NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

FRANCISCA DE ALMEIDA NUNES TUDELA MANOEL

Bachelor in Biomedical Engineering

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE ALGORITHMS FOR EARLY DIAGNOSIS OF SEPSIS

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
November, 2021

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

DEVELOPMENT OF ARTIFICIAL INTELLIGENCE ALGORITHMS FOR EARLY DIAGNOSIS OF SEPSIS

FRANCISCA DE ALMEIDA NUNES TUDELA MANOEL

Bachelor in Biomedical Engineering

Adviser: Prof. Dr. Carla Maria Quintão Pereira
Auxiliar Professor, NOVA University Lisbon

Co-adviser: Prof. Dr. Ricardo Nuno Pereira Verga e Afonso Vigário
Associate Professor, NOVA University Lisbon

Examination Committee

Chair: Prof. Dr. Célia Maria Reis Henriques
Associate Professor, FCT-NOVA

Rapporteur: Dr. André Valério Raposo Carreiro
Senior Scientist, Fraunhofer Portugal

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
November, 2021

Development of Artificial Intelligence Algorithms for Early Diagnosis of Sepsis

To my family.

Acknowledgements

This dissertation represents one of the biggest accomplishments in my academic life, which would not have been possible without the support of the people in my life.

Firstly, I would like to thank my advisers, Professor Carla Quintão and Professor Ricardo Vigário, for giving me the opportunity to work on this project, as well as for all the support. Your feedback was crucial for accomplishing this work's goals. I also want to express my gratitude to Professor Cláudia Quaresma, who has always been there with us, providing input and would always offer to help with whatever issue. Thank you all for your guidance, enthusiasm, and daily good humor, all of which have always kept me motivated. I would like to thank Dr. Paulo Barreto and everyone at *Centro Hospitalar Universitário Lisboa Central* involved in this partnership project, for helping me and making it possible to happen throughout these challenging pandemic times.

I would also like to give the biggest word of gratitude to my family, to whom I owe everything. To my father, Sebastião, who has always been my greatest encourager and the one I look up to. I am your biggest fan and I hope to make you proud. My mother, Sara, whose love and support have always been constant throughout my life, you inspire me to always be better. My stepmother, Verónica, who has always been there for me, ready to help with anything I ever needed. To my special one, Pedro, thank you for always believing in me. You are my best friend, and this journey would not have been the same without you. Thank you all for your patience and unconditional support during these past 5 years. This work is for you.

To all my closest friends, Clarisse, Maria, Catarina Martins, Núria, Cátia, and Catarina Passarinho, a huge thank you for all the years we have spent together, for all the laughs, the memories that I will forever cherish, for always caring for me and believing in me.

Abstract

Sepsis is a prevalent syndrome that manifests itself through an uncontrolled response from the body to an infection, that may lead to organ dysfunction. Its diagnosis is urgent since early treatment can reduce the patients' chances of having long-term consequences. Yet, there are many obstacles to achieving this early detection. Some stem from the syndrome's pathogenesis, which lacks a characteristic biomarker. The available clinical detection tools are either too complex or lack sensitivity, in both cases delaying the diagnosis. Another obstacle relates to modern technology, that when paired with the many clinical parameters that are monitored to detect sepsis, result in extremely heterogenous and complex medical records, which constitute a big obstacle for the responsible clinicians, that are forced to analyse them to diagnose the syndrome.

To help achieve this early diagnosis, as well as understand which parameters are most relevant to obtain it, an approach based on the use of Artificial Intelligence algorithms is proposed in this work, with the model being implemented in the alert system of a sepsis monitoring platform.

This platform uses a Random Forest algorithm, based on supervised machine learning classification, that is capable of detecting the syndrome in two different scenarios. The earliest detection can happen if there are only five vital sign parameters available for measurement, namely heart rate, systolic and diastolic blood pressures, blood oxygen saturation level, and body temperature, in which case, the model has a score of 83% precision and 62% sensitivity. If besides the mentioned variables, laboratory analysis measurements of bilirubin, creatinine, hemoglobin, leukocytes, platelet count, and C-reactive protein levels are available, the platform's sensitivity increases to 77%. With this, it has also been found that the blood oxygen saturation level is one of the most important variables to take into account for the task, in both cases. Once the platform is tested in real clinical situations, together with an increase in the available clinical data, it is believed that the platform's performance will be even better.

Keywords: Sepsis, Early Diagnosis, Artificial Intelligence, Machine Learning, Alert System, Monitoring.

Resumo

A sépsis é uma síndrome com elevada incidência a nível global, que se manifesta através de uma resposta desregulada por parte do organismo a uma infeção, podendo resultar em disfunções orgânicas generalizadas. O diagnóstico da mesma é urgente, uma vez que um tratamento precoce pode reduzir as hipóteses de consequências a longo prazo para os doentes. Apesar desta necessidade, existem vários obstáculos. Alguns deles advêm da patogenia da síndrome, que carece de um biomarcador específico. As ferramentas de deteção clínica são demasiado complexas, ou pouco sensíveis, em ambos os casos atrasando o diagnóstico. Outro obstáculo relaciona-se com os avanços da tecnologia, que, com os vários parâmetros clínicos que são monitorizados, resulta em registos médicos heterogéneos e complexos, o que constitui um grande obstáculo para os profissionais de saúde, que se vêm forçados a analisá-los para diagnosticar a síndrome.

Para atingir este diagnóstico precoce, bem como compreender quais os parâmetros mais relevantes para o alcançar, é proposta neste trabalho uma abordagem baseada num algoritmo de Inteligência Artificial, sendo o modelo implementado no sistema de alerta de uma plataforma de monitorização de sépsis.

