 2, 2018



UNIVERSITAT DE BARCELONA

*Abstract*

Facultat de Matemàtiques i Informàtica

MSc

**Proteomics Analysis of Septic and Cardiogenic Shock**

by Carme ZAMBRANA

**Introduction:** Shock is a life-threatening condition affecting about one third of patients in the ICU. The most common types of Shock are Septic and Cardiogenic, affecting 62% and 16% of Shock patients, respectively. A rapid and specialized treatment focused on the type of Shock is crucial for reducing its high mortality rate. Unfortunately, current therapies strive to reduce the medical signs present by the patients rather than target the cause of Shock. A better understanding of the mechanisms and pathophysiology of Shock is mandatory for improving its diagnosis. Omic data and Machine Learning techniques make the perfect combination to tackle this challenge.

**Methodology:** In this thesis, a two-step Machine Learning model has been proposed for analysing proteomic data. The model consists of a Feature Selection method, aimed at selecting relevant proteins, followed by a Classification method, whose purpose is to predict the type of Shock. A robust procedure has been designed for selecting the best model, i.e., stable, interpretable and accurate. Since there is no consensus on the best stability measure, an analysis of different metrics has been performed to decide which metric is more suitable for our problem.

**Conclusions:** Promising results have been obtained using the proteomic data collected in the European research project *ShockOmics* from Septic and Cardiogenic Shock patients. The best model, a combination of ReliefF and Random Forest, is capable of perfectly discriminate between these two types of Shock. On top of that, the proposed model selected meaningful proteins which have been extensively studied in the literature for its relation with Septic Shock.



## *Acknowledgements*

Foremost, I would like to express my sincere gratitude to my supervisor, Dr. Vicent Ribas Ripoll, for his immense knowledge, professional guidance and valuable support. I am genuinely grateful to him and *ShockOmics'* researchers for giving me the opportunity of analysing this dataset.

Besides my supervisor, I would like to thank the whole Eurecat eHealth department for their support and kind words which have made these two busy years a little easier. In particular, Dra. Eloisa Vargiu, for introducing this fascinating world called Research to me and being my mentor since then.

This thesis would not have come to a successful completion, without all I had learnt from UB's professors. I would also like to thank my Master's colleges, it has been a pleasure to share with you these two years.

A special thanks to Sergi Rovira Cisterna. His willingness to give his time so generously listening to my concerns and helping me with excellent suggestions has been very much appreciated.

Last but not the least, I would also like to thank my family: my parents for their encouragement during my studies and for giving me the values of hard work and Dra. Joana Zambrana and Narcís Puig for being my role model.

## *Acronyms*

<b>AdaBoost</b>	Adaptative Boosting
<b>AMI</b>	Acute Myocardial Infarction
<b>ANFIS</b>	Adaptive Neuro-Fuzzy Inference Systems
<b>ANN</b>	Artificial Neural Networks
<b>APACHE</b>	Acute Physiology And Chronic Health Evaluation
<b>AS</b>	Anaphylactic Shock
<b>CLS</b>	Classification
<b>CS</b>	Cardiogenic Shock
<b>DS</b>	Distributive Shock
<b>DT</b>	Decision Tree
<b>ESICM</b>	European Society of Intensive Medicine
<b>ET</b>	Extra Trees
<b>FS</b>	Feature Selection
<b>GA</b>	Genetic Algorithm
<b>GI</b>	Global Importance
<b>HS</b>	Hypovolemic Shock
<b>ICU</b>	Intensive Care Unit
<b>IG</b>	Information Gain
<b>KNN</b>	K-Nearest Neighbour
<b>LDA</b>	Linear Discriminant Analysis
<b>MI</b>	Mutual Information
<b>LR</b>	Logistic Regression
<b>ML</b>	Machine Learning
<b>MLP</b>	Multilayer Perceptron
<b>mRMR</b>	Minimum Redundancy Maximum Relevance
<b>MS</b>	Mass Spectrometry
<b>NB</b>	Naive Bayes
<b>NS</b>	Neurogenic Shock
<b>OOB</b>	Out-of-bag
<b>OS</b>	Obstrusive Shock

**PC** Principal Component

**QDA** Quadratic Discriminant Analysis

**RF** Random Forest

**RFE-SVM** Recursive Feature Elimination - Support Vector Machine

**SS** Septic Shock

**SOFA** Sequential Organ Failure Assessment

**SVM** Support Vector Machine

**ULB** Université Libre de Bruxelles (Hôpital Erasme)

**UNIGE** Université de Genève (Hôpitaux Universitaires de Genève)

**WPCA** Weighted Principal Component Analysis



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Thesis Motivation . . . . .	1
1.2 Thesis Objectives . . . . .	2
1.3 Thesis Structure . . . . .	2
<b>2 Background</b>	<b>5</b>
2.1 Medical Background . . . . .	5
2.1.1 Shock . . . . .	5
Septic Shock . . . . .	6
Cardiogenic Shock . . . . .	6
2.1.2 Clinical Scores . . . . .	7
SOFA Score . . . . .	7
APACHE II Score . . . . .	7
2.2 Omic Background . . . . .	7
2.2.1 Proteome . . . . .	8
2.2.2 Mass Spectrometry . . . . .	8
<b>3 Related Work</b>	<b>9</b>
<b>4 Methodology</b>	<b>11</b>
4.1 Proteomic Data Analysis . . . . .	11
4.2 Stability metrics analysis . . . . .	12
4.3 Methods . . . . .	13
4.4 Feature Selection . . . . .	13
4.5 Classification . . . . .	15
<b>5 Experimental results</b>	<b>17</b>
5.1 Dataset . . . . .	17
5.2 Stability metrics analysis . . . . .	19
5.2.1 Setting up . . . . .	19
5.2.2 Results . . . . .	20
5.3 Proteomic Data Analysis . . . . .	23
5.3.1 Setting up . . . . .	23
5.3.2 Results . . . . .	24
<b>6 Conclusions</b>	<b>29</b>
<b>A Stability metrics</b>	<b>31</b>
<b>B Best Configurations</b>	<b>35</b>

<b>C Selected features</b>	<b>43</b>
<b>Bibliography</b>	<b>49</b>

## Chapter 1

# Introduction

Circulatory Shock, also known as Shock, is a life-threatening generalized form of acute circulatory failure associated with inadequate oxygen utilization by the cells [11]. In other words, the circulation is unable to deliver sufficient oxygen to satisfy the tissues requirements, resulting in cellular injury and tissue malfunctioning. Shock is a common condition in critical care, affecting about one third of patients in the Intensive Care Unit (ICU) [71].

Depending on the underlying cause, Shock can be divided in four main types: Hypovolemic Shock (HS), Cardiogenic Shock (CS), Obstrusive Shock (OS) and Distributive Shock (DS). The last one can be further divided in Septic Shock (SS), Anaphylactic Shock (AS) and Neurogenic Shock (NS). SS is the most common form followed by CS and HS, affecting 62%, 16% and 16% of Shock patients, respectively. The other DS's subtypes (Anaphylactic and Neurogenic) occur in 4% of Shock patients. OS is relatively rare, affecting 2% of Shock patients [71].

The mortality rate of Shock is very high, especially for SS and CS. In Europe, it ranges from 42% to 56% for SS [6, 55, 60, 62] and from 48% to 65% for CS [3, 26, 34, 54]. The incidence<sup>1</sup> in the ICU ranges from 7% to 14% for SS and 6% to 9% for CS.

The support of patients in shock during the first hours is crucial. A rapid diagnosis and treatment is vital for avoiding permanent damage of the tissues. Indeed, resuscitation should be started even while investigation of the cause is ongoing [71]. Nevertheless, establishing the diagnosis of SS and CS is not always easy. Sepsis, the previous syndrome to SS, does not have a validated criterion standard diagnostic test [63]. Even, patients hospitalized with CS occasionally develop SS [11]. Most common CS is due to Acute Myocardial Infarction (AMI); however, the diagnosis of CS and its cause is not always clear [66].

High-throughput technologies, producing vast amount of Omic data<sup>2</sup>, and Machine Learning (ML) techniques, capable of analysing them, make the perfect combination to tackle the challenge of rapid diagnosis of shock patients. This thesis is an step forward to that direction.

### 1.1 Thesis Motivation

Many studies tried to typify Shock by analysing clinical variables through ML techniques [70]. In fact, Shock's types have heretofore been described using clinical scores which rely on clinical variables. Current therapies strive to reduce the medical

<sup>1</sup>These data has been obtained from the same studies cited for the mortality rate.

<sup>2</sup>Referring to data coming from a biology field ending in -omics, i.e, genomics, proteomics, metabolomics, transcriptomics ...

signs of Shock, but they are unable to target its cause. Therefore, a better comprehension of the mechanisms and pathophysiology of Shock is necessary to reduce the high mortality rate.

In the past decade, the volume of Omic data has expanded exponentially due to the advances in biotechnology [56]. These data extensively characterize the behaviour of cells, tissues and organs at the molecular level which is key for understanding the aetiology of human diseases. Consequently, the analysis of Omic data is essential for fathoming the underlying cause of Shock and, hence, improve its diagnosis and prognosis.

*ShockOmics*<sup>3</sup> was an European research project, which ended last year, aimed at identifying molecular biomarkers in acute heart failure induced by Shock. Omic and hemodynamic data from **CS**, **SS** and **HS** patients were collected in a multicenter prospective observational trial [5]. The two published findings were related to the therapy of **SS**. Blood pressure variability, heart rate variability, and baroreflex trends (hemodynamic variables) were studied for understanding the responsiveness to therapy in the early phase of **SS** [10]. In [9], the authors provided evidence that metabolic disturbances play an important role in individual patients' responses to infection.

The data analysed in this thesis is the *ShockOmics*' proteomic dataset. The researchers of the project analysed each type of Shock separately with the aim of depicting its responsiveness to early therapy<sup>4</sup>. In contrast, this thesis focuses on characterizing **SS** and **CS** through proteins.

## 1.2 Thesis Objectives

The aim of this thesis is to characterize **SS** and **CS** through proteomics data. In order to achieve this aim, the objectives of this project are:

- **Create a model capable of discriminating between **SS** and **CS** patients in early stages using proteomic data.** The diagnosis of the type of Shock is crucial for providing the correct treatment.
- **Analyse different stability metrics.** Although the stability of the methods is essential for creating a model capable of generalise to unseen data, there is no consensus on the best stability measure.
- **Identify key proteins for discriminating between **SS** and **CS** in early stages.** These proteins are fundamental for a better understanding of each type of Shock.

## 1.3 Thesis Structure

This thesis is organized in six Chapters.

**Chapter 2** provides the relevant background to understand the analysed data. It is divided in two parts: Medical and Omic Background. In the former part, the pathophysiological mechanisms, clinical signs and the type of Shock are explained, especially for **SS** and **CS**. Moreover, the two clinical scores most used

<sup>3</sup><http://www.shockomics.org/home>

<sup>4</sup>[https://cordis.europa.eu/result/rcn/213996\\_en.html](https://cordis.europa.eu/result/rcn/213996_en.html)

in the **ICU**, Sequential Organ Failure Assessment (**SOFA**) and Acute Physiology And Chronic Health Evaluation (**APACHE**) II, are defined. In the latter part, Omic technologies and types of Omic data are explained, focusing on Mass Spectrometry (**MS**) technique and proteomic data.

**Chapter 3** reviews previous relevant studies in which **MS** data were analysed. Since the disease or condition studied in the reviewed papers is not Shock, the focus of this chapter is the methodology used rather than the results obtained.

**Chapter 4** explains the methodology used for analysing proteomic data, including the procedure for selecting the best model and an analysis of different stability metrics. The proposed model consist on two steps: Feature Selection (**FS**) and Classification (**CLS**). A brief introduction to each step is given followed by a more detailed explanation of the experimented methods.

**Chapter 5** presents the setting-up and the obtained results of both experiments, analysis of stability metrics and proteomic data. A previous analysis of the dataset is given at the beginning of the chapter, comparing the descriptive variables of the different populations and visualising the proteins' intensities.

**Chapter 6** concludes the thesis discussing the obtained results and presenting the future lines of work.



## Chapter 2

# Background

The aim of this chapter is to provide the necessary background to understand the analysed data. Thus, the chapter is divided into two sections: Medical and Omic Background. In the former section, Shock, especially **SS** and **CS**, are described. In addition, the two main used Clinical Scores, **SOFA** and **APACHE II**, are detailed. In the later section, Omic technologies and types of Omic data are explained, focusing on **MS** technique and proteomic data.

## 2.1 Medical Background

The section 2.1.1 is mainly based on the Consensus on Circulatory Shock [11].

### 2.1.1 Shock

Shock is a clinical state of acute circulatory failure that can result from one, or a combination, of the following four pathophysiological mechanisms:

- decrease in venous return due to a loss of circulatory volume, i.e., internal or external fluid loss;
- failure of the pump function of the heart that results from a loss of contractility, resulting from ischemia, infarction, myopathy or myocarditis; or a major arrhythmia, such as ventricular tachycardia;
- obstruction due to a pulmonary embolism, tension pneumothorax or cardiac tamponade;
- loss of vascular tone that results in maldistribution of blood flow, due to sepsis, anaphylaxis or spinal injury.

These four mechanisms correspond to the four types of Shock enumerated in Chapter 1. The three first types, **HS**, **CS** and **OS**, are characterized by low cardiac output, whereas **DS** is associated with decreased peripheral resistance. Regarding the oxygen, the former three types are featured by inadequate oxygen transport and the latter type by altered oxygen extraction.

Shock is associated with evidence of inadequate tissue perfusion, clinical signs of this alteration can be visualized through three windows:

- peripheral window: cold, clammy and blue skin
- renal window: low urine output

- neurological window: altered mental characterized by obtundation, disorientation and confusion.

Shock patients typically present arterial hypotension. Nevertheless, low blood pressure is not a prerequisite for defining Shock, since Shock patients with chronic hypertension could have moderate hypotension.

The clinical signs might vary depending on the type of Shock. **DS** is usually characterized by an elevated cardiac output while the other types of shock are associated with low cardiac output. **HS** patients present low blood pressures and volumes, whereas these clinical signs are high in **CS** patients. **OS** is associated with increased pulmonary artery pressure and dilated right sides cavities.

In the following sections **SS** and **CS** are explained in more detail. As mentioned in Chapter 1, these two types of Shock have the worst prognosis among all the types and are the ones studied in this work.

### Septic Shock

**SS** is an advanced stage of Sepsis, which is a life-threatening organ dysfunction caused by a dysregulated host [63]. Sepsis occurs when the body attacks its own tissues and organs instead of the infecting pathogen. When Sepsis is aggravated with persisting hypotension, it becomes **SS**. What differences sepsis from an infection is a dysregulated host response and the presence of organ dysfunction, i.e., the organ does not perform its expected function.

The main origins of infection causing Sepsis, and, hence, **SS** are: respiratory, mainly lung infection; digestive, e.g., perforated viscus, ischemic bowel, cholecystitis, peritonitis, colitis; and urinary, e.g., pyelonephritis and obstructive uropathy [4, 45].

**SOFA** score, which will be explained in detail in section 2.1.2, is used for clinically characterize a Septic patient. The task force of [63] recommend using the difference between the baseline and the currently **SOFA** score for identifying Sepsis, instead of using the score value. More precisely, they suggest that a difference bigger than two points indicates that the patient has Sepsis. The baseline **SOFA** score should be assumed zero unless the patient is known to have preexisting organ dysfunction before the onset infection.

### Cardiogenic Shock

**CS** is a severe state of systemic hypoperfusion due to a cardiac dysfunction, often followed by multiorgan failure [31]. It results from abnormalities of myocardial structure and function, impairment of mechanical function of the heart or cardiac dysrhythmia [66].

The main cause of **CS** is **AMI**, but can also be caused by, among others, Acute Mitral Regurgitation, Ventricular Septal Rupture, Isolate Right Ventricular Failure.

**APACHE** II score, which will be explained in detail in section 2.1.2, correlates with the mortality of patients with **CS** [38]. The value of this score, at the time of admission, is significantly different between patients who survive from those who do not. Specifically, patients with a **APACHE** II score higher than 31.5 are at substantially higher risk of death.

### 2.1.2 Clinical Scores

The clinical scores main used in **SS** and **CS**, **SOFA** and **APACHE II**, are explained in this section, highlighting their main objective and how they are measured/computed.

#### SOFA Score

**SOFA** score was created in a consensus meeting organized by European Society of Intensive Medicine (**ESICM**) in 1994 [72]. At that time, the score was studied with Sepsis patients and, consequently, named *Sepsis-Related Organ Failure Assessment*. Nowadays, the score, renamed as *Sequential Organ Failure Assessment*, is broadly used for organ failure assessment.

The main objective of the consensus was to describe quantitatively and as objectively as possible the degree of organ dysfunction/failure<sup>1</sup> over time in groups of patients or even individuals. The authors pointed out that the score was meant to describe a sequence of complication in the **ICU** patient, not only for predicting mortality.