Esta plataforma utiliza um classificador *Random Forest* baseado em aprendizagem automática supervisionada, capaz de diagnosticar a síndrome de duas formas. Uma deteção mais precoce pode ocorrer através de cinco parâmetros vitais, nomeadamente frequência cardíaca, pressão arterial sistólica e diastólica, nível de saturação de oxigénio no sangue e temperatura corporal, caso em que o modelo atinge valores de 83% de precisão e 62% de sensibilidade. Se, para além das variáveis mencionadas, estiverem disponíveis análises laboratoriais de bilirrubina, creatinina, hemoglobina, leucócitos, contagem de plaquetas e níveis de proteína C-reativa, a sensibilidade da plataforma sobre para 77%. Concluiu-se que o nível de saturação de oxigénio no sangue é uma das variáveis mais importantes a ter em conta para o diagnóstico, em ambos os casos. A partir do momento que a plataforma venha a ser utilizada em situações clínicas reais, com o consequente aumento dos dados disponíveis, crê-se que o desempenho venha a ser ainda melhor.

Palavras-chave: Sépsis, Diagnóstico Precoce, Inteligência Artificial, Aprendizagem Automática, Sistema de Alerta, Monitorização.

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

.csv	Comma Separated Values
Acc	Accuracy
Adaboost	Adaptive Boosting
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUROC	Area Under the ROC curve
bilir	Bilirubin
CHULC	Centro Hospitalar Universitário de Lisboa Central
CNN	Convolutional Neural Network
creat	Creatinine
CRP	C-reactive protein
DBP	Diastolic Blood Pressure
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
ED	Emergency Department
EHR	Electronic Health Records
FFNN	Feed-Forward Neural Network
FiO_2	Fraction of inspired oxygen
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GBDT	Gradient Boosting Decision Tree
GCS	Glasgow Coma Scale

hemo	Hemoglobin
HR	Heart Rate
ICU	Intensive Care Unit
KNN	K-Nearest Neighbours
leuko	Leukocytes
LR	Logistic Regression
LSTM	Long Short-Term Memory based Neural Network
MBP	Mean Blood Pressure
MDA	Mean Decrease in Accuracy
MDI	Mean Decrease in Impurity
MIMIC-III	Medical Information Mart of Intensive Care III
ML	Machine Learning
NEWS	National Early Warning Score
NHS	National Health Service
NN	Neural Network
PaO_2	Partial pressure of oxygen
PCT	Procalcitonin
plat	Platelet count
qSOFA	quick-SOFA
RF	Random Forest
RNN	Recurrent Neural Network
ROC curve	Receiver Operating Characteristics curve
RR	Respiratory Rate
SaO_2	Blood oxygen saturation
SBFS	Sequential Backward Floating Selection
SBP	Systolic Blood Pressure
SBS	Sequential Backward Selection
SD	Standard Deviation
SFFS	Sequential Forward Floating Selection
SFS	Sequential Forward Selection

SIRS Systemic Inflammatory Response Syndrome
SOFA Sequential Organ Failure Assessment

temp Body temperature
TN True Negative
TP True Positive
TPR True Positive Rate

var Variance

XGBoost eXtreme Gradient Boosting

Chapter 1

Introduction

1.1 Context and Motivation

Sepsis is defined as a potentially life-threatening organ dysfunction caused by an unregulated response to an infection [1]. This definition has been evolving over the past decades due to its complexity [1]–[3], with this syndrome being the leading cause of mortality from infection [1]. It is estimated that in 2017, it had a world incidence of approximately 49 million cases, resulting in 11 million deaths. Thus, it accounted for about 20% of global mortality [4].

Several studies have shown that the earlier the patient is diagnosed and begins treatment, the better are the chances of recovery [1], [5]–[10], which reflects the urgency in its recognition. However, the complex pathology of the syndrome makes it difficult for clinicians to, not only get a timely diagnosis, but also deliver effective treatment [6], [11], [12]. The presence of inflammation can be caused by several pathological processes, and the unregulated response of the patient varies greatly with each individual [6].

There are several criteria for its clinical detection, each being dependent on certain clinical parameters. The current criterion of choice is the Sequential Organ Failure Assessment (SOFA) score [1] which determines the degree of organic dysfunction of the patient. This screening tool (as well as others) has the disadvantage of requiring information from laboratory results, which may delay the diagnosis. Simpler criteria have been created to easily assess the risk of sepsis, namely the criterion quick-SOFA (qSOFA), but the need for tools that allow for a rapid diagnosis remains, due to their low sensitivity in the earliest stages of the syndrome [13]–[15].

In addition, continuous patient monitoring generates Electronic Health Records (EHR) with large quantities of data. Although they allow for great monitoring possibilities, they can also be seen as an obstacle, as healthcare professionals must work with and

comprehend large amounts of data before making decisions [16], [17]. EHRs consist of many types of clinical parameters that may or may not be collected with different protocols and at different times. This makes them very irregular and difficult to extract patterns from, especially regarding time dependency [18].

Artificial Intelligence (AI) has proven to be increasingly useful, not only in overcoming these obstacles [17] but also in assisting clinicians to deliver timely treatment [19]. In the realm of medicine, it has been used for drug discovery, personalized diagnostics and therapies, molecular biology, bioinformatics, and medical imaging [20]. Machine Learning (ML) has shown great results for many clinical purposes, from the analysis of these EHRs to the development of prediction models, not only for disease progression, but also mortality risk assessment [20], [21], and has shown better performance than the most widely used detection criteria, such as SOFA [16], [22]–[24].

1.2 Objectives

This dissertation project focuses on the development of AI algorithms, that aim to achieve an early diagnostic and classification of sepsis patients. Contrasting with patients that are already in ICU settings, this retrospective study has the goal of detecting the syndrome through classification of patients in early care settings, with the lowest amount of clinical parameters as possible. This way, two different types of EHRs will be used: the first holds clinical information from patients within the ICU department, and will be used to verify the models' ability to distinguish patients that most likely have a more severe stage of sepsis, while the second contains information from both the intermediate care and infirmary settings of Centro Hospitalar Universitário de Lisboa Central (CHULC) (with whom this dissertation project was developed in partnership) to then compare and conclude whether it is possible to distinguish the syndrome in its early stages, with the available clinical data. The best performing classifiers will then be integrated in a sepsis detecting platform, developed in a previous work by Miguel [25], to help its alert system in identifying patients at risk and alert responsible clinicians. As a secondary goal, it is also intended to determine which indicators are most relevant for the detection of the syndrome. Therefore, the scope of this work includes:

- Construction and evaluation of ML models that classify and distinguish the septic population from the non-septic, within the early care setting's dataset;
- Analyse the best performing models, in order to detect the most important features for the classification task;
- Selecting the best performing model to implement in the platform [25];

1.3 Thesis Overview

This dissertation has six Chapters and two Appendices. The present Chapter introduces the context in which the problem of sepsis detection arises, as well as the goals that allow to propose a solution-approach for this problem. Chapter 2 presents the most important theoretical concepts, that provide a better understanding of the described framework. Then, in Chapter 3, a review of state-of-the-art approaches that have recently been proposed is presented, as well as some background studies that support the claim of the existing problem, regarding the clinical detection of sepsis. Chapter 4 describes the datasets that were used during the development of the algorithms, as well as the preprocessing methods that were deployed on this data. Then, the results regarding the performance of the trained models, the feature importance analysis, and their implementation in the platform, are all demonstrated and discussed in Chapter 5. This dissertation ends with Chapter 6, in which a summary of the achieved results, their limitations, and suggestions for future work, are given. Appendix A and Appendix B provide additional information regarding the results obtained in the data preprocessing approach (Chapter 4) and the model's performance evaluation and feature analysis (Chapter 5), respectively.

Chapter 2

Theoretical Concepts

In this chapter, the most relevant theoretical concepts related to this dissertation are described in detail. It begins by defining the evolution of the clinical definition of sepsis, to then describe it in light of the latest accepted definition, as well as its detection criteria/screening tools. Then, the fields of Artificial Intelligence (AI) and Machine Learning (ML) are delved into, focusing on the fundamentals of ML, as well as the models that have been recently used in this context. The area of Deep Learning (DL) is briefly mentioned, due to its recent results and potential for future use. Finally, both the concept of model overfitting and important metrics of evaluation of AI algorithms are explained in detail.

2.1 Sepsis

In 1991, at the American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference, the inflammatory reaction of the host to an infection was defined as sepsis. When there would be a progression to organ dysfunction, the term severe sepsis would arise, which could in turn lead to septic shock, a sepsis-induced hypotension, even with adequate fluid resuscitation [1], [26]. The diagnosis could be achieved through the Systemic Inflammatory Response Syndrome (SIRS) criteria, which assesses the following clinical parameters: body temperature, heart rate, respiratory rate, and white blood cell count [26]. This definition of the syndrome showed to be inefficient to describe septic patients, due to the variety of medical causes for infection, with SIRS having shown to have insufficient specificity [1], [2], as up to 90% of patients admitted to the ICU settings meet the defined thresholds [26].

In 2016, the Third International Consensus Definitions for Sepsis and Septic Shock [1] took place, where the most recent definitions of the syndrome and its detection criteria were then agreed upon. Sepsis is now defined as a potentially fatal organ dysfunction

caused by an uncontrolled response from the host to an infection, being dependent on parameters related to the host (such as genetics, comorbidities, and race) and the pathogen. The term severe sepsis was found to be redundant, and septic shock is now defined as a subset of sepsis, in which there are underlying circulatory and cellular or metabolic abnormalities that substantially increase the risk of mortality [1].

According to the World Health Organization [27], the syndrome manifests itself with the following symptoms: fever or low body temperature, altered mental state, difficulty breathing or high respiratory rate, high heart rate, low blood pressure, low urine output, cyanosis, or mottled skin (with irregular spots), and intense body pain or discomfort. This goes to show how hard it would be to diagnose sepsis from its symptoms, as they are very common to many other diseases and conditions.

Adding to the low specificity of the symptoms, the syndrome also has a complicated pathogenesis, with many influencing mediators, both immune and non-immune. The presence of inflammation can result from various underlying disease processes, as well as prior use of antibiotics that result in cultures being negative. Some biomarkers that have been studied for sepsis detection are C-reactive protein (CRP) and Procalcitonin (PCT) [6]. CRP, thought to be the most commonly researched biomarker for the syndrome [6], was initially found to be effective for sepsis detection [28] but was later concluded that it lacked in specificity as it can rise in many inflammatory illnesses [29]. PCT plasma concentrations have also been found to be promising [30], [31] and even better than CRP [32], but septic shock was shown to be difficult to diagnose through this biomarker [31]. Thus, there is no molecular profile of specific biomarkers that can unambiguously identify a patient with the illness or predict their future prognosis [2].