To make the score accessible to every institution, the evaluation of organ dysfunction/failure was based on simple variables, obtained through easily and routinely measurements. Hence, the number of variables is limited to study six organs: respiratory, cardiovascular, coagulation, central nervous system, liver and renal. Other important organs, such as the gut, are not included because of its complexity. For each organ, its dysfunction/failure is assessed by ranging it from 0, meaning normal function, to 4, meaning most abnormal function. More details of how to assess the score for each organ can be found in [72].

#### APACHE II Score

**APACHE II** score was proposed by Knaus et al. [41] in 1985 for measuring severity of acute disease. This score simplifies the previous version, **APACHE** [40], improving its clinical usefulness.

The score was developed for objectively stratify acutely ill patients prognostically by risk of death. Thus, it can be viewed as a severity disease classification system for evaluating the use of hospital resource or comparing the efficacy of intensive care in different hospitals or over time.

The score is mainly computed by quantifying the degree of abnormality of various routine physiologic measurements. The age of the patients and his/hers previous health status related severe organ system insufficiency and immunodeficiency are also factored into the computation of the score. In particular, **APACHE II** measures 12 physiologic variables: temperature, mean arterial pressure, heart rate, respiratory rate, oxygenation, arterial pH, serum sodium, serum potassium, serum creatine, percentage of hematocrits, white blood counts and Glasgow coma score. Each of this variables is weighted depending of its importance. The final score it ranges from 0 to 71, being 71 the worst prognostic.

## 2.2 Omic Background

Omic technologies utilise high-throughput screening techniques, which are able to conduct millions of experiments simultaneously, to produce high quality data. These

---

<sup>1</sup>Organ failure is organ dysfunction to such a degree that normal homeostasis cannot be maintained without external clinical intervention.

technologies collect, transform and integrate Omic information. Before applying **ML** techniques, the preprocessing of the data using Quality Assurance and Quality Control techniques is advisable [39].

Omic experiments are data-driven, differing from more traditional biological studies, which are hypothesis-driven [33]. In data-driven studies holistic approaches are applied without any previous hypothesis to huge amounts of collected data, obtaining a hypothesis as a conclusion which should be further tested.

Omic data can be divided in four main types: genome, transcriptome proteome and metabolome. Each type represents the complete set of a specific type of molecule: DNA, RNA, protein and metabolites, respectively. In the following section proteome, the type of Omic data used in this thesis, is further explained.

### 2.2.1 Proteome

The proteome is the entire set of expressed proteins in a given type of cell or organism, at a given time and under defined conditions. The term was coined by *Marc Wilkins* in 1994 [74]. Proteomics is the study of the proteome.

Proteins are large complex biomolecules, which perform a critical role within organisms, contributing to the structure, function and regulation of the tissues and organs. Proteins are formed by long chains of amino acids combined in 3-dimensional structures. They can be classified depending on their function: antibody, enzyme, messenger, structural components and transport/storage.

There exist numerous technologies to extract quantitative proteins information from biological samples: **MS**, two-dimensional gel electrophoresis, enzyme-linked, immunosorbent assays, protein arrays and affinity separation. The proteomic data analysed in this thesis were obtained by **MS** technique, thus this technique is extensively explained in the next section.

### 2.2.2 Mass Spectrometry

**MS** is an analytic technique for the characterization of biological samples. The high-throughput ability of this technique makes it suitable for large-scale proteomic profiling, i.e., large lists of proteins are identified from samples that are analysed.

In particular, **MS** measures the mass-to-charge ratio ( $m/z$ ) of molecules. In order to be able to measure it, the molecules must first be electrically charged and changed to gas phase. After that, their  $m/z$  ratios are measured by their movements through an electric or magnetic field, this occurs in a mass analyser. Once measured, the  $m/z$  values are visualised as mass spectra, which describes the molecules present through peaks at the relevant  $m/z$  ratios [65].

Data from proteomic **MS** are typically characterized by very high-dimensional feature spaces but relatively small numbers of samples. Therefore, before applying any discriminative technique the dimensionality of the data must be reduced. Thus, the most common **ML** pipeline applied on proteomic datasets obtained by **MS** consists of a **FS** method followed by a **CLS** method.

## Chapter 3

# Related Work

To the best of my knowledge there has not been any attempt to discriminate between **SS** and **CS** by analysing **MS** proteomic data through **ML** techniques. Nevertheless, proteomic data and **ML** techniques have been extensively used for studying the prognosis and diagnosis of other diseases, e.g., cancer, rheumatoid arthritis, inflammatory bowel disease. The methodology applied in these studies is relevant for this thesis since they used the same type of data. Thus, this chapter revises the methodology of those studies in which **MS** proteomic data is analysed through **ML** techniques, even though the target disease or condition is not Shock.

As mentioned in the previous chapter, a two-step model combining a **FS** method followed by a **CLS** method is the most common **ML** pipeline used to analyse proteomic data. Most studies combine different methods of each step to find the best model. A great variety of **FS** methods has been used: filters and wrappers methods [53, 73, 76], embedded methods [24] and biological based algorithms [46, 57]. Contrary to all these studies, in which the model contains just one **CLS** method, Bhanot et al. [7] presented an interesting voting-based meta-classifier model combining the prediction of multiple well-known classifiers: Artificial Neural Networks (**ANN**), Support Vector Machine (**SVM**), K-Nearest Neighbour (**KNN**), Decision Tree (**DT**) and Logistic Regression (**LR**). Table 3.1 summarises the details of these studies: disease/condition, number of participants, **FS** and **CLS** methods experimented and cross-validation procedure used.

New **FS** methods have been developed using Omic data. Recursive Feature Elimination - Support Vector Machine (**RFE-SVM**), which is one of the most well-known **FS** methods, was proposed in [29] for microarray data. Many variants of this method have been used with proteomic data, e.g. [35, 79]. In [50, 30, 77], the authors presented a new method for reducing the high-dimensionality of proteomics data based on Discrete Wavelet Transformation. However, dimensionality reduction techniques are not advisable in Omic data due to the loss of interpretability [59].

All of the aforementioned studies used a cross-validation procedure, either k-fold cross-validation or leave-one-out cross-validation, for validating its results. However, they do not specified if the **FS** method had been applied inside or outside the validation procedure. As it is pointed out in [37], both steps of the pipeline, **FS** and **CLS**, must be applied inside the validation procedure. Overestimated results would be obtained if the **FS** method is applied over all the dataset, i.e., outside the cross-validation procedure..

The stability of the **FS** method is crucial in Omic data due to their high dimensionality [32]. An stable method is robust to small perturbation in data, i.e., the selected features are not affect. However, only one of the above-mentioned studies reported the stability of the **FS** method selected, [53]. The importance of reporting the stability is highlighted in [36] by comparing different **FS** methods over various datasets.

Study	Disease/Condition	Num. Participants	Feature Selection Methods	Classification Methods	Cross-validation Procedure
[7]	Prostate cancer	332 (259)	Peak selection	meta-classifier	10-fold
[24]	Rheumatoid arthritis	206 (68)	DT	Bagging RF	Leave-one-out
	Inflammatory bowel disease	480 (240)	Bosting	ExtraTrees Bosting	
[46]	Ovarian cancer	200 (100)	GA	KNN	Randomly
[53]	Stroke	42 (21)	IG	DT	10-fold
			ReliefF	KNN	
			SVM	SVM MLP	
[57]	Hepatocellular carcinoma	411 (199)	PSO	SVM	10-fold
[73]	Prostate cancer	326 (82)	Peak selection	LDA	10-fold
				QDA	
				KNN	
				SVM	
[76]	Ovarian cancer	89 (47)	T-statistic RF	LDA	10-fold
				QDA	
				KNN SVM	

TABLE 3.1: Details of Mass Spectrometry proteomic data studies. All the studies are binary trying to discriminate between patients and controls. In [24], the authors used all the methods for analysing both datasets. The third column reports the total number of participants, specifying the number of patients in parenthesis. The abbreviations of the methods are: IG (Information Gain), PSO (Particle swarm optimization), GA (Genetic Algorithm) RF (Random Forest), DT (Decision Trees), KNN (K-Nearest Neighbour), SVM (Support Vector Machine), MLP (Multilayer Perceptron), LDA (Linear Discriminant Analysis), QDA (Quadratic Discriminant Analysis).

## Chapter 4

# Methodology

Due to the lack of a standard methodology for analysing proteomic data found in the literature, this work details a robust procedure for selecting the best model. The focus has been put on stability of the selected features, since it is crucial for this type of data. The proposed procedure has been experimented with state-of-the-art methods.

### 4.1 Proteomic Data Analysis

Motivated by the related work explain in Chapter 3, the model proposed in this thesis for analysing proteomic data is a two-step **ML** pipeline combining a **FS** method and a **CLS** method. The best model is defined by three characteristics: stability of the **FS** method, performance of the **CLS** method and interpretability of the model. Since there is no consensus on the best stability measure, an analysis of different stability metrics has been performed to decide which metric is more suitable for our problem, see section 4.2 for the details. Regarding the metric for measuring the **CLS** method's performance, since both classes are equally important and almost balanced, *Accuracy* metric has been chosen for measuring the performance of the classifier.

For robustly selecting the best model, the data have been stratified randomly split, with 70-30%, in two sets: training and testing set. The former set has been used to select the best model; and the latter set to estimate the performance over unseen data, once the best model has been chosen.

Regarding the selection of the best model, a stratified randomly sampling aimed at selecting the most stable method has been nested in a 5-fold cross-validation procedure for selecting the classifier with the best *Accuracy*. To empirically estimate the stability, the different **FS** methods have been fitted with small perturbations of the original dataset. These new datasets have been obtained by randomly stratified sampling the original dataset, using a 75% of the total amount of samples. This procedure has been done  $B = 100$  runs. Hence, for each **FS** method 100 different subsets has been used for estimating its stability by averaging the similarity of all the pairs of features subsets (i.e.,  $B(B - 1)/2 = 4950$  pairs). Then, the best **FS** method has been applied to the training set of the fold. The resulting dataset with the selected features has been used to train the **CLS** methods and its performance has been reported using the *Accuracy* over the validation set of that fold. Finally, the estimate of the *Accuracy* for each model have been computed averaging the *Accuracies* among the five folds. Once the best model has been selected its performance has been reported over the testing set. Figure 4.1 summarises the proposed procedure.

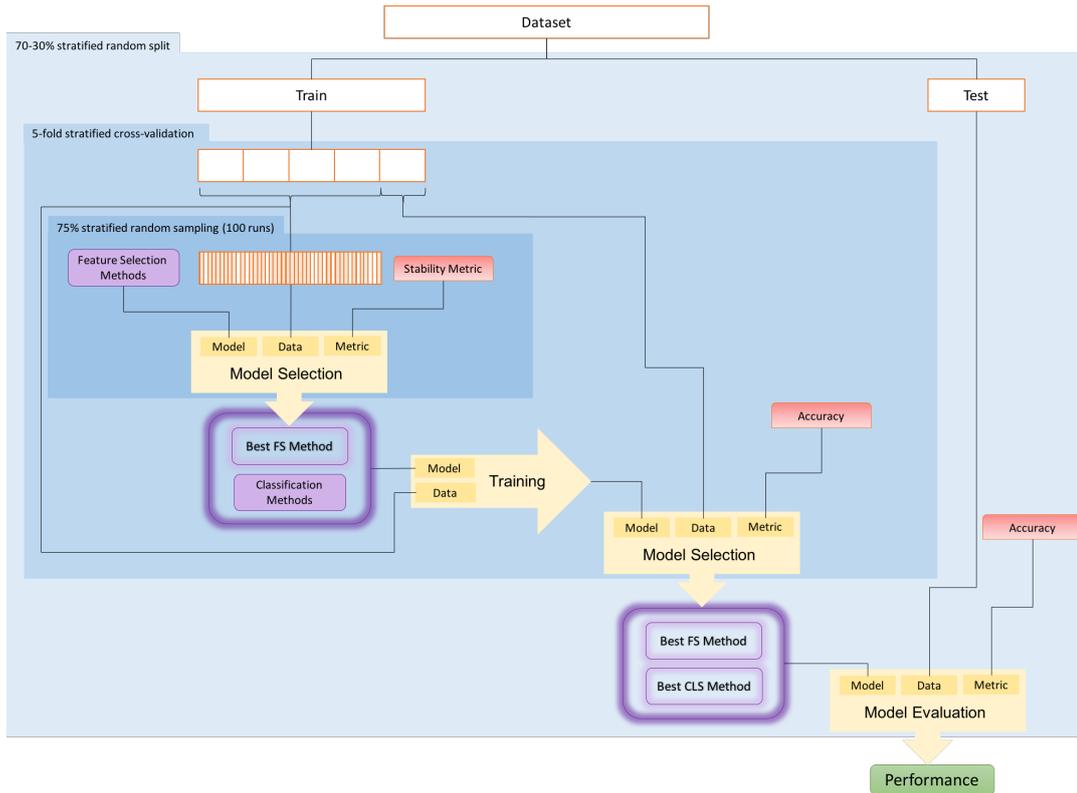


FIGURE 4.1: Diagram explaining the methodology used to select the best model.

As highlighted in [36], choosing the FS method inside the cross-validation procedure decrease the bias of the results, due to the training set used for training the classifier is the same used for selecting the more stable FS method.

## 4.2 Stability metrics analysis

As mentioned in the section 4.1, although there exist various metrics to assess the stability of a FS method, there is no consensus on the best stability measure. These metrics are mainly divided in three groups, depending on how the selected features are reported: subset, ranking or weight/score. Some of the stability metrics reviewed in [32] have been empirically compared in this thesis. In particular, four of the ten stability metrics for subset of features and one of the three for raking features. The weight/score metric has not been used because not all the proposed FS methods for this work report scores. Moreover, as the authors pointed out this metric is the less commonly used.

Regarding the subset metrics, the *Length adjusted stability* has been discarded because it uses a parameter which is not well defined and the default value, given in the original paper [17] is not suitable for our dataset. From the other nine metrics, there are several that are equivalent when comparing two sets with the same cardinality. In particular, *Tanimoto distance* is equivalent to *Consistency*; and *Percentage of overlapping features*, *Dice-Sorensen's index*, *Ochiai's index*, *Relative Hamming distance* and *Weighted consistency* are equivalent among them (see Appendix A for the mathematical proof). Thus, the four subset metric empirically compared in this thesis are: *Tanimoto distance*, *Percentage of overlapping features*, *Percentage of overlapping features*

related and Kuncheva's stability measure. Table 4.1 shows their formulation. All these metrics take values in  $[0, 1]$ , 1 indicates that both sets of features are identical and 0 indicates the sets are completely disjoint.

As regards the ranking features, only the *Canberra distance* has been used because the other metrics take into account the whole dataset and we are only interested in the best features. The smaller the *Canberra distance*, the more stable the method. Table 4.1 shows how this metric is computed.

Metric	Formulation
Tanimoto distance	$1 - \frac{ S  +  S'  - 2 S \cap S' }{ S  +  S'  -  S \cap S' }$
Percentage of overlapping features	$\frac{ S \cap S' }{ S }$
Percentage of overlapping features related	$\frac{ S \cap S'  + c}{ S }$
Kuncheva's stability measure	$\frac{ S \cap S' m -  S }{ S (m -  S )}$
Canberra distance	$\sum_{i=1}^p \frac{ \min\{R(i), k+1\} - \min\{R'(i), k+1\} }{\min\{R(i), k+1\} + \min\{R'(i), k+1\}}$

*c* is the number of features in *S* that are not in *S'* but are significantly positively correlated with at least one feature in *S'*  
*m* is the total number of features in the dataset  
*k* is the number of features selected

TABLE 4.1: Stability metrics between two subsets of features (*S* and *S'*) and Canberra distance between two rankings of features (*R* and *R'*).

In order to decide which of the aforementioned metrics is more suitable for our problem, the four subset metric, *Tanimoto distance*, *Percentage of overlapping features*, *Percentage of overlapping features related* and *Kuncheva's stability measure* and the *Canberra distance* have been empirically compared. All of them have been computed following the stratified random sampling procedure explained in 4.1.

### 4.3 Methods

Five different methods have been experimented for each step of the pipeline. ReliefF, Minimum Redundancy Maximum Relevance (**mRMR**), Weighted Principal Component Analysis (**WPCA**), **RFE-SVM** and **DT** for the first step, **FS**, **SVM**, Random Forest (**RF**), **KNN**, **LR** and Adaptive Boosting (**AdaBoost**) for the second one, **CLS**. This section briefly describes both steps of the pipeline and the methods experimented in each one. A more detailed explanation can be found in the cited original paper of each method.

### 4.4 Feature Selection

**MS** technique for obtaining proteomic data produces large amounts of data. The well-known curse of dimensionality would affect any classifier trained with all these data. For avoiding this curse, the dimensionality of the data must be reduced. Dimensionality reduction techniques, such as those based on projection or compression, are not suitable because the original features are modified and the models are no longer interpretable. Since **FS** methods do not alter the original representation

of the features, but merely select a subset of them, they are more suitable when interpretability is needed. In the literature, FS methods are usually divided in three groups: *filters*, *wrappers* and *embedded* [28]. *Filters* methods assess the relevance of features by looking only at the intrinsic properties of the data. These methods can be further divided depending on how many features are taken into account for selecting the best features: univariate and multivariate methods. The goodness of the selected features can be measured by different metrics, e.g., distance, information, principal component. *Wrappers* methods use learning algorithms to evaluate different subsets of features according to their predictive power. *Embedded* methods perform FS in the process of training a ML method. One method from each of these categories have been experimented in this thesis.