The syndrome requires urgent detection, as a patient with a suspected infection that has even a moderate degree of organ dysfunction, is 10% more likely to die while getting treatment [1]. The treatment focuses on delivering antibiotics on time, resuscitation, and controlling the cause of infection [2], [12], but a rushed treatment is not desirable, as it depends on the stage of the syndrome and may not be an easy decision to make. The administration of antibiotics, for instance, is almost always justified for the adequate treatment for patients with a suspected bacterial infection. Despite this, there is an increased risk of general antimicrobial resistance with treatments delivered in a rush [12]. It has also been shown that precipitated fluid resuscitation might not be as effective as previously thought and might be associated with an increase in the risk of organ dysfunction [11]. This serves to demonstrate how difficult it is to diagnose the syndrome, as well as to clinically manage it.

Despite this, there is strong evidence that early treatment of sepsis, achieved through early diagnosis, leads to better outcomes for patients [2], [5]–[10]. Therefore, it is critical to be able to detect the syndrome as soon as possible, so that adequate therapy can begin early in the disease's course, avoiding further deterioration.

2.1.1 Detection Criteria

The detection criteria of sepsis, along with its clinical definition, have evolved. As mentioned before, SIRS is no longer considered accurate enough for the task. Despite this, three other scores are widely used: Sequential Organ Failure Assessment (SOFA), quick-SOFA (qSOFA) and National Early Warning Score (NEWS).

SOFA and qSOFA

The current detection criterion of choice [1] for the screening of the syndrome is the SOFA score [33], which determines the degree of organic dysfunction of the patient. Table 2.1 represents its calculation, which is as follows: for each system affected and considering the respective clinical interventions, there is an addition to the baseline score of a value between 0 and 4. Organ dysfunction is identified with an overall SOFA score of 2 or more points.

Table 2.1: SOFA scoring system. Adapted from [33].

	Score				
System	0	1	2	3	4
Respiration					
PaO_2/FiO_2, mmHg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) w/ respiratory support	<100 (13.3) w/ respiratory support
Coagulation					
Platelets, $\times 10^3 \mu$ L	≥150	<150	<100	<50	<20
Liver					
Bilirrubin, mg/dL (μmol/L)	<1.2 (20)	1.2-1.9 (20-32)	2.0-5.9 (33-101)	6.0-11.9 (102-204)	>12.0 (204)
Cardiovascular	MBP ≥70 mmHg	MBP <70 mmHg	Dopamine <5 or dobutamine (any dose)	Dopamine 5.1-15 or epinephrine ≤0.1 or norepinephrine ≤ 0.1	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1
Central nervous system					
GCS score	15	13-14	10-12	6-9	<6
Renal					
Creatinine mg/dL (μmol/L)	<1.2 (110)	1.2-1.9 (110-170)	2.0-3.4 (171-299)	3.5-4.9 (300-440)	>5.0 (440)
Urine output, *mL*/day				<500	<200

As seen in Table 2.1, this criterion involves laboratory-dependent results (such as the partial pressure of oxygen, platelet count, etc.), so it may take long to assess, and preclude its use outside the hospital. For patients with a suspected infection, whose chances of staying in the ICU for a long time are high, a prompt diagnosis can be achieved with

qSOFA, whose parameters, shown in Table 2.2, can be quickly analysed [1].

Table 2.2: qSOFA scoring system. Adapted from [1].

At least two of the following:
Respiratory rate $\geq$ 22/min
Altered level of consciousness
Systolic blood pressure $\leq$ 100 mmHg

Both SOFA and qSOFA assess the level of consciousness of the patient. The first criterion implicitly uses the Glasgow Coma Scale (GCS), which is a clinical score that, similarly to SOFA, sums a value to the baseline score for each of several assessments made to the patient's eye-movements, as well as motor and verbal responses. The lower the sum, the more severe is the altered mental state, as seen in Table 2.1. In order to reduce the burden of measuring this score, qSOFA, considers a more general state of altered mentation. This, in essence, translates to a GCS that is inferior to 15 [1]. Despite this, for both criteria, there is an inherent subjectivity to these assessments, as they require the presence of expert healthcare professionals, to evaluate each case. That subjectiveness is, in part, the reason why they are not monitored through sensors. The responsible clinicians cannot be present at all times to assess this parameter, in particular when considering ambulatory conditions, which can be an obstacle to early diagnosis.

NEWS

Early Warning Scores are commonly used to detect patients at risk of clinical deterioration [3], which occurs when there are anomalies in the patient's vital signs, a likelihood of adverse outcomes, and there is a risk of consequences such as mortality, transfer to a higher level of care and prolonged hospital stay [34]. Because sepsis is a potential cause of severe illness, using these tools for detection of the syndrome has been seen as promising [3].

The NEWS is a criterion developed by the Royal College of Physicians to improve the detection and response to clinical deterioration in adult patients [35]. Its use with suspected sepsis patients has been recommended by the National Health Service (NHS) of England [35]. Similarly to SOFA and qSOFA, NEWS is composed of several parameters that are routinely monitored, as represented in Table 2.3, each one receiving a different score depending on its variation. The greater the score, the greater the risk of mortality. A score of 5 or more should prompt an urgent clinical response and is indicative of a potential case of sepsis.

Sepsis management is a complicated, time-dependent task that requires highly experienced and trained professionals. Its detection and progression are influenced by many types of clinical data, which makes it difficult to obtain a rapid diagnosis. However, as AI continues to innovate in the medical field, it is expected that some of these choices may

eventually be delegated to machines that can be seen as "intelligent". This is where it may be of valuable help, in assisting in clinical practice and improving patient outcomes [6].

Table 2.3: NEWS scoring system. Taken from [35].

Physiological parameter	Score 3	2	1	0	1	2	3
Respiration rate (per minute)	≤8		9–11	12–20		21–24	≥25
SpO_2 Scale 1 (%)	≤91	92–93	94–95	≥96			
SpO_2 Scale 2 (%)	≤83	84–85	86–87	88–92 ≥93 on air	93–94 on oxygen	95–96 on oxygen	≥97 on oxygen
Air or oxygen?		Oxygen		Air			
Systolic blood pressure (mmHg)	≤90	91–100	101–110	111–219			≥220
Pulse (per minute)	≤40		41–50	51–90	91–110	111–130	≥131
Consciousness				Alert			CVPU
Temperature (°C)	≤35.0		35.1–36.0	36.1–38.0	38.1–39.0	≥39.1	

2.2 Machine Learning

AI is defined as the field that aims to construct and view computational systems with properties that mimic human intelligence [36].

ML consists of a specific area of AI, in which machines make use of data that is provided to them and, combining statistical and analytic techniques with computer science, create algorithms that can learn or extract information from this data [20]. This learning process can be categorized in several ways, such as the dichotomy: **supervised** and **unsupervised**.

2.2.1 Supervised Learning

Having a labeled dataset $(X_l, Y_l) = \{(x_1, y_1), ..., (x_l, y_l)\}$, where $x_i \in \mathbb{R}^D$ is the i-th D-dimensional data vector, and $y_i \in \mathbb{R}$ or $y_i \in \{1, ..., M\}$ is the class of the data vector x_i [37], the two most common types of problems that supervised ML solves can be split in [38]:

- **Classification** problems: when the goal is to distinguish or predict the **label** of the input from a given set of possibilities, and therefore, y_i is the respective class of x_i, from the M possible classes [37] (for example, we speak of binary classification when $M = 2$, with the outcome class usually being either positive or negative, and multiclass classification if there are $M > 2$ possible labels for the outcome);

- **Regression** problems: when the goal is to predict a continuous value or a real number, given a set of rules that the model follows, being y_i the fitting of x_i.

The goal of a supervised ML model is to create an algorithm capable of predicting an output, when presented with a given input dataset $(X_{train}, Y_{train}) = \{(x_1, y_1), ..., (x_{train}, y_{train})\}$, so that after being trained with a large enough amount of data, it will deliver accurate predictions when exposed to new and previously unseen data [20], $(X_{test}, Y_{test}) = \{(x_1, y_1), ..., (x_{test}, y_{test})\}$. The methodology for the construction of the models may differ according to the problem or task at hand, but it generally consists of the following steps:

1. **Preprocessing** the dataset, to create what is often called the training data;
2. **Learning** the algorithm, which includes training the model to reach the desired outcomes; and validating it, to reduce the probability of overfitting to specific data sets;
3. **Testing** its performance on predicting the outcome when it receives new data, often called testing or evaluation data, with some evaluation metrics (for example, its accuracy).

For the scope of this work, the detection of sepsis is seen as a supervised classification task since, as previously mentioned in Chapter 1.2, the goal is to identify septic and nonseptic patients in the used dataset, as early as possible.

2.2.1.1 Supervised Classification Models

ML models have been used in medicine for a wide variety of applications [20]. Decision Tree (DT) based models in particular (namely Random Forest (RF), Gradient Boosting Decision Tree (GBDT), Adaptive Boosting (Adaboost) and eXtreme Gradient Boosting (XGBoost)) have shown good results related to sepsis diagnosis [22]–[24], [39]–[43].

Decision Trees

DT is thought to be the most often used model, when it comes to decision-making, in the realm of ML [44]. It consists of a tree-based algorithm with several nodes, each corresponding to a feature, that eventually lead to a result. Each node is iteratively split into branches that are associated with the outcome possibilities, often denominated as leaves [38]. Figure 2.1 represents an example of a decision tree algorithm within the context of sepsis detection, with binary nodes where each node has two branches associated with the true and false value of the condition. It is important to note that the figure might not be clinically accurate and is used as an example, purely for explanation purposes.

The use of this classifier has the drawback of often resulting in an overfit model, which means it lacks the ability to correctly predict outcomes with data it has not been trained

with (the concept of overfitting will be explained in further detail in Chapter 2.2.5). On the other hand, its branching structure allows for a rapid interpretation of the learnt classification strategy [38].

The fundamental concept behind a DT is to find features that hold the most information related to the target label, and thus splitting the dataset in such a way that the most informative parameter is the one that best distinguishes the target located at the resultant leaf nodes.

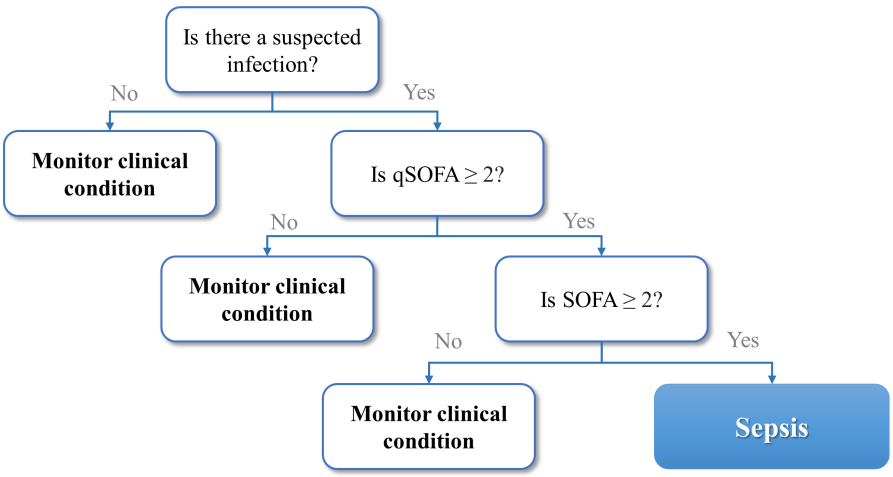


Figure 2.1: Example of a DT.