Regarding *filter* methods, only multivariate techniques have been explored since the complexity of diseases cannot be captured by a single protein [16]. Three *filter* methods with different evaluation measures have been used: ReliefF (distance measure), mRMR (information measure) and WPCA (principal component measure).

**ReliefF** estimates the quality of features according to how well their values distinguish between instances that are near to each other. Given a sample, the algorithm searches for its  $k$  nearest neighbours from the same class, called nearest hits, and its  $k$  nearest neighbours from the other class, called nearest misses. Then, the quality of each feature is updated by adding the distances between the sample and each nearest miss and subtracting the distances between the sample and each nearest hit. The key idea is that having a different value between the sample and its nearest hits (misses) means that this feature separates two sample of the same (different) class which is bad (good). This procedure can be done for all the sample (if there are small number of sample) or for a group of randomly selected samples. [58].

**mRMR** is a twofold procedure consisting of maximizing the relevance and the minimizing the redundancy of the selected features. Both concepts are computed by using Mutual Information (MI), which measures the level of similarity. The relevance, defined as the importance of a feature, is computed by measuring the level of discriminant power, i.e., measuring MI between the feature and the class. The redundancy is computed as the similarity between features. The first step of the algorithm selects a feature with maximum relevance. Then, the rest of the features are incrementally selected using the following criterion: the selected feature is the one with the maximum relevance and the minimum redundancy among the previously selected features. This criterion can be combined in just one optimization problem by maximizing the difference or the quotient of the these quantities, relevance and redundancy. Although there exist methods for computing MI on continuous data, better results are reported in the original paper using the discrete version of MI on discretized continuous data. The data is discretized by  $mean \pm t * std$  where  $t$  is a threshold which normally takes the following values 0.5, 1 and 2 [19].

**WPCA** is based on the principal component weights as a measure of the feature importance. The number of Principal Component (PC) to use can be select by the user or it can be used as many features as needed to have at least 90% of the information present in the data. The weights of the selected PC are used to compute the Global Importance (GI) of each feature; together with the importance of the component and the discriminant power of the feature. The first step of the algorithm chooses the feature with higher GI. Then, for

choosing the following features the **GI** of all the remaining features is updated with the correlation between the selected feature and the remaining ones. The feature with the highest **GI** is selected [14].

The well-known *wrapper* method for gene selection **RFE-SVM** has been experimented. Although the method was developed for selecting genes on microarray data, it is also suitable for other Omic data, such as **MS** proteomic data.

**RFE-SVM** recursively eliminates the less important features according to the weights of **SVM**. Firstly, a Linear **SVM** is trained using the whole set of features. Then, the features are ranked by using the weights assigned to them by the classifier. The least important features are eliminated, in this step just one or a set of features can be removed. The procedure is repeated until the required number of features to select is reached [29].

As an *embedded* method a forest of **DT** has been used. After training a forest of trees and importance score for each feature can be calculated and used as a **FS** method.

**DT** are grown to its maximum depth when they are used as **FS** method. At each node, a given number of features are randomly chosen. Then, the samples are split by randomly choosing a value from the range of values of the chosen feature [25]. The score of importance of a given feature is the increasing in mean of the error of a tree in the forest when the observed values of this feature are randomly permuted in the Out-of-bag (**OOB**) samples [23].

## 4.5 Classification

As most of the problems in Omic data, our main objective is to classify binary labelled samples, i.e., we are dealing with a binary supervised classification problem. There exist many **ML** methods to tackle this kind of problem. In this work, five state-of-the-art classifiers have been experimented: **SVM**, **RF**, **KNN**, **LR** and **AdaBoost**. Since all these methods are conceptually different, their complexity and interpretability are different. Although complex methods have good predictive results, they usually suffer from overfitting. Interpretability can be defined as the ability to understand the rationale behind an obtained result by a model. Among these five classifiers, the most complex is **SVM** and the most interpretable is **RF** [42].

**SVM** conceptually classifies the samples by using a hyperplane. When the training set is not perfectly separable, the regularization parameter,  $C$ , determines the weight of the misclassified samples. The shape of the hyperplane is defined by a kernel, which could be, among others, linear or Gaussian, also known as Radial Basis Function (RBF). With RBF kernel, the parameter  $\gamma$  can be used to define the influence of the training samples [13].

**RF** is an averaging ensemble learning method, which builds several independent estimators and then averages their predictions. Each of these estimators is a decision tree, which learns simple decision rules inferred from the features. The size of the random subset of features considered when splitting a node can be delimited by fixing its maximum number of features in the subset [68].

**KNN** is a simple algorithm that uses the  $k$  closest training samples in the feature space to predict the class of a new given sample. Distance between samples is calculated using the Euclidean distance. For a given sample, its predicted

class is the most common one among its neighbours' classes. A weight can be added to the contributions of the neighbours, so that the contribution of the samples is proportional to the distance from the given one. The number of neighbours ( $k$ ) should be a positive integer. In binary classification, when the neighbours contribute with the same weight,  $k$  should be an odd number, avoiding possible ties [15].

**LR** tries to model the posterior probabilities of the class belonging to a particular category. These probabilities are modelled by using a logistic function. Despite its name, it is a linear method for classification. The penalty in the cost function can be done by using either  $L_1$ -norm or  $L_2$ -norm. The regularization parameter,  $C$ , works as in **SVM**, smaller values specify stronger regularization [22].

**AdaBoost** is an adaptive version of the boost by majority algorithm. It sequentially combines a bunch learning algorithms, called weaker learners, into a proper method, called strong learner. For the first weak learner, all the samples have the same weights. For the following learners, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set. The final strong learner is a weighted majority vote of the weaker learners [61].

## Chapter 5

# Experimental results

This chapter starts with an overview of the dataset used in the experiments. Then, the setting up and the results of both experiments analysis of stability metrics and the proteomics data analysis, has been reported.

### 5.1 Dataset

As mentioned before, the data used in this work comes from the European research project *ShockOmics*, where several data sources were collected. More details of the different sources can be found in [5]. Among these sources, human blood samples were collected and pre-processed for Omics analysis. This thesis is focussed on analysing the proteomic data.

More specifically, the blood samples were collected from 48 **ICU** patients: 29 Université de Genève (Hôpitaux Universitaires de Genève) (**UNIGE**) and 19 from Université Libre de Bruxelles (Hôpital Erasme) (**ULB**). The collecting last from October 2014 to March 2016 and during this period a total number of three blood samples were acquired:

- **T1**: acute response shock (16h after **ICU** admission)
- **T2**: steady-state after administration of therapy against the onset shock (48h after **ICU** admission)
- **T3**: steady-state of relevant system level consequences (7 days before **ICU** admission or before **ICU** discharge or before discontinuing therapy (death))

Samples from T1 point were collected before the therapy has taken effect, when the Shock has already activated the main pathophysiological cascades of inflammation and disease. Moreover, as stated before the objective of this thesis is to identify the type of Shock in early stages because a rapid diagnosis during the first hours of the Shock is of paramount importance for avoiding permanent damage of the tissues. Therefore, only T1 data have been used in this study.

As the data come from patients of two different **ICU**, it is sensible to check if there are any significant differences in the main descriptive variables of the two populations. Table 5.1 shows the sex, **ICU** outcome, i.e., alive or dead, the age of the patient when he/she was admitted in the **ICU**, the SOFA score at T1, the APACHE II Score at T1, the number of affected organs and number of days at the **ICU**. All these variables have been computed for the patients grouped by **ICU** and by type of Shock, as well as, for all the patients.

	All	UNIGE	ULB	SS	SC
n	48	29	19	28	20
Sex (Male)	35	23	12	19	16
ICU Outcome (Alive)	39	27	12	23	16
Age	65.71 ± 17.15	67.55 ± 16.57	62.89 ± 18.08	66.21 ± 20.01	65.00 ± 12.57
SOFA	11.75 ± 2.70	12.14 ± 2.62	11.16 ± 2.79	11.82 ± 2.71	11.65 ± 2.76
APACHE II	24.17 ± 7.03	24.83 ± 6.52	23.16 ± 7.81	23.75 ± 6.58	24.75 ± 7.75
Affected Organs	4 ± 0.90	3.95 ± 0.94	4.05 ± 0.85	4.07 ± 0.90	3.9 ± 0.91
Days in ICU	8.13 ± 5.80	7.62 ± 5.62	8.89 ± 6.14	8.61 ± 6.06	7.45 ± 5.50

TABLE 5.1: Descriptive variables of five populations: whole group of patients, those coming from the Hôpitaux Universitaires de Genève, Université de Genève (UNIGE), those from the Hôpital Erasme, Université Libre de Bruxelles (ULB), those who has Cardiogenic Shock (CS) and those with Septic Shock (SS). The categorical variables are reported using the number of samples of the category in parenthesis. The continuous variables are reported with its mean and standard deviation

The *Mann-Whitney U test* has been used for assessing if the continues variables have significant differences between the populations of ICU (UNIGE and ULB) and the population of type of Shock (SS and CS). The null hypothesis of this test is that the two independent samples were selected from populations having the same distribution. As shown in Table 5.2, there are not significant differences between the populations. Thus, the dataset used for the analysis will include all these 48 patients.

	UNIGE-ULB	SS-CS
Age	0.123025	0.189791
SOFA	0.090319	0.491591
APACHE II	0.190137	0.307372
Affected organs	0.403328	0.207085
Days in ICU	0.194995	0.380074

TABLE 5.2: Results (p-values) from the *Mann-Whitney U test* for assessing if the continuous variables have significant differences between the populations of UCI (the Hôpitaux Universitaires de Genève, Université de Genève (UNIGE) and Hôpital Erasme, Université Libre de Bruxelles (ULB)) and between the population of type of shock (Septic Shock (SS) and Cardiogenic Shock (CS)).

The proteomic data consist of 261 different labelled proteins obtained by MS. These data were already preprocessed by *ShockOmics* researchers, so the ML techniques can be applied without further biological processing. Figure 5.1 shows the intensities of all proteins, y-axis, for each patient, x-axis. The patients have been divided by type of Shock and by ICU. In general, the proteins intensities of the SS patients are higher than those of the CS patients, which indicates that the type of shock could be characterized using these proteins. Regarding differences between the samples coming from different ICU, the proteins intensities of the UNIGE's patients are a little bit higher than those of the ULB's patients, especially for the SS patients.

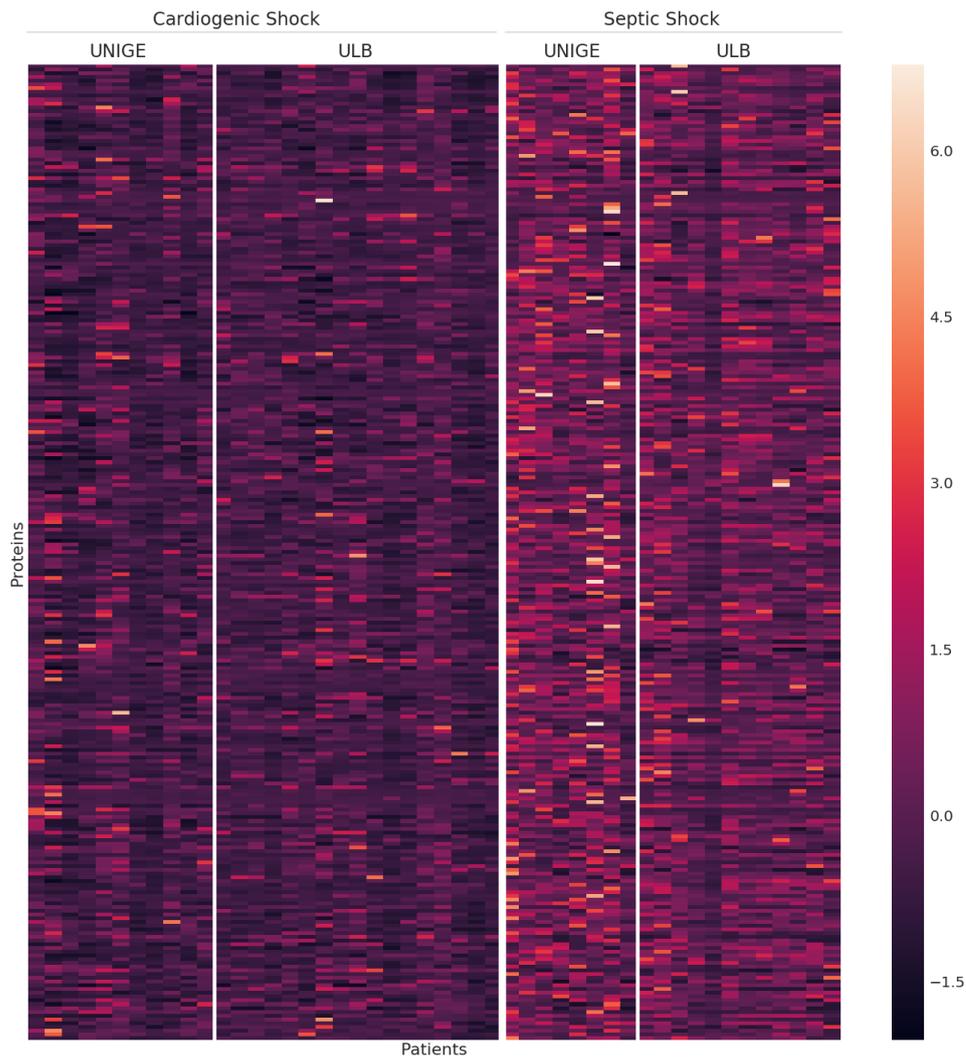


FIGURE 5.1: Heatmap representing the intensities of the proteins for each patient. Each vertical line represents a patients and each horizontal line represents a protein. The patients have been divided by type of Shock (i.e., Cardiogenic and Septic Shock) and by ICU (i.e., UNIGE: Hôpitaux Universitaires de Genève, Université de Genève and ULB: Hôpital Erasme, Université Libre de Bruxelles)

To sum up, the dataset has 48 samples for 261 features. Since the data have more features than samples, the experiments performed must ensure the generalization of the estimated result. Performing a features selection step before applying the classifier will prevent the classifier from suffering the curse of dimensionality.

## 5.2 Stability metrics analysis

### 5.2.1 Setting up

The four subset metric, *Tanimoto distance*, *Percentage of overlapping features*, *Percentage of overlapping features related* and *Kuncheva's stability measure* and the *Canberra distance* have been empirically compared over the five FS methods, ReliefF, mRMR, WPCA,

**RFE-SVM** and **DT**. The default value of its hyper-parameters has been used except for the number of features to select, for which the following values has been experimented 2, 3, 4, 5, 10, 15, 20. Table 5.3 shows the default values of the other hyper-parameter.

Method	Parameter	Range
ReliefF	Number neighbours	10
DT	Number estimators	10
mRMR	Discretization threshold	1
	Method	Mutual Information Difference
RFE-SVM	C of SVM estimator	10

TABLE 5.3: Default hyper-parameters for Feature Selection methods used in the stability analysis.

## 5.2.2 Results

The comparison among the subset metrics is reported using line charts, which are suitable for comparing different trends. In this comparison, the specific number obtained by each metric is not relevant, what is important is to compare the behaviour of the stability among metrics. Figure 5.2 shows the comparison among the four metrics: *Tanimoto distance* (on the top left), *overlapping percentage* (on the top right), *related overlapping percentage* (on the bottom left) and *Kuncheva measure* (on the bottom right). The x-axis of each chart represents the number of selected features and each line correspond to one **FS** method. As shown in the Figure 5.2, the four metrics have similar behaviour among the five **FS** methods. In other words, given a number of features to be select and ordering the five **FS** methods by its stability, the same ranking is obtained for all the metrics. Therefore, one of these four metrics can be used as a representation of the subset metric. We will use *Tanimoto distance*, since it is the most well-known among the four metrics and it is used in [36].

A bar chart has been used to better compare the results obtained by the *Tanimoto distance* and the *Canberra distance* among the five **FS** methods for a given number of feature to be select. Figure 5.3 shows the results of the *Tanimoto distance* for each of the five **FS** methods, each method represented by a different colour. The x-axis represents the number of selected features. The results shows that ReliefF is most stable for all the number of selected features, but for selecting 3 features. In this case, the **WPCA** method is more stable, by 0.049. This method is the second more stable when selecting 2 and 4 features, but its stability decrease considerably when selecting 5, 10, 15 and 20. For these subsets of features the **DT** method is the second more stable, and it is almost as stable as the ReliefF method when selecting 10 features.

Figure 5.4 shows the results of the *Canberra distance* for each of the five **FS** methods, each method represented by a different colour. The x-axis represents the number of selected features. Recall that the smaller the *Canberra distance* the more stable the method. These results show that ReliefF method is the most stable method for all the different number of selected features. In this case, when 3 features are selected, the instability value of **WPCA** is higher than ReliefF by 0.0624 and when 10 features are selected, ReliefF is much more stable than **DT**.