Consider the first node of Figure 2.1 to be node m, with its data represented by Q. At the moment of splitting the node, it will result in the right and left child nodes, Q_r and Q_l, in this case, in the Yes and No nodes. The **impurity** at node m, which is given by a function H, measures how helpful this split of the data is, for the classification of the patient. The best split will minimize the weighted average of impurity of Q_l and Q_r [45]. In a scenario where it is desired to detect sepsis with varying degrees of disease intensity, for node m, the output y can have values of $1, 2, ..., K$, with the proportion of each class k in node m being given by Equation 2.1:

$$p_{mk} = \frac{1}{n_m} \sum x_i \in m(y_i = k) \tag{2.1}$$

The measurement of impurity at this node can be done with different criteria, being the **Gini impurity** one of them. It achieves this by subtracting the sum of squared probabilities of each class from one [45]:

$$H(m) = 1 - \sum_{k=1}^{K} p_{mk}^2 \tag{2.2}$$

Thus, with every split, during DT construction, measurements of impurity in that node will make sure that the combined impurity of the two children nodes is lower than

that of the parent node. The split of the data that minimizes this combined impurity, is the best split, i.e., the one that holds more information [45]. The concept of impurity will also be relevant in the context of feature selection, which will be explained further in Chapter 2.2.3.

Besides DTs, ensemble models based on these are often used, which combine several models to create a better performing one. There are two popular techniques of ensemble [45]:

- **Bagging**: where the data used for training of the models is resampled, each tree resulting from a different sample from the training data;
- **Boosting**: where many trees are sequentially created, giving the subsequent trees the ability to correct errors from any previous one.

Random Forests

A RF is based on bagging ensemble, and consists of many DTs connected, belonging to the same "forest", each tree being somewhat different from the other since a different sample of the data is used. Therefore, the many trees compensate each individual tree's overfitting, while maintaining its prediction power. This is achieved by splitting each node not according to the best features, i.e. with lowest impurity, but according to the best features among a subset of predictors chosen by chance at that node [46].

Gradient Boosting DTs and eXtreme Gradient Boosting

Both GBDT and XGBoost are ensemble boosting models, as their name suggests.

GBDT adds a new simple DT to the main tree with each iteration, that considers the errors made in the previous iterations, employing a gradient descent technique to minimize loss when adding the new trees. The model is then capable of learning from previous errors and compensates for them [38].

XGBoost follows the same principle as GBDT but optimizes computing speed, learning performance, and memory resources. It does this by adding a regularization term that influences the intricacy of the model [47].

Adaptive Boosting

The Adaboost classifier, also a boosting ensemble model, begins by creating many simple DTs, that are characterized by having a weak predictive power and are, therefore, weak learners. Each next learner to be created will be influenced by the error from the previous one. All learners are initially considered to have the same weight for the final ensemble, but the influence of wrongly categorized data points is increased each round, forcing the weak learner to improve its ability in predicting the challenging examples in the training set [48]. This way, this classifier combines many weak learners, similarly to RF,

with different values/weights, each accounting for the mistakes of the previous one, like GBDT.

K-Nearest Neighbours

Many more models exist in the area of ML, including less complex models that are not based on DTs. The KNN model, for instance, can predict a value for a new data point by locating the closest data points in the training dataset, with k corresponding to the number of surrounding data points. It, therefore, presumes that similar data points will exist near each other. There are several methods for measuring the distance between them, the most common of which is the Euclidian distance. One of its virtues is its simplicity and how easy it is to comprehend. Besides, it typically does not need a lot of parameter tuning to produce acceptable results. Despite all of this, it is often slow, and cannot handle a wide range of features [38].

2.2.2 Unsupervised Learning

Opposite to supervised learning is the unsupervised learning approach, where the algorithm itself unveils characteristics within the dataset, without being given information about the results upfront. The dataset can then be represented by $X_{train} = X_u = \{x_1, ..., x_u\}$, and is therefore, unlabelled. This is useful to discover specific features of the dataset that are not evident [36]. Because of this, unsupervised learning models are often useful during the preprocessing of data [38], which was the case for this work, as will be mentioned in Chapter 4.3.3.

An example of an unsupervised ML model is **K-means clustering**, illustrated in Figure 2.2. The main objective of clustering models is to group the data in such a way that points within a particular cluster are very similar to each other, and points in different clusters are not [38]. K-means clustering, in particular, works by assigning each data point to a cluster center, based on the mean distance to it, with k being the number of centroids.

The following four steps explain the process of a k-means clustering model [49]:

1. It begins by randomly assigning k data points as cluster centroids;
2. Each of the remaining data points is then assigned to a cluster, according to the mean distance to its center;
3. It recalculates the centroid as the mean distance of all data points assigned to the cluster;
4. The process repeats until the assignment of centroids does not change.