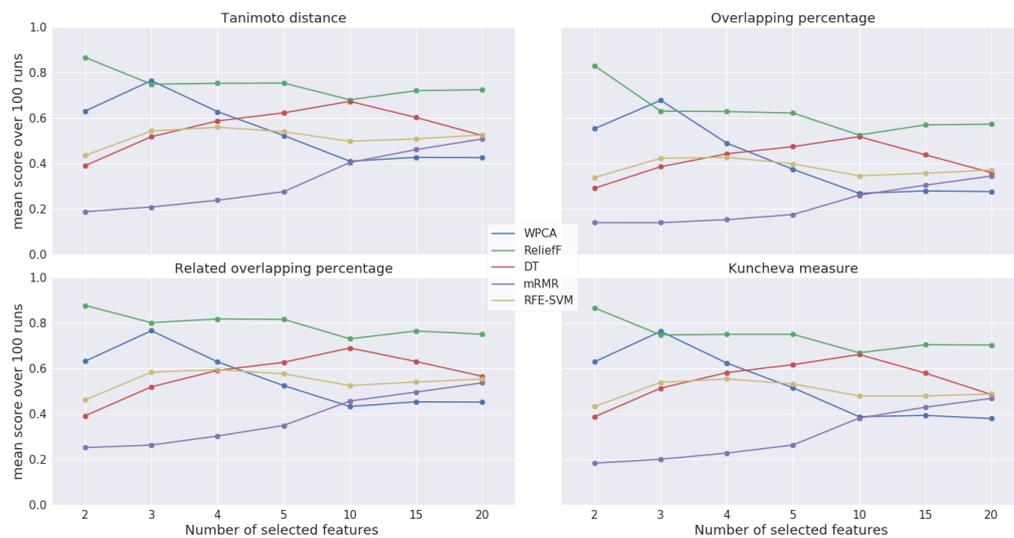


FIGURE 5.2: Comparison between different stabilities metrics for features subsets. Each metric is display in a different plot: *Tanimoto distance* (on the top left), *overlapping percentage* (on the top right), *related overlapping percentage* (on the bottom left) and *Kuncheva measure* (on the bottom right). Each line correspond to one Feature Selection method. The stability has been computed for the following number of selected features: 2, 3, 4, 5, 10, 15, 20.

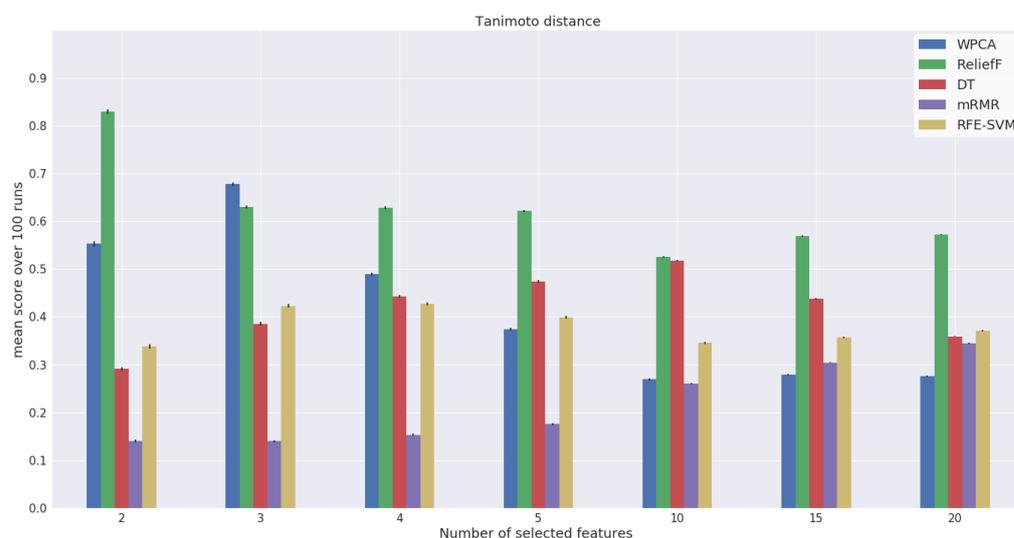


FIGURE 5.3: Results of the *Tanimoto distance* for each of the five Feature Selection methods and for the following number of selected features: 2, 3, 4, 5, 10, 15, 20.

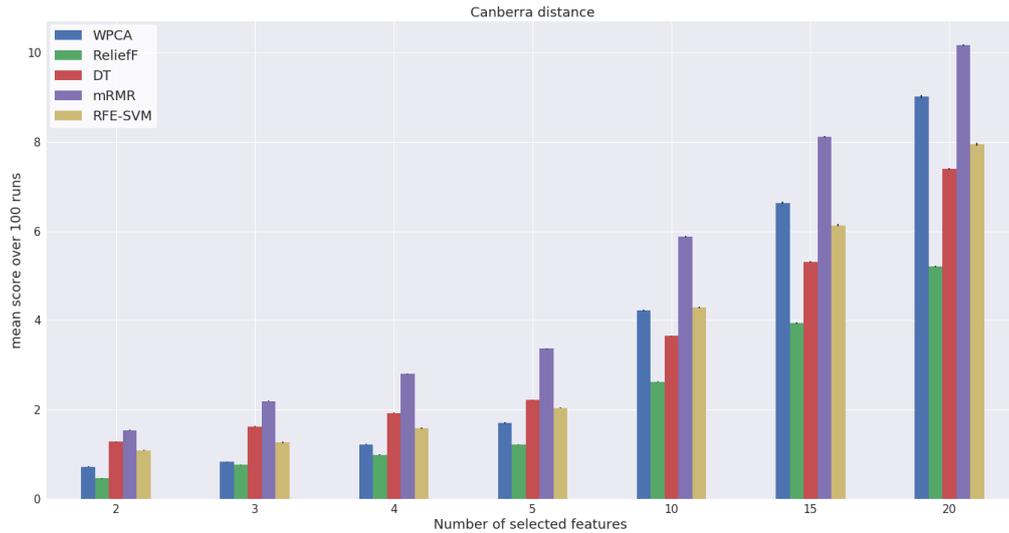


FIGURE 5.4: Results of the *Canberra distance* for each of the five Feature Selection methods and for the following number of selected features: 2, 3, 4, 5, 10, 15, 20.

*Tanimoto distance* and *Canberra distance* are not directly comparable, since the scales are completely different. The former metric ranges from 0 to 1 and the latter metric takes values in the positives. For comparing its results the methods will be ranked from more to less stable. Figure 5.5 shows the comparison between the stability ranking of the FS methods for each feature to be selected. As shown in Figure 5.5, both rankings are equal when selecting 2 features and completely different except for the first position when selecting 20 features. In the other cases only two position of the rankings are interchanged. Focusing on the position through different numbers of features to be selected, both metrics coincide in the first position of the ranking all the times but one, four times for the second and third position, only two times in the fourth and five times in the last position. Thus, there are no big differences in the selected method for both metrics.

Since both metrics have similar results and in the pipeline the CLS method is applied over all the selected subset, i.e., the order in which the features was been selected is not important, *Tanimoto distance* has been used for estimating the stability in the nested cross-validation procedure.

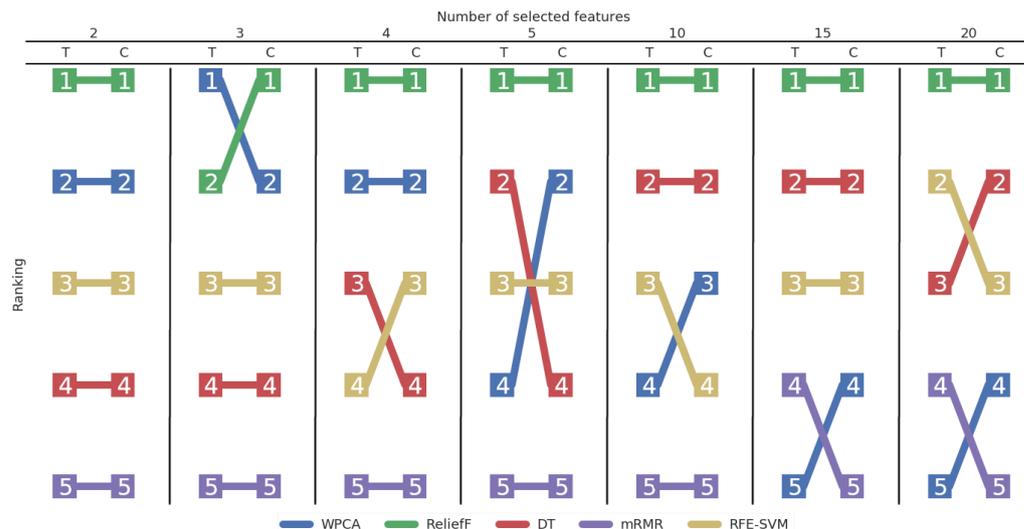


FIGURE 5.5: Comparison of *Tanimoto distance* and *Canberra distance*. For each feature to be selected the position in the stability ranking of the different methods has been compared.

## 5.3 Proteomic Data Analysis

### 5.3.1 Setting up

As a conclusion of the section 5.2, the metric chosen for estimating the stability of the FS methods in the nested cross-validation procedure is *Tanimoto distance*. The performance of the CLS method has been measured by *Accuracy*. For each method, different values of its hyper-parameters have been experimented, Table 5.4 shows them.

Method	Parameter	Range
Relieff	Number neighbours	3, 5, 7, 9, 11, 13
DT	Number estimators	5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47
mRMR	Discretization threshold	0.5, 1, 2
	Method	Mutual Information Quotient, Mutual Information Difference
RFE-SVM	C of SVM estimator	0.01, 0.1, 1, 10, 100, 1000, 10000
SVM	kernel	Linear, RBF
	C	0.01, 0.1, 1, 10, 100, 1000, 10000
	$\gamma$ (RBF)	0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000
RF	Number estimators	5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47
	Maximum features	From 1 to maximum of features
KNN	Number neighbours	3, 5, 7, 9, 11, 13
	Weight	Uniform, Distance
LR	C	0.01, 0.1, 1, 10, 100, 1000, 10000
	Penalty	$L_1, L_2$
AdaBoost	Number estimators	25, 28, 31, 34, 37, 40, 43, 46, 49, 52, 55, 58, 61, 64, 67, 70, 73

TABLE 5.4: Range of each hyper-parameter.

Since the number of sample in the testing set is limited, just 15 samples, the results obtained could be biased from the true generalization error. Permutation test has been used to validate if the obtained results are due to chance. Therefore, the

best model has been retrained 100 times permuting the values of the class. The p-value obtained with this procedure represents the fraction of randomized samples where the model behaves better in random data than in the original data [51].

As in the feature selection experiments, different numbers of features to be selected has been explored. Thus, a total of seven models have been created, each one with a different number of features to be selected: 2, 3, 4, 5, 10, 15 and 20. From now on, each experiment will be called *Exp-x*, where x represent the number of features to be selected.

### 5.3.2 Results

ReliefF has been selected in all the experiments as the most stable method. Table 5.5 shows the average stability of the method over the five folds for each experiment.

	Stability
Exp-2	$0.53 \pm 0.31$
Exp-3	$0.62 \pm 0.26$
Exp-4	$0.45 \pm 0.16$
Exp-5	$0.48 \pm 0.10$
Exp-10	$0.51 \pm 0.10$
Exp-15	$0.57 \pm 0.06$
Exp-20	$0.52 \pm 0.05$

TABLE 5.5: Relief average stability over the five folds for each experiment.

For each experiment, Table 5.6 shows, the cross-validation results, reporting the best accuracy obtained by each family of classifiers and the percentage of configurations which obtained this performance. The results obtained are very high, almost all above the 0.95. As shown, at least one SVM configuration is capable of classifying perfectly all the samples of the validation sets, except in *Exp-20*. This result is expected since complex method have good predictive power, especially when the number of features is small. RF and LR results are also great, obtaining a perfect classification in all the experiments but in two, in *Exp-2* and *Exp-4* for RF and in *Exp-2* and *Exp-3* for LR. In these cases the performance is also good, being  $Accuracy = 0.97 \pm 0.06$ . KNN, the simplest method, is not able to make a perfect classification, though its results are good. AdaBoost obtained the same result for all the configuration, ranging from  $Accuracy = 0.94 \pm 0.11$  to the perfect classification depending on the number of features to be selected.

	SVM		RF		KNN		LR		AdaBoost	
Exp-2	$1.00 \pm 0.00$	(7.94)	$0.97 \pm 0.06$	(100.00)	$0.97 \pm 0.06$	(25.00)	$0.97 \pm 0.06$	(92.86)	$0.97 \pm 0.06$	(100.00)
Exp-3	$1.00 \pm 0.00$	(15.87)	$1.00 \pm 0.00$	(86.67)	$0.94 \pm 0.07$	(100.00)	$0.97 \pm 0.06$	(7.14)	$0.97 \pm 0.06$	(100.00)
Exp-4	$1.00 \pm 0.00$	(7.94)	$0.97 \pm 0.06$	(86.67)	$0.97 \pm 0.06$	(50.00)	$1.00 \pm 0.00$	(7.14)	$1.00 \pm 0.00$	(100.00)
Exp-5	$1.00 \pm 0.00$	(7.94)	$1.00 \pm 0.00$	(95.00)	$0.97 \pm 0.06$	(50.00)	$1.00 \pm 0.00$	(50.00)	$1.00 \pm 0.00$	(100.00)
Exp-10	$1.00 \pm 0.00$	(22.22)	$1.00 \pm 0.00$	(58.52)	$0.97 \pm 0.06$	(16.67)	$1.00 \pm 0.00$	(35.71)	$0.97 \pm 0.06$	(100.00)
Exp-15	$1.00 \pm 0.00$	(23.81)	$1.00 \pm 0.00$	(63.33)	$0.94 \pm 0.07$	(100.00)	$1.00 \pm 0.00$	(21.43)	$1.00 \pm 0.00$	(100.00)
Exp-20	$0.97 \pm 0.06$	(25.40)	$1.00 \pm 0.00$	(58.95)	$0.94 \pm 0.07$	(100.00)	$1.00 \pm 0.00$	(21.43)	$0.94 \pm 0.11$	(100.00)

TABLE 5.6: Best CV-Accuracy for each experiment. The percentage of Configurations with this performance has been reported, between parentheses

Since the interpretability is key in this work and the most interpretable family of classifiers, **RF**, obtained promising results, we will focus on this family. For each experiment, the less complex configuration from the ones which obtained higher CV-*Accuracy* has been evaluated over the testing set. The complexity of the **RF** method is measured by the number of estimators and the maximum features. The higher the number values of those parameters, the higher the complexity.

Table 5.7 shows the best configuration for each experiment and its *Accuracy* over the testing set. The ReliefF's hyper-parameter is the number of neighbours and the **RF**'s hyper-parameter are the number of estimators and the maximum features. All the models perfectly classify the testing set. The complete list of configurations which achieved the best CV-results can be found in the Appendix B, Tables B.1, B.2, B.3, B.4, B.5, B.6, B.7.

	Best Model		Test Accuracy
	Relief	RF	
Exp-2	11	5, 1	1.00
Exp-3	5	8, 1	1.00
Exp-4	5	8, 3	1.00
Exp-5	7	5, 2	1.00
Exp-10	11	17, 4	1.00
Exp-15	13	5, 1	1.00
Exp-20	9	5, 2	1.00

TABLE 5.7: Testing results of the best configuration for each experiment. The ReliefF's hyper-parameter is the number of neighbours. The **RF**'s hyper-parameter are the number of estimators and the maximum features.

Figure 5.6 shows the permutation test result for validating the *Accuracy* obtained over the testing set in each experiment. As shown, the p-values are smaller than 0.05 in all the cases and the results obtained by permuting the target values are less than 0.85. Thus, the results obtained over the testing set are not obtained by chance.

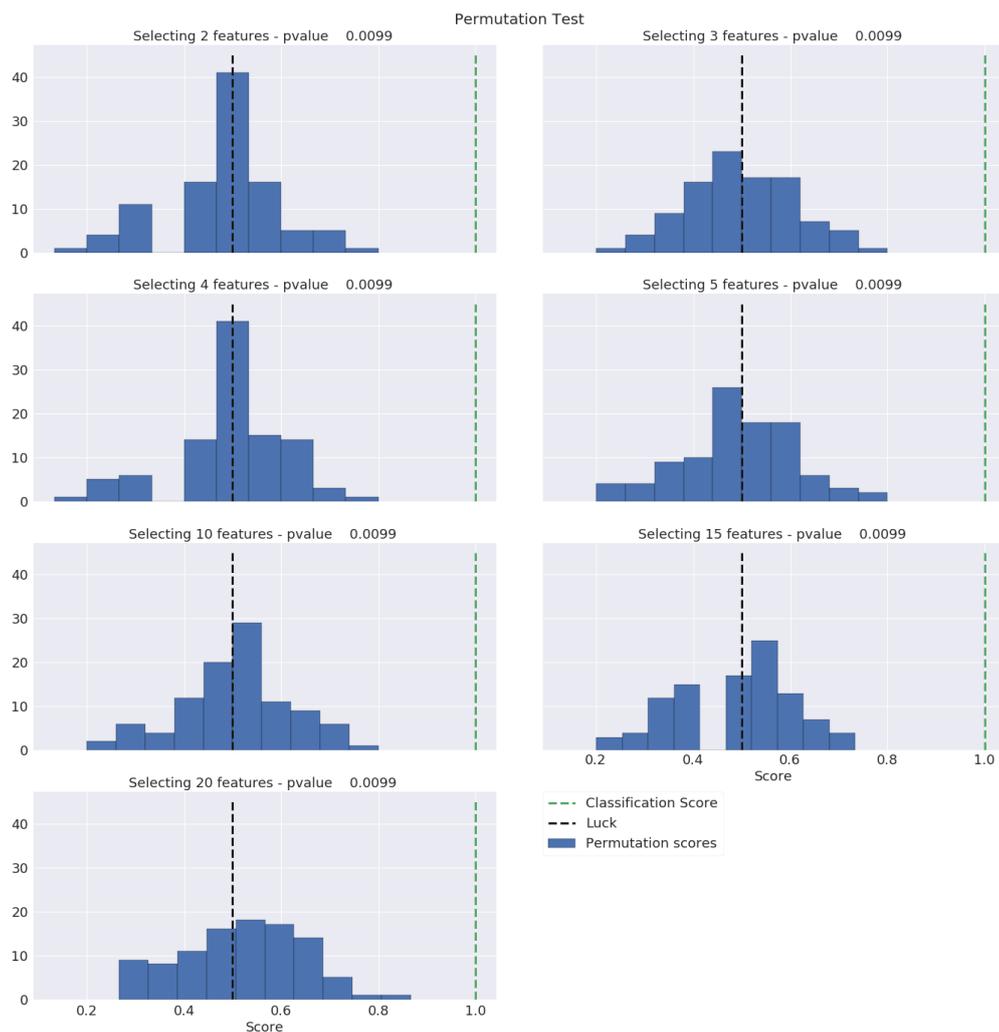


FIGURE 5.6: Permutation test results of the best model for each experiment.

Table 5.8 shows the features selected in each experiment by the best model, ReliefF. All the proteins selected in an experiment have also been selected in the experiments with larger number of features to select. The only exception is the protein *Q15582*, which is selected in the experiment *Exp-15* but not in the *Exp-20*.

Exp-2	Exp-3	Exp-4	Exp-5	Exp-10	Exp-15	Exp-20
P05160						
P03952						
	P43652	P43652	P43652	P43652	P43652	P43652
		P03951	P03951	P03951	P03951	P03951
			P01042	P01042	P01042	P01042
				P54108	P54108	P54108
				P04217	P04217	P04217
				P00748	P00748	P00748
				P05155	P05155	P05155
				Q92954	Q92954	Q92954
					P01031	P01031
					P11597	P11597
					P17936	P17936
					P08253	P08253
					Q15582	Q9UHG3
						P49908
						P05154
						P22891
						P02753
						P02766

TABLE 5.8: Proteins for the best model in each experiment.

Figure 5.7 shows the five selected features in the *Exp-5* using the heatmap in which the intensities of the proteins are plotted for each patient. Figures C.1, C.2, C.3, C.4, C.5 and C.6, in the Appendix C, show the heatmap highlighting the selected features in *Exp-2*, *Exp-3*, *Exp-4*, *Exp-10*, *Exp-15* and *Exp-20*, respectively. The selected features show a clear pattern: the intensities of these proteins are higher for **SS** patients than for **CS** patients.

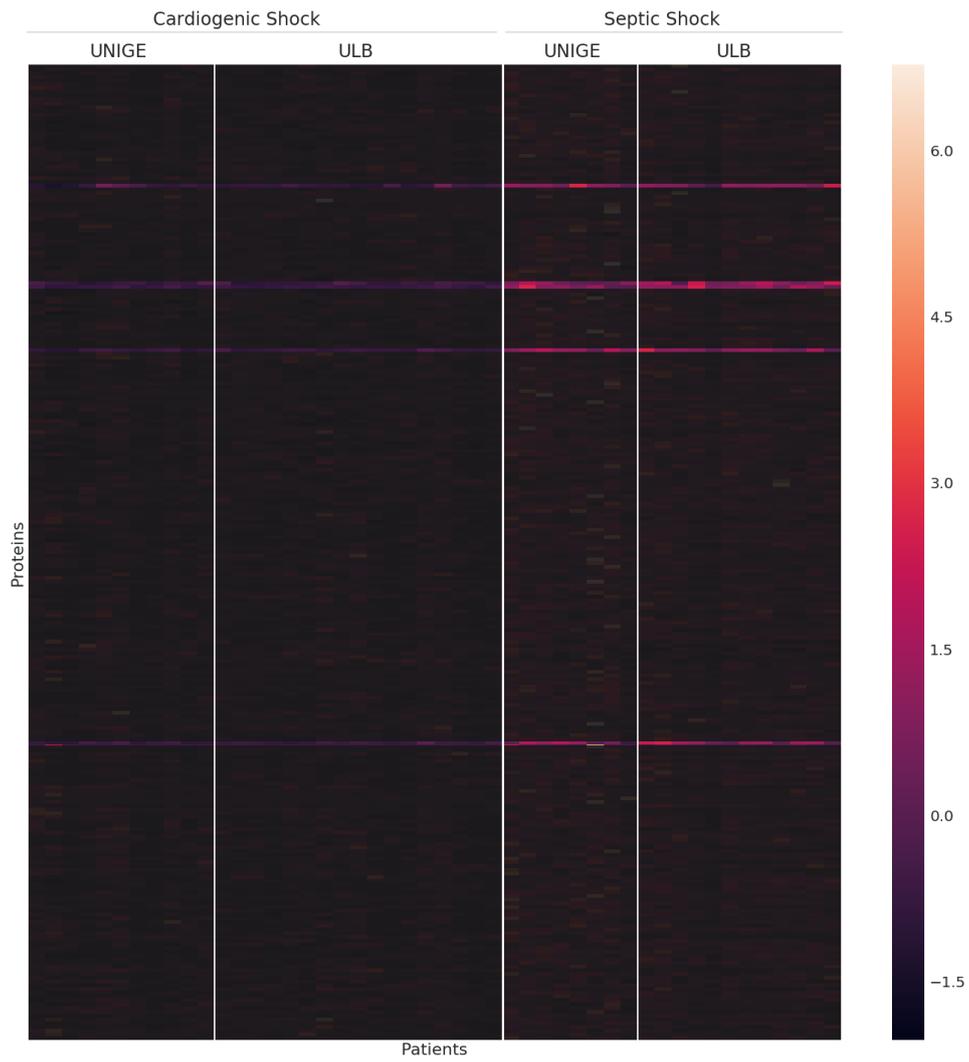


FIGURE 5.7: Heatmap representing the intensities of the proteins,  $y$ -axis, for each patient,  $x$ -axis. The selected proteins selected in *Exp-5* are highlighted

## Chapter 6

# Conclusions

A two-step **ML** model has been proposed for analysing proteomic data. In the first step a **FS** method aimed at selecting relevant proteins and in the second a **CLS** method, whose purpose is to predict the type of Shock. The best model has been obtained by maximizing the stability of the **FS** method and the predictive power of the **CLS** method in a nested validation procedure. The interpretability has also been factored into choosing the best model.

Different stability metrics have been analysed to choose the most suitable one for our problem. Five stability metrics, four subset metrics and one raking metric, have been empirically compared through five **FS** methods and different numbers of features to be selected, 2, 3, 4, 5, 10, 15, 20. All the subset metrics analysed have an equivalent behaviour. Moreover, no big differences have been found between *Tanimoto distance* (subset metric) and *Canberra distance* (ranking metric). In fact, both metrics choose ReliefF as the most stable method in all the cases, except for selecting 3 features. In this case, *Tanimoto distance* selects **WPCA** as the most stable method followed by ReliefF. Nevertheless, the difference between the stability of these two methods is 0.049.

Seven different models have been created, each one with a different number of features to be selected, 2, 3, 4, 5, 10, 15, 20. The best model obtained in all the experiments is composed by ReliefF and **RF**. For all the experiments, the best model perfectly classifies the testing set. Permutation test with p-values smaller than 0.05 ensure that the results obtained over that testing set are not obtained by chance.

Although these results are encouraging, the main limitation of this study is the limited size of the population. Only samples from 48 patients, 28 with **CS** and 20 with **SS** has been analysed. On top of that, the testing set only contains 15 samples, 9 with **CS** and 6 with **SS**. Therefore, we should focus on the proteins selected for obtaining these results rather than the prediction performance per se.

The selected proteins are coherent through the different experiments, meaning that all the proteins selected in an experiment have also been selected in the experiments with larger number of features to select. The only exception is the protein *Q15582*, which is selected in the experiment *Exp-15* but not in the *Exp-20*. The levels of the selected proteins are higher for **SS** patients than for **CS** patients. Table 6.1 contains the description of all the selected proteins ordered by experiment, meaning the two first proteins where selected for the *Exp-2*, the three first for *Exp-3*, and so forth.

Several studies relating most of the selected features with **SS** can be found in the literature: *Coagulation factor XIII B chain* [8, 67, 75, 78], *Plasma kallikrein*, [1, 48, 62], *Afamin* [44], *Coagulation factor XI* [18, 69, 64], *Kininogen-1* [44], *Alpha-1B-glycoprotein* [43], *Complement C5* [12, 49], *Insulin-like growth factor-binding protein 3* [2, 27, 47, 52] and *Selenoprotein P* [20, 21, 80]. Most of these papers analyse the behaviour of the protein comparing survival versus non-survival of **SS** patients or **SS** patients versus

controls. In [48], the authors also point out that *Plasma kallikrein* does not play a role in CS.

Identifier	Protein Description
P05160	Coagulation factor XIII B chain
P03952	Plasma kallikrein
P43652	Afamin
P03951	Coagulation factor XI
P01042	Kininogen-1
P54108	Cysteine-rich secretory protein 3
P04217	Alpha-1B-glycoprotein
P00748	Coagulation factor XII
P05155	Plasma protease C1 inhibitor
Q92954	Proteoglycan 4
P01031	Complement C5
P11597	Cholesteryl ester transfer protein
P17936	Insulin-like growth factor-binding protein 3
P08253	72 kDa type IV collagenase
Q15582	Transforming growth factor-beta-induced protein ig-h3
Q9UHG3	Prenylcysteine oxidase 1
P49908	Selenoprotein P
P05154	Plasma serine protease inhibitor
P22891	Vitamin K-dependent protein Z
P02753	Retinol-binding protein 4
P02766	Transthyretin

TABLE 6.1: Description of the selected proteins ordered by experiment, meaning the two first proteins where selected for the *Exp-2*, the three first for *Exp-3*, and so forth.

To further validate the results obtained in this thesis two main future lines are proposed. Firstly, the models should be applied over other datasets to validate its capability of generalisation to unseen data. Secondly, the validation of the selected features can be assess by comparing the features selected for models trained over different datasets. Another interesting future line is to amplify the analysis of the stability metrics by comparing more FS methods and using other datasets.

It is worth mentioning that the methodology presented in this thesis is general i.e. it does not assume any particularity of the proteomic data, further than its high dimensionality. Therefore, it can be used not only for analysing proteomic data but also for analysing other kinds of Omic data.

## Appendix A

# Stability metrics

This appendix is intended to prove that several subset metrics reviewed in [17] are equivalent when comparing two sets with the same cardinality. In particular, *Tanimoto distance* is equivalent to *Consistency*; and *Percentage of overlapping features*, *Dice-Sorensen's index*, *Ochiai's index*, *Relative Hamming distance* and *Weighted consistency* are equivalent among them. Table A.1 shows the formulation of these metrics.

Metric	Formulation
Tanimoto distance	$1 - \frac{ S  +  S'  - 2 S \cap S' }{ S  +  S'  -  S \cap S' }$
Consistency*	$\frac{1}{ S \cup S' } \left[ \sum_{f \in \{S, S'\}} \text{freq}(f) - 1 \right]$
Percentage of overlapping features	$\frac{ S \cap S' }{ S }$
Sorensen's index	$\frac{2 S \cap S' }{ S  +  S' }$
Ochiai's index	$\frac{ S \cap S' }{\sqrt{ S  S' }}$
Relative Hamming distance	$1 - \frac{ S \setminus S'  +  S' \setminus S }{ S  +  S' }$
Weighted consistency*	$\sum_{f \in \{S, S'\}} \frac{\text{freq}(f)}{ S  +  S' } \cdot \text{freq}(f) - 1$

*freq(f)* denotes the number of occurrences of feature *f* in  $\{S, S'\}$   
 \* in [17] the sum is divided by the number of compared subset minus 1, in our case the denominator is reduced to 1, thus it is omitted

TABLE A.1: Stability metrics between two subsets of features (*S* and *S'*).

In order to prove these equalities we will use the well-known property of cardinalities between sets given in Lemma 1

### Lemma 1.

$$|S \cup S'| = |S| + |S'| - |S \cap S'|$$

*Proof.*

$S \cup S'$  can be break up in three different disjoint sets,  $S \cup S' = (S \setminus S') \cup (S' \setminus S) \cup (S \cap S')$ , as is shown in Figure A.1.

Then, using the property that the cardinality of a disjoint union is the sum of its cardinalities, we obtain:

$$|S \cup S'| = |S \setminus S'| + |S' \setminus S| + |S \cap S'|$$

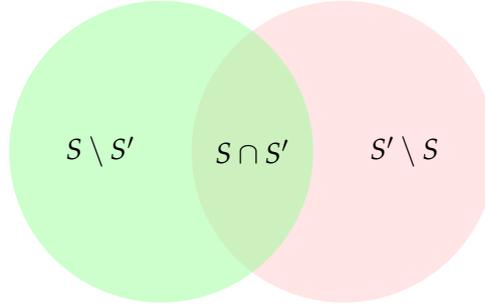


FIGURE A.1: Union of sets  $S$ , green circle, and  $S'$ , red circle.

As shown in Figure A.1,  $S = (S \setminus S') \cup (S \cap S')$  and  $S' = (S' \setminus S) \cup (S \cap S')$ . The cardinality of  $S$  and  $S'$  can be computed using again the property of cardinality of disjoint sets:

$$|S| = |S \setminus S'| + |S \cap S'|$$

$$|S'| = |S' \setminus S| + |S \cap S'|$$

Now, substituting  $|S \setminus S'|$  and  $|S' \setminus S|$  in the first formula, the desired formula is obtained:

$$\begin{aligned} |S \cup S'| &= |S \setminus S'| + |S \setminus S'| + |S \cap S'| \\ &= |S| - |S \cap S'| + |S'| - |S' \setminus S| + |S \cap S'| + |S \cap S'| \\ &= |S| + |S'| - |S \cap S'| \end{aligned}$$

□

The following proposition proves that *Tanimoto distance* is equivalent to *Consistency* when they are computed for two sets.

**Proposition 1.**

*Tanimoto distance is equivalent to Consistency when they are computed for two sets.*

*Proof.*

Let be  $S$  and  $S'$  two sets.

Then, *Tanimoto distance* can be directly simplified as:

$$1 - \frac{|S| + |S'| - 2|S \cap S'|}{|S| + |S'| - |S \cap S'|} = \frac{|S \cap S'|}{|S| + |S'| - |S \cap S'|}$$

Regarding *Consistency* metric, its term  $freq(f)$  is equal to 2 if  $f \in S \cap S'$  and 1 otherwise. Thus, the following equality holds:

$$\frac{1}{|S \cup S'|} \left[ \sum_{f \in \{S, S'\}} freq(f) - 1 \right] = \frac{1}{|S \cup S'|} \left[ \sum_{f \in S \cap S'} 1 \right] = \frac{|S \cap S'|}{|S \cup S'|}$$

Then, for Lemma 1 *Tanimoto distance* and *Consistency* are equals

□

In the following propositions, we will pair-wisely prove that *Relative Hamming distance*, *Dice-Sorensen's index*, *Ochiai's index* and *Percentage of overlapping features* are equivalent.

**Proposition 2.**

*Dice-Sorensen's index is equivalent to Percentage of overlapping features when comparing two sets with the same cardinality.*

*Proof.*

Let be  $S$  and  $S'$  two sets of the same cardinality, i.e.,  $|S| = |S'|$ .

Then, *Dice-Sorensen's index* is directly equivalent to *Percentage of overlapping features*:

$$\frac{2|S \cap S'|}{|S| + |S'|} = \frac{2|S \cap S'|}{2|S|} = \frac{|S \cap S'|}{|S|}$$

□

**Proposition 3.**

*Ochiai's index is equivalent to Percentage of overlapping features when comparing two sets with the same cardinality.*

*Proof.*

Let be  $S$  and  $S'$  two sets of the same cardinality, i.e.,  $|S| = |S'|$ .