There are specific methods that determine the optimal number of clusters. One of them is called the **elbow method**, where the sum of squared errors between each data point and the cluster center is calculated for a range of values for k. The function will

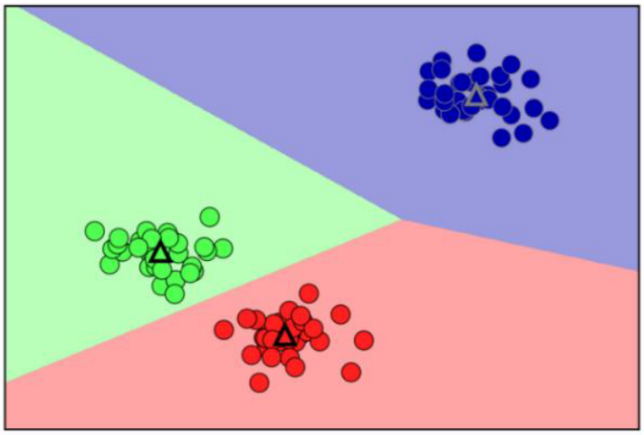


Figure 2.2: Representation of a K-means clustering with $k = 3$. The black triangle represents the cluster centroid. From [38].

have an elbow shape since this sum value decreases rapidly as the number of clusters approaches its ideal value. It will continue to decrease beyond the ideal point, but it will become much slower, allowing for the determination of the optimal value for k. This method is simple and easy to understand, but it has drawbacks, namely if the inflection point is not obvious, in which case, k cannot be determined [50]. The sum of squared errors is often called inertia.

2.2.3 Data Preprocessing

As mentioned previously, the first step to develop a supervised learning model is the **preprocessing** of the data that will be used to train it.

The steps to achieve this can vary according to the given problem or task. Despite this, four main steps are often used and are transversal to many ML applications:

1. Perform **data imputation** techniques, to deal with missing values within the dataset;
2. Perform **feature engineering** techniques, to extract features from the raw data;
3. Perform **dimensionality reduction** or **feature selection** techniques on the dataset, especially for datasets with a high number of features;
4. **Scale** the features within the dataset.

Data imputation

One of the biggest obstacles of working with EHRs containing clinical information, whether that is for simple data analysis or for building prediction models, is the frequent high percentage of missing data [51]. These EHRs have shown to be very inconsistent and

incomplete [51]–[54]. Certain medical parameters might have been measured for a certain type of patient, but not for others. The frequency of such measurements also varies greatly, meaning that the collected data can vary greatly from patient to patient [52].

Excluding the cases or data points that are missing is the most common and simplest technique for handling these missing values [52]. Nevertheless, this can reduce the amount of information taken into account for the learning task, which is usually not desirable. There are methods for imputing data that prevent this, such as replacing it with its previous value, its next available value, or some other value, like its mean. The fact that a value is missing, might even be a feature. Besides the simpler methods, more complex approaches have been used for missing data imputation [51], including methods that make use of ML models [52], [53].

Feature Engineering

Feature engineering can broadly be described as the process of finding the representation of the data that optimizes the model's ability to learn the dataset's characteristics, and better helps the task in question. This can consist of encoding a certain feature, like a binary encoding of the patients' gender in a dataset. It can also consist of constructing features that are made up of the interaction between two variables [38], like the calculation of sepsis detection criteria scores (such as SOFA and qSOFA scores) and using these as features.

Dimensionality Reduction

Dimensionality reduction consists of transforming a high-dimension dataset into a lower-dimension representation of the same data, assuming that this lower-dimension representation has most of the relevant information that describes the original data. This helps to eliminate unnecessary information and improve learning accuracy [55].

A common technique used for this is **Principal Component Analysis**. It makes use of new variables, called the principal components, that are linear functions of the variables in the original dataset, but are uncorrelated with each other, and that maximize the amount of variability by them represented while decreasing the dimension of the original data [38]. A drawback of this method, though, is that these components are generally difficult to interpret, and it usually does not provide information regarding the influence of each original feature on the new ones [55].

Feature selection is one of the most basic methods for reducing dimensionality [55]. It also focuses on finding a subset of features within the dataset, according to a specific condition, that represents most of its information. The two approaches differ, however, since feature selection does not transform the dataset per se, but rather chooses which features to use for the training of the models, without changing them. As a result, they maintain the original meaning of the variables, providing the benefit of easy interpretability [56].

The methods often studied within the realm of feature selection for ML can generally be described in three main techniques: **filter**, **wrapper**, and **embedded techniques**.

Filter-based approaches evaluate the importance of the features solely based on intrinsic attributes of the data. It usually involves the calculation of a feature significance score that is then used as the criterion with which to exclude, or not, certain features. It often is a computationally simple task to implement, since it is performed before training of the models, and is also applicable to high-dimensional data. Despite these benefits, it usually does not consider possible dependencies among features [56].

Wrapper methods, often called greedy methods, create and evaluate many different variations of the feature subset, using them for several training and validating tasks of a certain model. These consider feature dependencies and are generally associated with better performances when compared to filter methods, but involve a higher computational cost, and are more prone to overfitting [56].

Examples of these are the Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS). The first, SFS, iteratively adds the feature which optimizes the accuracy of the model, to the feature subset, beginning with the feature that alone provides the highest value for this optimization. It is possible to then think of this first feature being added, as the most important. In SBS, the approach is similar, but instead of including features, it iteratively removes them from the subset. The first feature to be removed is the feature, among the complete subset, that contributes less, and i.e., is the least important. Two more methods that are relevant in the context of this project are the previous methods' floating versions, Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS). These apply the same process as mentioned earlier, but have, in each iteration, an additional element that evaluates for each feature already chosen/excluded, if removing/adding another feature improves the performance [57].