Then, *Ochiai's index* is directly equivalent to *Percentage of overlapping features*:

$$\frac{|S \cap S'|}{\sqrt{|S||S'|}} = \frac{|S \cap S'|}{\sqrt{|S|^2}} = \frac{|S \cap S'|}{|S|}$$

□

**Proposition 4.**

*Relative Hamming distance is equivalent to Dice-Sorensen's index when comparing two sets.*

*Proof.*

Using the fact that  $|S \cup S'| = |S \setminus S'| + |S' \setminus S| + |S \cap S'|$ , seen in the proof of Lemma 1, *Relative Hamming distance* can be rewritten as:

$$1 - \frac{|S \setminus S'| + |S' \setminus S|}{|S| + |S'|} = 1 - \frac{|S \cup S'| - |S \cap S'|}{|S| + |S'|}$$

This formula can be further simplified by using Lemma 1 :

$$1 - \frac{|S \cup S'| - |S \cap S'|}{|S| + |S'|} = \frac{|S| + |S'| - |S \cup S'| + |S \cap S'|}{|S| + |S'|} = \frac{2|S \cap S'|}{|S| + |S'|}$$

The simplification obtained is indeed the *Dice-Sorensen's index*. □

**Proposition 5.**

*Weighted consistency is equivalent to Dice-Sorensen's index when comparing two sets with the same cardinality.*

*Proof.*

As the denominator is not affected by the summatory, *Weighted consistency* can be rewritten as follows:

$$\sum_{f \in \{S, S'\}} \frac{freq(f)}{|S| + |S'|} \cdot freq(f) - 1 = \frac{1}{|S| + |S'|} \left[ \sum_{f \in \{S, S'\}} freq(f)^2 - freq(f) \right]$$

Then, the *Weighted consistency* can be simplified, using the fact that when two datasets are compared,  $freq(f)^2 - freq(f)$  is 2 if  $f \in S \cap S'$  and 0 otherwise.

$$\frac{1}{|S| + |S'|} \left[ \sum_{f \in \{S, S'\}} freq(f)^2 - freq(f) \right] = \frac{1}{|S| + |S'|} \left[ \sum_{f \in S \cap S'} 2 \right] = \frac{2|S \cap S'|}{|S| + |S'|}$$

The simplification obtained is indeed the *Dice-Sorensen's index*. □

## Appendix B

# Best Configurations

The complete list of configurations which achieved the best CV-results for the experiments *Exp-2*, *Exp-3*, *Exp-4*, *Exp-5*, *Exp-10*, *Exp-15* and *Exp-20* are shown in Tables [B.1](#), [B.2](#), [B.3](#), [B.4](#), [B.5](#), [B.6](#) and [B.7](#), respectively. All the list are ordered from less complex to more complex configurations.

Number estimators	Maximum features
5	1
8	1
11	1
14	1
17	1
20	1
23	1
26	1
29	1
32	1
35	1
38	1
41	1
44	1
47	1

TABLE B.1: RF configurations which obtain the best *CV-Accuracy* for selecting 2 features.

---

Number estimators	Maximum features
8	1
8	2
11	1
11	2
17	1
17	2
20	1
20	2
23	1
23	2
26	1
26	2
29	1
29	2
32	1
32	2
35	1
35	2
38	1
38	2
41	1
41	2
44	1
44	2
47	1
47	2

---

TABLE B.2: RF configurations which obtain the best *CV-Accuracy* for selecting 3 features.

Number estimators	Maximum features
8	1
8	2
8	3
11	1
11	2
11	3
14	1
14	2
14	3
17	1
17	2
17	3
23	1
23	2
23	3
26	1
26	2
26	3
29	1
29	2
29	3
32	1
32	2
32	3
35	1
35	2
35	3
38	1
38	2
38	3
41	1
41	2
41	3
44	1
44	2
44	3
47	1
47	2
47	3

TABLE B.3: RF configurations which obtain the best CV-Accuracy for selecting 4 features.

Number estimators	Maximum features	Number estimators	Maximum features
5	2	29	1
5	2	29	2
5	3	29	3
5	4	29	4
8	1	32	1
11	1	32	2
11	2	32	3
11	3	32	4
11	4	35	1
14	1	35	2
14	2	35	3
14	3	35	4
14	4	38	1
17	1	38	2
17	2	38	3
17	3	38	4
17	4	41	1
20	1	41	2
20	2	41	3
20	3	41	4
20	4	44	1
23	1	44	2
23	3	44	3
23	4	44	4
23	5	47	1
26	1	47	2
26	2	47	3
26	3	47	4
26	4		

TABLE B.4: RF configurations which obtain the best CV-Accuracy for selecting 5 features.

Number estimators	Maximum features	Number estimators	Maximum features
17	4	35	3
17	5	35	4
17	6	35	5
17	7	35	6
17	8	35	7
17	9	35	8
23	3	35	9
23	4	38	1
23	5	38	2
23	6	38	3
23	7	38	4
23	8	38	5
23	9	38	6
26	1	38	7
26	3	38	8
26	4	38	9
26	5	41	1
26	6	41	2
26	7	41	3
26	8	41	4
26	9	41	5
29	1	41	6
29	3	41	7
29	4	41	8
29	5	41	9
29	6	44	4
29	7	44	5
29	8	44	6
29	9	44	7
32	1	44	8
32	2	44	9
32	3	47	2
32	4	47	3
32	5	47	4
32	6	47	5
32	7	47	6
32	8	47	7
32	9	47	8
35	1	47	9
35	2		

TABLE B.5: RF configurations which obtain the best CV-Accuracy for selecting 10 features.

Number estimators	Maximum features	Number estimators	Maximum features	Number estimators	Maximum features
5	1	17	3	26	6
5	2	17	4	26	7
5	3	17	5	26	8
5	4	17	6	26	9
5	5	17	7	26	10
5	6	17	8	26	11
5	7	17	9	26	12
5	8	17	10	26	13
5	9	17	11	26	14
5	10	17	12	29	1
5	11	17	13	29	2
5	12	17	14	29	3
5	13	20	1	29	4
5	14	20	2	29	5
8	1	20	3	29	6
11	1	20	4	29	7
11	2	20	5	29	8
11	3	20	6	29	9
11	4	20	7	29	10
11	5	20	8	29	11
11	6	20	9	29	12
11	7	20	10	29	13
11	8	20	11	29	14
11	9	20	12	32	1
11	10	20	13	32	2
11	11	20	14	32	3
11	12	23	1	32	4
11	13	23	2	32	5
11	14	23	3	32	6
14	1	23	4	32	7
14	2	23	5	32	8
14	3	23	6	32	9
14	4	23	7	32	10
14	5	23	8	32	11
14	6	23	9	32	12
14	7	23	10	32	13
14	8	23	11	32	14
14	9	23	12	35	1
14	10	23	13	35	2
14	11	23	14	35	3
14	12	26	1	38	1
14	13	26	2	38	2
14	14	26	3	38	3
17	1	26	4		
17	2	26	5		

TABLE B.6: RF configurations which obtain the best CV-Accuracy for selecting 15 features.

Number estimators	Maximum features	Number estimators	Maximum features	Number estimators	Maximum features
5	2	17	7	26	13
5	3	17	8	26	14
5	4	17	9	26	15
5	5	17	10	26	16
5	6	17	11	26	17
5	7	17	12	26	18
5	8	17	13	26	19
5	9	17	14	29	3
5	10	17	15	29	4
5	11	17	16	29	5
5	12	17	17	29	6
5	13	17	18	29	7
5	14	17	19	29	8
5	15	20	3	29	9
5	16	20	4	29	10
5	17	20	5	29	11
5	18	20	6	29	12
5	19	20	7	29	13
11	3	20	8	29	14
11	4	20	9	29	15
11	5	20	10	29	16
11	6	20	11	29	17
11	7	20	12	29	18
11	8	20	13	29	19
11	9	20	14	32	4
11	10	20	15	32	5
11	11	20	16	32	6
11	12	20	17	32	7
11	13	20	18	32	8
11	14	20	19	32	9
11	15	23	3	32	10
11	16	23	4	32	11
11	17	23	5	32	12
11	18	23	6	32	13
11	19	23	7	32	14
14	3	23	8	32	15
14	4	23	9	32	16
14	5	23	10	32	17
14	6	23	11	32	18
14	7	23	12	32	19
14	8	23	13	35	4
14	9	23	14	35	5
14	10	23	15	35	6
14	11	23	16	35	7
14	12	23	17	35	8
14	13	23	18	35	9
14	14	23	19	35	10
14	15	26	4	35	11
14	16	26	5	35	12
14	17	26	6	35	13
14	18	26	7	35	14
14	19	26	8	35	15
17	3	26	9	35	16
17	4	26	10	35	17
17	5	26	11	35	18
17	6	26	12	35	19

TABLE B.7: RF configurations which obtain the best CV-Accuracy for selecting 20 features.



## Appendix C

# Selected features

Figures C.1, C.2, C.3, C.4, C.5 and C.6 show the selected features in experiments *Exp-2*, *Exp-3*, *Exp-4*, *Exp-5*, *Exp-10*, *Exp-15* and *Exp-20*, respectively. The selected features are highlighted in heatmaps representing the intensities of the proteins, y-axis, for each patient, x-axis.

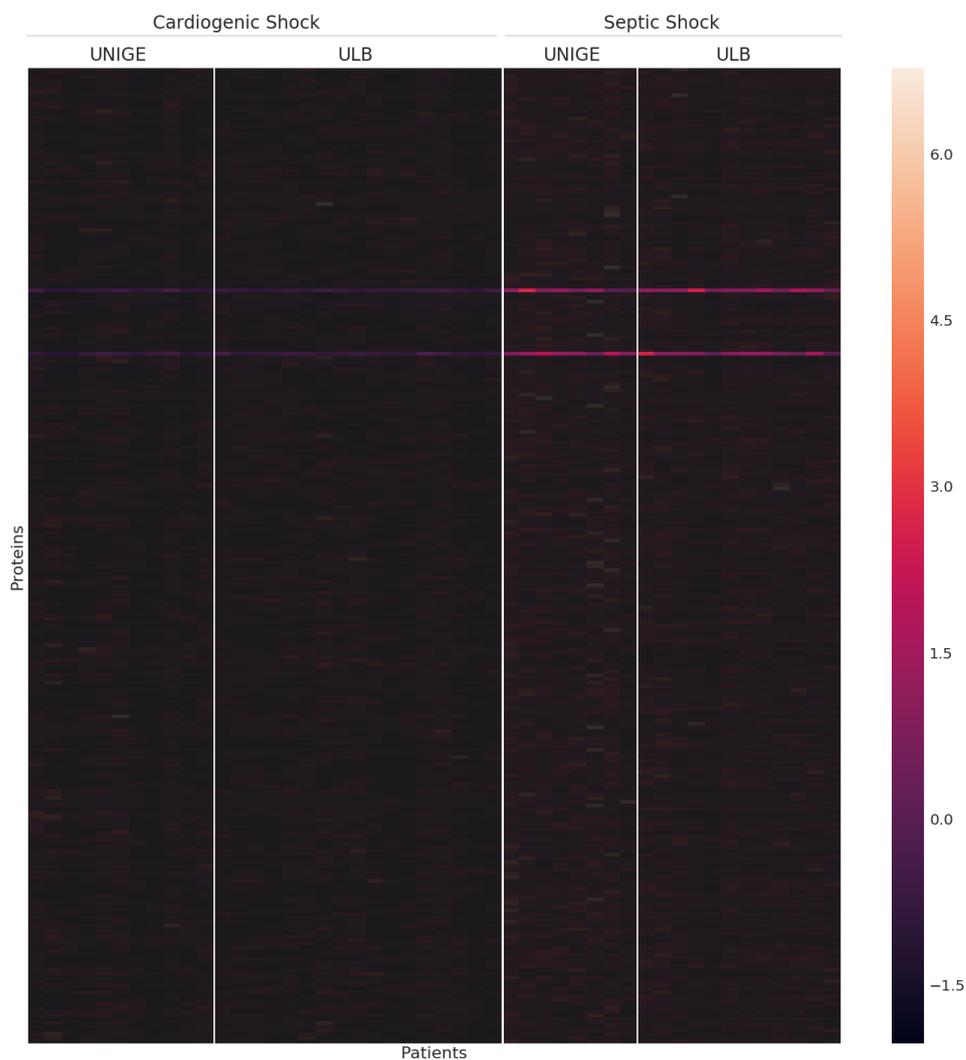


FIGURE C.1: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-2* are highlighted

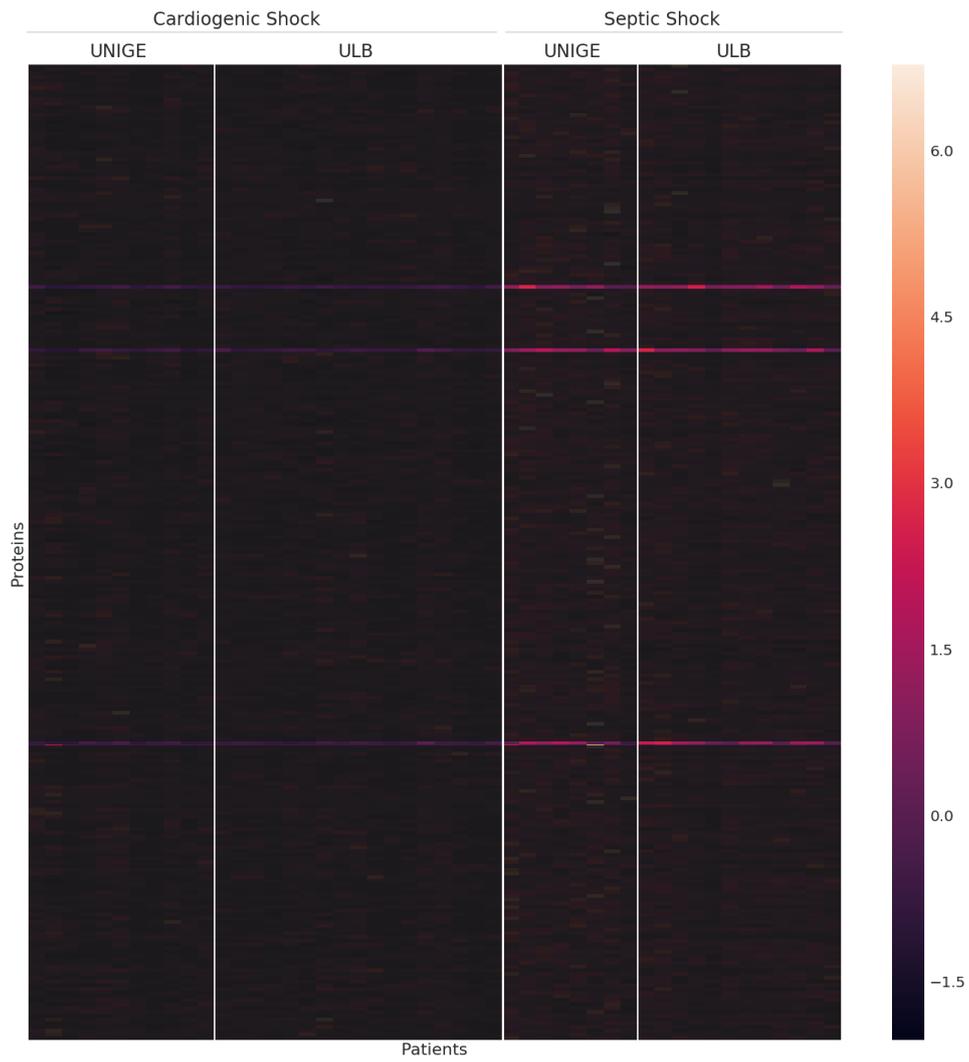


FIGURE C.2: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-3* are highlighted

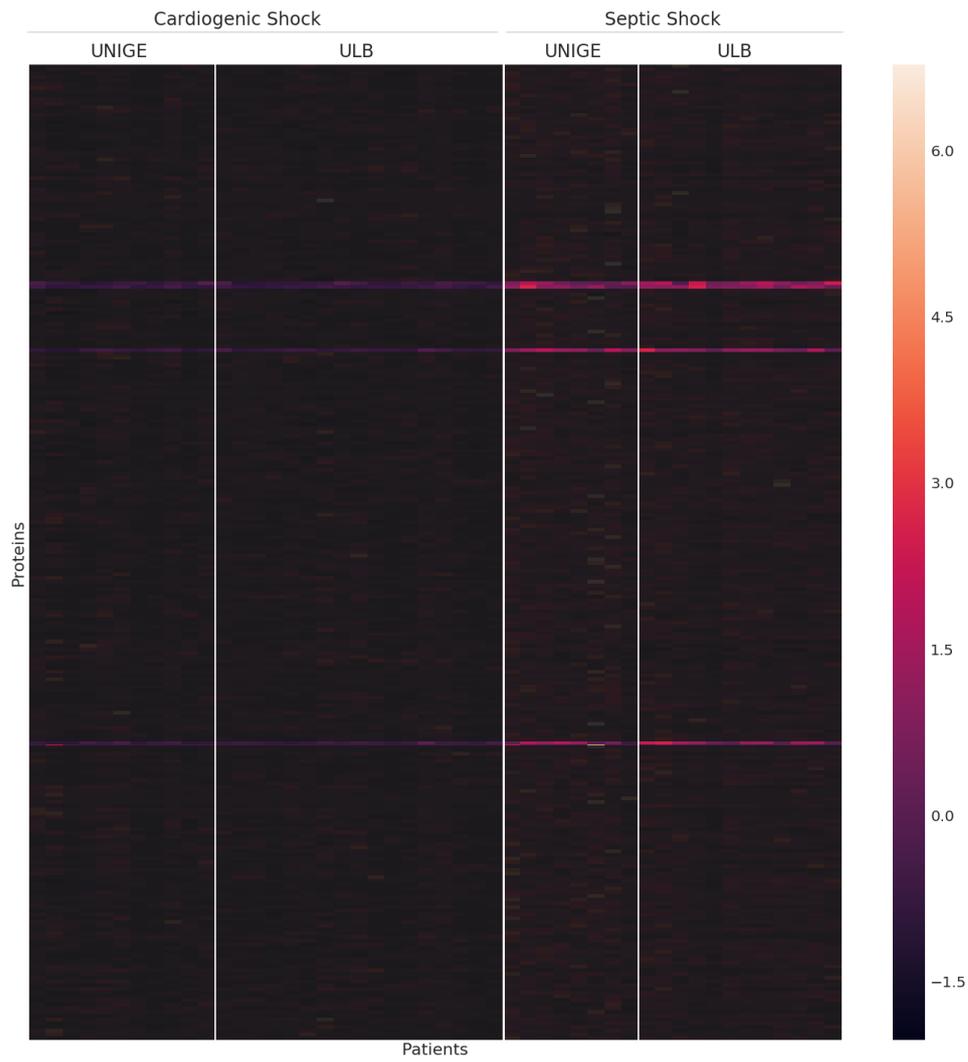


FIGURE C.3: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-4* are highlighted

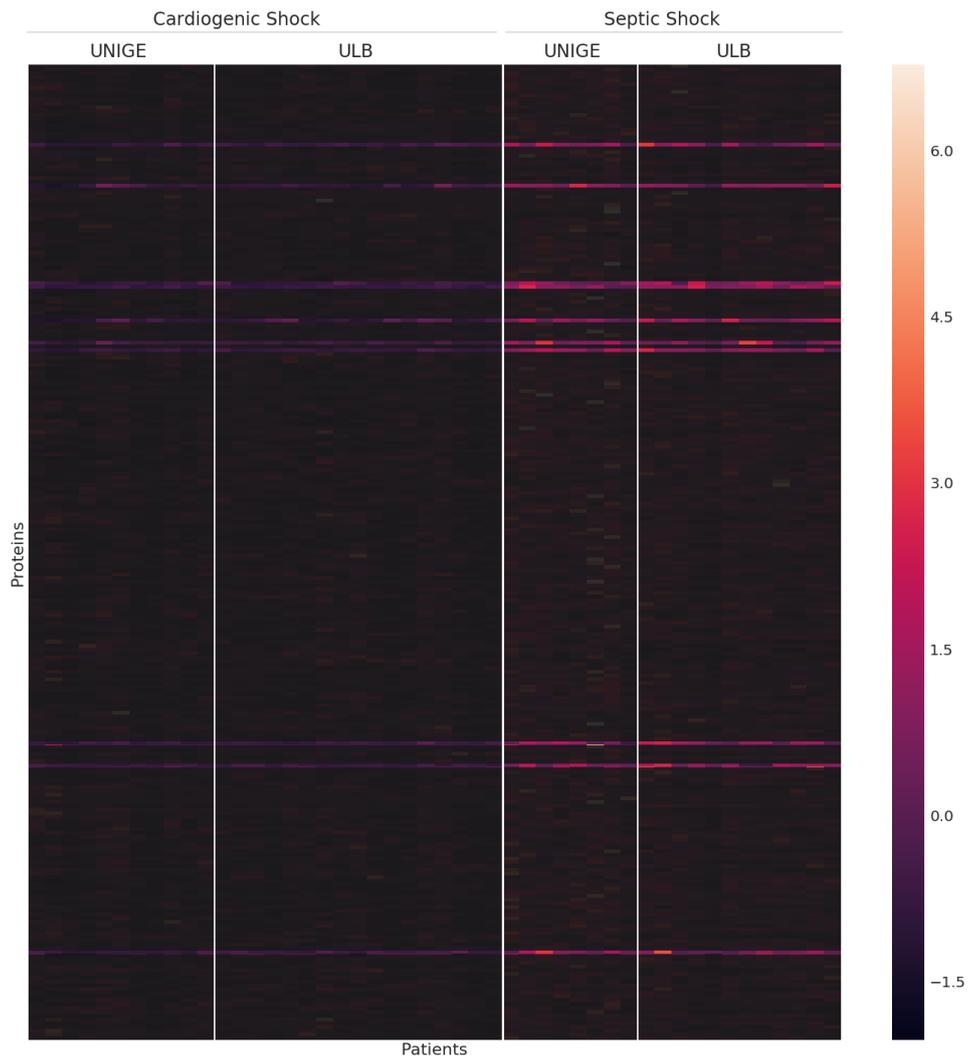


FIGURE C.4: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-10* are highlighted.

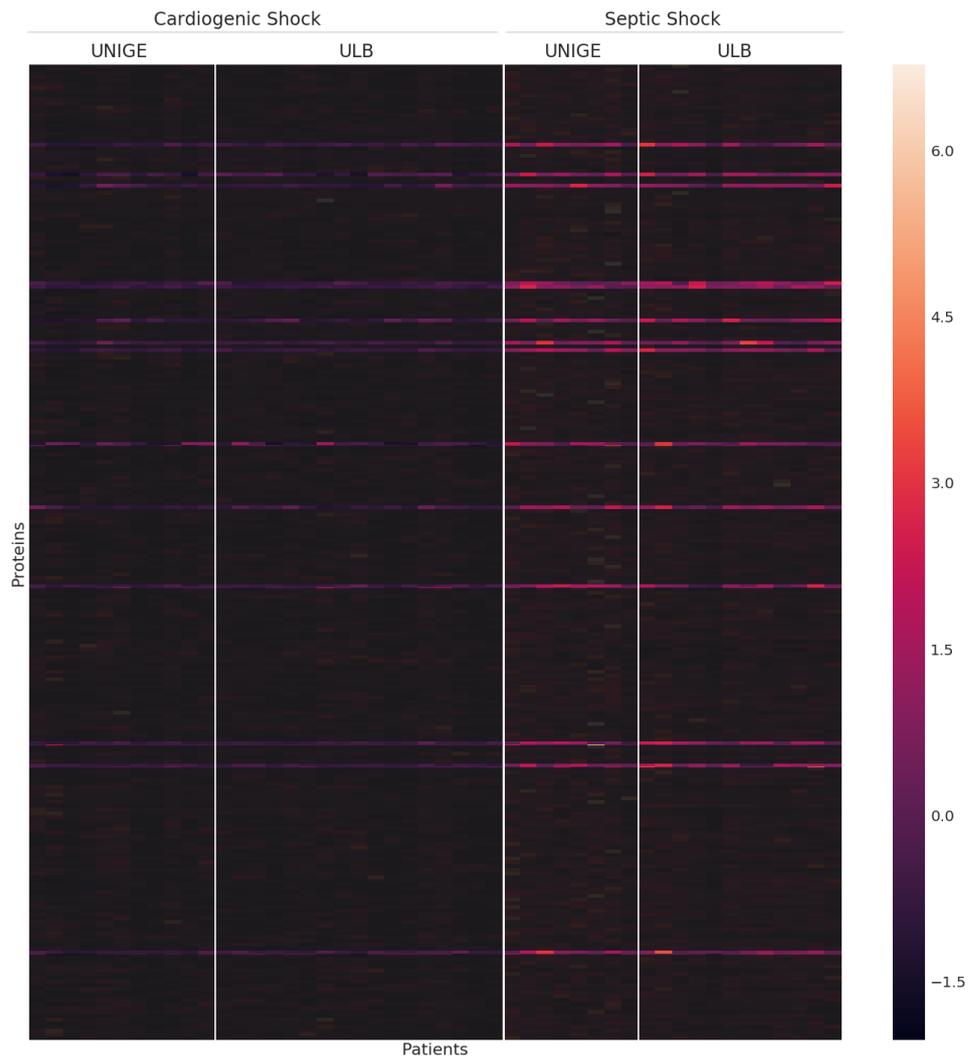


FIGURE C.5: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-15* are highlighted.

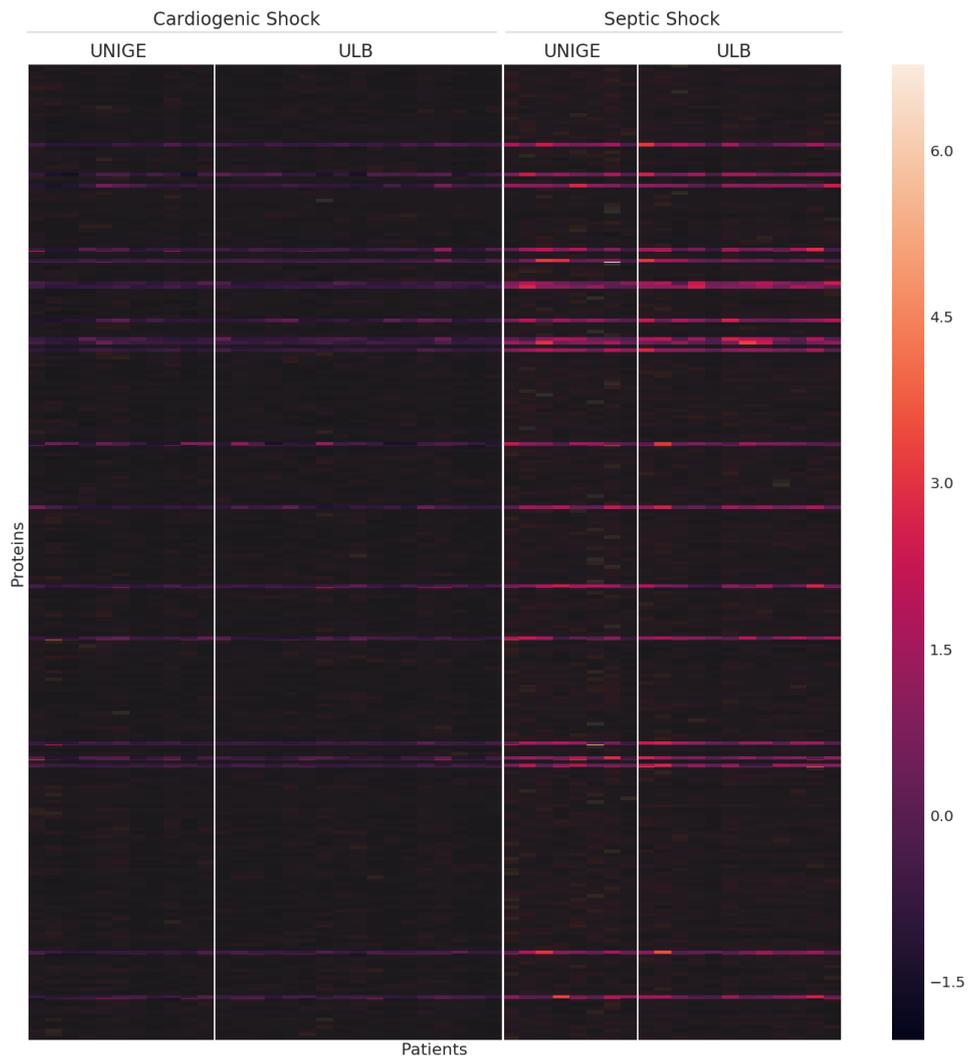


FIGURE C.6: Heatmap representing the intensities of the proteins, y-axis, for each patient, x-axis. The selected proteins in *Exp-20* are highlighted.

# Bibliography

- [1] AO Aasen et al. "Studies on components of the plasma kallikrein-kinin system in plasma samples from normal individuals and patients with septic shock". In: *Advances in shock research* 4 (1980), pp. 1–10.
- [2] Amy M. Ahasic et al. "Predictors of circulating insulin-like growth factor-1 and insulin-like growth factor-binding protein-3 in critical illness". In: *Critical Care Medicine* (2015). ISSN: 15300293. DOI: [10.1097/CCM.0000000000001314](https://doi.org/10.1097/CCM.0000000000001314).
- [3] Nadia Aissaoui et al. "Improved outcome of cardiogenic shock at the acute stage of myocardial infarction: A report from the USIK 1995, USIC 2000, and FAST-MI French Nationwide Registries". In: *European Heart Journal* 33.20 (2012), pp. 2535–2543. ISSN: 0195668X. DOI: [10.1093/eurheartj/ehs264](https://doi.org/10.1093/eurheartj/ehs264).
- [4] Corinne Alberti et al. "Epidemiology of sepsis and infection in ICU patients from an international multicentre cohort study". In: *Intensive Care Medicine* 28 (2002), 108–121. DOI: [10.1007/s00134-001-1143-z](https://doi.org/10.1007/s00134-001-1143-z).
- [5] Federico Aletti et al. "ShockOmics: Multiscale approach to the identification of molecular biomarkers in acute heart failure induced by shock". In: *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 24.1 (2016), p. 9. DOI: [10.1186/s13049-016-0197-4](https://doi.org/10.1186/s13049-016-0197-4).
- [6] Djillali Annane et al. "Current epidemiology of septic shock: The CUB-Réa network". In: *American Journal of Respiratory and Critical Care Medicine* 168.2 (2003), pp. 165–172. ISSN: 1073449X. DOI: [10.1164/rccm.2201087](https://doi.org/10.1164/rccm.2201087).
- [7] Gyan Bhanot et al. "A robust meta-classification strategy for cancer detection from MS data". In: *Proteomics* 6.2 (2006), pp. 592–604. ISSN: 16159853. DOI: [10.1002/pmic.200500192](https://doi.org/10.1002/pmic.200500192).
- [8] Jürgen Birnbaum et al. "Effects of coagulation factor XIII on intestinal functional capillary density, leukocyte adherence and mesenteric plasma extravasation in experimental endotoxemia". In: *Critical Care* 10.1 (2006), R29. DOI: [10.1186/cc3994](https://doi.org/10.1186/cc3994).
- [9] Alice Cambiaghi et al. "Characterization of a metabolomic profile associated with responsiveness to therapy in the acute phase of septic shock". In: *Scientific Reports* 7.1 (2017). ISSN: 20452322. DOI: [10.1038/s41598-017-09619-x](https://doi.org/10.1038/s41598-017-09619-x).
- [10] Marta Carrara et al. "Baroreflex Sensitivity and Blood Pressure Variability can Help in Understanding the Different Response to Therapy During Acute Phase of Septic Shock". In: *Shock* 50.1 (2018), pp. 78–86. DOI: [10.1097/shk.0000000000001046](https://doi.org/10.1097/shk.0000000000001046).
- [11] Maurizio Cecconi et al. "Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine". In: *Intensive Care Medicine* 40.12 (2014), pp. 1795–1815. ISSN: 14321238. DOI: [10.1007/s00134-014-3525-z](https://doi.org/10.1007/s00134-014-3525-z).
- [12] Jean Charchafliet et al. "The role of complement system in septic shock". In: *Clinical and Developmental Immunology* (2012). ISSN: 17402522. DOI: [10.1155/2012/407324](https://doi.org/10.1155/2012/407324).

- [13] Corinna Cortes, Vladimir Vapnik, and Lorenza Saitta. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297.
- [14] Joaquim F Pinto da Costa, Hugo Alonso, and Luis Roque. "A Weighted Principal Component Analysis and Its Application to Gene Expression Data". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8.1 (2011), pp. 246–252. ISSN: 1545-5963. DOI: [10.1109/TCBB.2009.61](https://doi.org/10.1109/TCBB.2009.61).
- [15] Thomas M Cover and Peter E Hart. "Nearest Neighbour Pattern Classification". In: *IEEE transactions on information theory* 24.1 (1967), pp. 335–342.
- [16] Mohammed Dakna et al. "Addressing the Challenge of Defining Valid Proteomic Biomarkers and Classifiers". In: *BMC Bioinformatics* 11.1 (2010), p. 594. DOI: [10.1186/1471-2105-11-594](https://doi.org/10.1186/1471-2105-11-594).
- [17] Chad A. Davis et al. "Reliable gene signatures for microarray classification: Assessment of stability and performance". In: *Bioinformatics* 22.19 (2006), pp. 2356–2363. ISSN: 13674803. DOI: [10.1093/bioinformatics/btl1400](https://doi.org/10.1093/bioinformatics/btl1400).
- [18] Xavier Delabranche, Julie Helms, and Ferhat Meziani. "Immunohaemostasis: a new view on haemostasis during sepsis". In: *Annals of Intensive Care* 7.1 (2017), p. 117. DOI: [10.1186/s13613-017-0339-5](https://doi.org/10.1186/s13613-017-0339-5).
- [19] Chris Ding and Hanchuan Peng. "Minimum Redundancy Feature Selection From Microarray Gene Expression Data". In: *Journal of Bioinformatics and Computational Biology* 3.2 (2005), pp. 185–205. DOI: [10.1142/S0219720005001004](https://doi.org/10.1142/S0219720005001004).
- [20] X Forceville. "Effects of high doses of selenium, as sodium selenite, in septic shock patients a placebo-controlled, randomized, double-blind, multi-center phase II study – Selenium and Sepsis". In: *Journal of Trace Elements in Medicine and Biology* 21 (2007), pp. 62–65. DOI: [10.1016/j.jtemb.2007.09.021](https://doi.org/10.1016/j.jtemb.2007.09.021). URL: [www.elsevier.de/jtemb](http://www.elsevier.de/jtemb).
- [21] X. Forceville et al. "Selenoprotein P, rather than glutathione peroxidase, as a potential marker of septic shock and related syndromes". In: *European Surgical Research* 43.4 (2009), pp. 338–347. DOI: [10.1159/000239763](https://doi.org/10.1159/000239763).
- [22] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics New York, NY, USA, 2001.
- [23] Robin Genuer et al. "Variable selection using Random Forests". In: *Pattern Recognition Letters* 31.14 (2012), pp. 2225–2236.
- [24] P. Geurts et al. "Proteomic mass spectra classification using decision tree based ensemble methods". In: *Bioinformatics* 21.14 (2005), pp. 3138–3145. DOI: [10.1093/bioinformatics/bti494](https://doi.org/10.1093/bioinformatics/bti494).
- [25] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees". In: *Machine Learning* 63.1 (2006), pp. 3–42. DOI: [10.1007/s10994-006-6226-1](https://doi.org/10.1007/s10994-006-6226-1).
- [26] Robert J. Goldberg et al. "Thirty-year trends (1975 to 2005) in the magnitude of, management of, and hospital death rates associated with cardiogenic shock in patients with acute myocardial infarction a population-based perspective". In: *Circulation* 156.2 (2008), pp. 227–233. DOI: [10.1161/CIRCULATIONAHA.108.814947](https://doi.org/10.1161/CIRCULATIONAHA.108.814947).

- [27] F. de Groof et al. "Acute Stress Response in Children with Meningococcal Sepsis: Important Differences in the Growth Hormone/Insulin-Like Growth Factor I Axis between Nonsurvivors and Survivors". In: *The Journal of Clinical Endocrinology & Metabolism* 87.7 (2002), pp. 3118–3124. DOI: [10.1210/jcem.87.7.8605](https://doi.org/10.1210/jcem.87.7.8605).
- [28] Isabelle Guyon and André Elisseeff. "An Introduction to Variable and Feature Selection". In: *Journal of Machine Learning Research (JMLR)* 3.3 (2003), pp. 1157–1182. ISSN: 00032670. DOI: [10.1016/j.aca.2011.07.027](https://doi.org/10.1016/j.aca.2011.07.027).
- [29] Isabelle Guyon et al. "Gene selection for cancer classification using Support Vector Machines". In: *Machine Learning* 43.1–3 (2002), pp. 389–422. DOI: [10.1108/03321640910919020](https://doi.org/10.1108/03321640910919020).
- [30] Henry Han. "A Novel Profile Biomarker Diagnosis for Mass Spectral proteomics". In: *Biocomputing* (2014), pp. 340–351. URL: [www.worldscientific.com](http://www.worldscientific.com).
- [31] Veli-Pekka Harjola et al. "Clinical picture and risk prediction of short-term mortality in cardiogenic shock". In: *European Journal of Heart Failure* 17.5 (2015), pp. 501–509. DOI: [10.1002/ejhf.260](https://doi.org/10.1002/ejhf.260).
- [32] Zengyou He and Weichuan Yu. "Stable feature selection for biomarker discovery". In: *Computational Biology and Chemistry* 34.4 (2010), pp. 215–225. DOI: [10.1016/j.compbiolchem.2010.07.002](https://doi.org/10.1016/j.compbiolchem.2010.07.002).
- [33] Richard P Horgan and Louise C Kenny. "'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics". In: *Gynaecologist* 13.3 (2011), pp. 189–195. DOI: [10.1576/toag.13.3.189.27672](https://doi.org/10.1576/toag.13.3.189.27672).
- [34] Raban Jeger et al. "Ten-year incidence and treatment of cardiogenic shock". In: *Annals of Internal Medicine* 149.9 (2008), pp. 618–626. DOI: [10.5167/uzh-5876](https://doi.org/10.5167/uzh-5876).
- [35] K. Jong et al. "Feature Selection in Proteomic Pattern Data with Support Vector Machines". In: *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on*. 2004, pp. 41–48. ISBN: 9789491027246. DOI: [10.1063/1.3033202](https://doi.org/10.1063/1.3033202).
- [36] A. Kalousis, J. Prados, and M. Hilario. "Stability of Feature Selection Algorithms". In: *Fifth IEEE International Conference on Data Mining (ICDM'05)* (2005), pp. 218–225. ISSN: 15504786. DOI: [10.1109/ICDM.2005.135](https://doi.org/10.1109/ICDM.2005.135). URL: <http://ieeexplore.ieee.org/document/1565682/>.
- [37] Alexandros Kalousis, Julien Prados, and Melanie Hilario. "Stability of feature selection algorithms: A study on high-dimensional spaces". In: *Knowledge and Information Systems* 12.1 (2007), pp. 95–116. DOI: [10.1007/s10115-006-0040-8](https://doi.org/10.1007/s10115-006-0040-8).
- [38] P. Kellner et al. "APACHE II, APACHE III, Elebute-Stoner, SOFA und SAPS II". In: *Medizinische Klinik - Intensivmedizin und Notfallmedizin* (2013). DOI: [10.1007/s00063-013-0234-2](https://doi.org/10.1007/s00063-013-0234-2).
- [39] Minseung Kim and Ilias Tagkopoulos. "Data integration and predictive modeling methods for multi-omics datasets". In: *Molecular Omics* 14.1 (2018), pp. 8–25. DOI: [10.1039/C7M000051K](https://doi.org/10.1039/C7M000051K).
- [40] William A Knaus et al. "APACHE-acute physiology and chronic health evaluation: a physiologically based classification system." In: *Critical care medicine* 9.8 (1981), pp. 591–597.
- [41] William A Knaus et al. "APACHE II: a severity of disease classification system". In: *Critical care medicine* 13.10 (1985), pp. 818–829.

- [42] S B Kotsiantis. "Supervised Machine Learning: A Review of Classification Techniques". In: *Informatica* 31 (2007), pp. 249–268.
- [43] Pentti Kuusela et al. "Changes in plasma protein levels as an early indication of a bloodstream infection". In: *PLoS ONE* 12.2 (2017). DOI: [10.1371/journal.pone.0172987](https://doi.org/10.1371/journal.pone.0172987).
- [44] Raymond J. Langley et al. "An integrated clinico-metabolomic model improves prediction of death in sepsis". In: *Science Translational Medicine* 5.195 (2013), 195ra95–195ra95. DOI: [10.1126/scitranslmed.3005893](https://doi.org/10.1126/scitranslmed.3005893).
- [45] Aleksandra Leligdowicz et al. "Association between source of infection and hospital mortality in patients who have septic shock". In: *American Journal of Respiratory and Critical Care Medicine* 189.10 (2014), pp. 1204–1213. DOI: [10.1164/rccm.201310-18750C](https://doi.org/10.1164/rccm.201310-18750C).
- [46] Leping Li et al. "Application of the GA/KNN method to SELDI proteomics data". In: *Bioinformatics* 20.10 (2004), pp. 1638–1640. DOI: [10.1093/bioinformatics/bth098](https://doi.org/10.1093/bioinformatics/bth098).
- [47] David J Marquardt et al. "Failure to recover somatotrophic axis function is associated with mortality from pediatric sepsis-induced multiple organ dysfunction syndrome". In: *Pediatric Critical Care Medicine* 11.1 (2010), pp. 18–25. DOI: [10.1097/PCC.0b013e3181b06046](https://doi.org/10.1097/PCC.0b013e3181b06046).
- [48] F Martinez-Brotons et al. "Plasma kallikrein-kinin system in patients with uncomplicated sepsis and septic shock-comparison with cardiogenic shock". In: *Thrombosis and haemostasis* 57.02 (1987), pp. 709–713. DOI: [10.1055/s-0038-1645960](https://doi.org/10.1055/s-0038-1645960).
- [49] Hajime Nakae et al. "Chronological changes in the complement system in sepsis". In: *Surgery today* 26.4 (1996), pp. 225–229.
- [50] Thanh Nguyen et al. "Mass spectrometry cancer data classification using wavelets and genetic algorithm". In: *FEBS Letters* 589 (2015), pp. 3879–3886. DOI: [10.1016/j.febslet.2015.11.019](https://doi.org/10.1016/j.febslet.2015.11.019).
- [51] Markus Ojala and Gemma C Garriga. "Permutation Tests for Studying Classifier Performance". In: *Journal of Machine Learning Research* 11.Jun (2010), pp. 1833–1863.
- [52] Chrysoula Papastathi et al. "Insulin-like Growth Factor I and its binding protein 3 in sepsis". In: *Growth Hormone and IGF Research* 23.4 (2013), pp. 98–104. DOI: [10.1016/j.ghir.2013.03.005](https://doi.org/10.1016/j.ghir.2013.03.005).
- [53] Julien Prados et al. "Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents". In: *Proteomics* 4.8 (2004), pp. 2320–2332. ISSN: 16159853. DOI: [10.1002/pmic.200400857](https://doi.org/10.1002/pmic.200400857).
- [54] Etienne Puymirat et al. "Cardiogenic shock in intensive care units: evolution of prevalence, patient profile, management and outcomes, 1997–2012". In: *European Journal of Heart Failure* 19.2 (2017), pp. 192–200. DOI: [10.1002/ejhf.646](https://doi.org/10.1002/ejhf.646).
- [55] Jean Pierre Quenot et al. "The epidemiology of septic shock in French intensive care units: The prospective multicenter cohort EPISS study". In: *Critical Care* 17.2 (2013), R65. DOI: [10.1186/cc12598](https://doi.org/10.1186/cc12598).
- [56] Kalpana Raja et al. *A Review of Recent Advancement in Integrating Omics Data with Literature Mining towards Biomedical Discoveries*. 2017. DOI: [10.1155/2017/6213474](https://doi.org/10.1155/2017/6213474).

- [57] Habtom W. Ressonm et al. "Analysis of mass spectral serum profiles for biomarker selection". In: *Bioinformatics* 21.21 (2005), pp. 4039–4045. ISSN: 13674803. DOI: [10.1093/bioinformatics/bti670](https://doi.org/10.1093/bioinformatics/bti670).
- [58] M Robnik-Sikonja and I Kononenko. "Theoretical and empirical analysis of Relief and ReliefF". In: *Machine Learning* 53.1-2 (2003), pp. 23–69.
- [59] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics". In: *Bioinformatics* 23.19 (2007), pp. 2507–2517. DOI: [10.1093/bioinformatics/btm344](https://doi.org/10.1093/bioinformatics/btm344).
- [60] Yasser Sakr et al. "Does dopamine administration in shock influence outcome? Results of the Sepsis Occurrence in Acutely Ill Patients (SOAP) Study". In: *Critical Care Medicine* 34.3 (2006), pp. 589–597. DOI: [10.1097/01.CCM.0000201896.45809.E3](https://doi.org/10.1097/01.CCM.0000201896.45809.E3).
- [61] Robert E. Schapire. "A brief introduction to boosting". In: *IJCAI International Joint Conference on Artificial Intelligence*. 1999, pp. 1401–1406. ISBN: 3540440119. DOI: [citeulike-article-id:765005](https://doi.org/citeulike-article-id:765005).
- [62] SepNet Critical Care Trials Group. "Incidence of severe sepsis and septic shock in German intensive care units: the prospective, multicentre INSEP study". In: *Intensive Care Medicine* 42.12 (2016), pp. 1980–1989. ISSN: 14321238. DOI: [10.1007/s00134-016-4504-3](https://doi.org/10.1007/s00134-016-4504-3).
- [63] Mervyn Singer et al. "The third international consensus definitions for sepsis and septic shock (sepsis-3)". In: *JAMA - Journal of the American Medical Association* 315.8 (2016), pp. 801–810. ISSN: 15383598. DOI: [10.1001/jama.2016.0287](https://doi.org/10.1001/jama.2016.0287).
- [64] N Smith-Erichsen, AO Aasen, and E Amundsen. "The functional inhibition of plasma kallikrein. A critical factor in septic shock". In: *Advances in experimental medicine and biology* 156 (1983), pp. 1049–1054.
- [65] Anna Louise Swan et al. "Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology". In: *OMICS: A Journal of Integrative Biology* 17.12 (2013), pp. 595–610. DOI: [10.1089/omi.2013.0017](https://doi.org/10.1089/omi.2013.0017).
- [66] Filip M Szymanski and Krzysztof J Filipiak. "Cardiogenic shock — diagnostic and therapeutic options in the light of new scientific data". In: *REVIEWS Anaesthesiology Intensive Therapy* 46.4 (2014), pp. 301–306. DOI: [10.5603/AIT.2014.049](https://doi.org/10.5603/AIT.2014.049).
- [67] Hiroshi Tanaka et al. "Role of granulocyte elastase in tissue injury in patients with septic shock complicated by multiple-organ failure". In: *Annals of Surgery* 213.1 (1991), p. 81. DOI: [10.1097/0000658-199101000-00014](https://doi.org/10.1097/0000658-199101000-00014).
- [68] Tin Kam Ho. "Random decision forests". In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. 1995, pp. 278–282. ISBN: 0-8186-7128-9. DOI: [10.1109/ICDAR.1995.598994](https://doi.org/10.1109/ICDAR.1995.598994).
- [69] Erik I. Tucker et al. "Survival Advantage of Coagulation Factor XI-Deficient Mice during Peritoneal Sepsis". In: *The Journal of Infectious Diseases* 198.2 (2008), pp. 271–274. DOI: [10.1086/589514](https://doi.org/10.1086/589514).
- [70] Alfredo Vellido et al. "Machine Learning for Critical Care: An Overview and a Sepsis Case Study". In: *International Conference on Bioinformatics and Biomedical Engineering*. 2017, pp. 15–30. ISBN: 978-3-319-56148-6. DOI: [10.1007/978-3-319-56148-6\\_{\\_}2](https://doi.org/10.1007/978-3-319-56148-6_{_}2).

- [71] Jean-Louis Vincent and Daniel De Backer. "Circulatory Shock". In: *New England Journal of Medicine* 369.18 (2013), pp. 1726–1734. DOI: [10.1056/NEJMra1208943](https://doi.org/10.1056/NEJMra1208943).
- [72] Jean-Louis Vincent et al. *The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure*. Tech. rep. 1996, pp. 707–710.
- [73] Michael Wagner et al. "Computational protein biomarker prediction: A case study for prostate cancer". In: *BMC Bioinformatics* 5.1 (2004), p. 26. DOI: [10.1186/1471-2105-5-26](https://doi.org/10.1186/1471-2105-5-26).
- [74] Valerie C. Wasinger et al. "Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*". In: *ELECTROPHORESIS* 16.1 (1995), pp. 1090–1094. ISSN: 15222683. DOI: [10.1002/elps.11501601185](https://doi.org/10.1002/elps.11501601185).
- [75] J. Witte et al. "Disturbances of selected plasma proteins in hyperdynamic septic shock". In: *Intensive Care Medicine* 8.5 (1982), pp. 215–222. DOI: [10.1007/BF01694524](https://doi.org/10.1007/BF01694524).
- [76] Baolin Wu et al. "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data". In: *Bioinformatics* 19.13 (2003), pp. 1636–1643. ISSN: 13674803. DOI: [10.1093/bioinformatics/btg210](https://doi.org/10.1093/bioinformatics/btg210).
- [77] J. S. Yu et al. "Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data". In: *Bioinformatics* 21.10 (2005), pp. 2200–2209. ISSN: 13674803. DOI: [10.1093/bioinformatics/bti370](https://doi.org/10.1093/bioinformatics/bti370).
- [78] Sacha Zeerleder et al. "Factor XIII in severe sepsis and septic shock". In: *Thrombosis Research* 119.3 (2007), pp. 311–318. DOI: [10.1016/j.thromres.2006.02.003](https://doi.org/10.1016/j.thromres.2006.02.003).
- [79] Xuegong Zhang et al. "Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data". In: *BMC Bioinformatics* 7.1 (2006), p. 197. DOI: [10.1186/1471-2105-7-197](https://doi.org/10.1186/1471-2105-7-197).
- [80] Yongzhong Zhao et al. "Selenoprotein P neutralizes lipopolysaccharide and participates in hepatic cell endoplasmic reticulum stress response". In: *FEBS Letters* 590.24 (2016), pp. 4519–4530. DOI: [10.1002/1873-3468.12494](https://doi.org/10.1002/1873-3468.12494).