Finally, **embedded** techniques consist of hybrid approaches, that combine the advantages of filter and wrapper feature selection. They are usually specific to a certain algorithm since the task of finding the ideal subset of features is built into the classifier [56].

DT-based models constitute an example of algorithms that perform embedded feature selection [56]. Two main types of variable importance can be determined from these, which are the Mean Decrease in Impurity (MDI) and Mean Decrease in Accuracy (MDA) [45]. The first, MDI, was already explained earlier in this Chapter (refer back to Equation 2.2) as it is one of the existing criteria that these models can use to build the trees since the splitting of the branches is done through minimizing the feature impurity. As the name suggests, it determines the importance by averaging the decrease of impurity of each feature, across the whole forest or ensemble of trees. It has the drawback, though, of being biased towards continuous features that increase the number of possible splits in the tree [46]. MDA, on the other hand, is usually calculated during a specific feature

importance method called **permutation**, which consists of iteratively replacing the instances of each feature with random values. This way, a feature can be seen as important if the change in its values led to a mean decrease in the accuracy of the model [46].

Dataset Scaling

Scaling the dataset is important to ensure that different features have the same importance, regardless of their value range. There are many methods to achieve this, the most common ones being **normalization**, in which every feature is scaled to be in the same range of values, and **standardization**, in which the features are scaled by subtracting from mean and dividing by standard deviation [38].

2.2.4 Deep Learning

As it was mentioned already, ML is an area within the context of AI. DL is also an area within this context, specifically within the area of ML, that is based on Artificial Neural Network (ANN) algorithms, which use the logic of human reasoning, based on the communication between neurons, through synapses [58].

An ANN has, at least, three layers, with a variable number of neurons in each one, as represented in Figure 2.3: an input layer, that receives the data; at least one hidden layer, where the most relevant characteristics of the input data will be extracted (in Figure 2.3, three hidden layers are visible); ending with an output layer, that gives the result of the prediction.

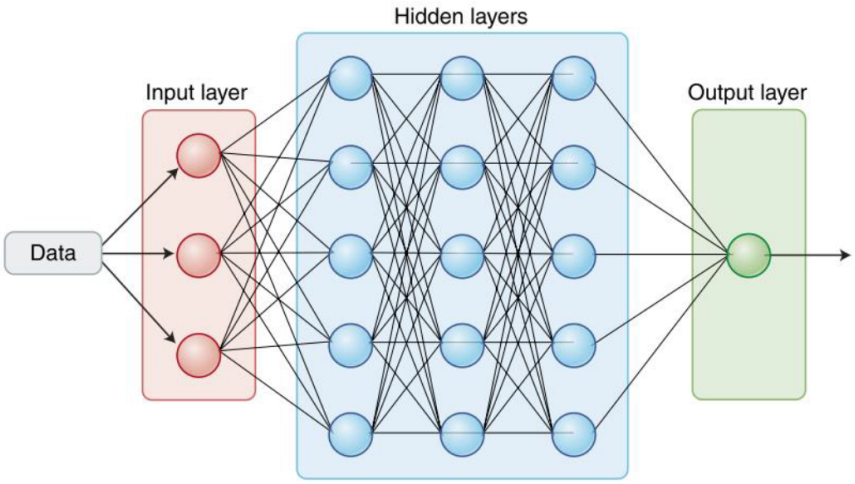


Figure 2.3: Typical structure of a NN with 3 hidden layers. From [17].

Each neuron can receive a signal, process it, and transmit it. In a fully connected network, each neuron is then connected to each neuron in the layer that succeeds its own. The connections have weights which reflect the degree of importance of the connection

through a real number. Each neuron (except for the input ones) receives signals from all the neurons from preceding layers. That signal, for each efferent neuron, is itself the weighted sum of its inputs, to which is applied a given activation function.

Thus, a DL model consists of an ANN (also called a Deep Neural Network (DNN)), with a high number of hidden layers, and non-linear data transformations that may help not only in the extraction of important features, but also in the suppression of less relevant ones, and so improve the overall network performance. Despite this, these algorithms have their drawbacks, namely the lack of interpretability inherent to them due to their complexity (so much so, that they are often called "black box algorithms") [17]. In a DNN, there can be hundreds of millions of adjustable weights, which imposes, as well, the need for a great number of data samples for the training procedure [59]. All of this leads to long training times, often computationally demanding, as well as a lot of parameter tuning [38]. Even so, unlike general ML algorithms, DL does not usually need a mandatory pre-stage of feature extraction processing, prior to the classification task. This happens because the earlier layers in the DNN perform that task, implicitly. That ensures a greater ability to handle complex data and to extract features from it [59].

DNNs and DL, in the context of sepsis detection, have shown good results for the early diagnosis of the syndrome [60]–[62], which is the reason why they are being mentioned in this dissertation project. Even though for our method the focus is on the previously mentioned ML models, as well as the data preprocessing, it is believed that almost all health professionals will come into contact with this technology in the future, thanks to its great potential [17] and for this reason, they are being introduced in this Chapter.

Some specific architectures that have been found useful are:

- Convolutional Neural Network (CNN): widely used in image recognition, these NNs use layers of convolution where the input data is filtered via the convolution operation, performing the dot product of the input matrix with a filter matrix, resulting in maps of the extracted features. These are followed by pooling layers, in which each feature map is sub-sampled, and the number of trainable parameters is significantly reduced. They usually end with a fully connected layer that provides the final output. These components ensure its ability to generalize and learn features with a high degree of abstraction [63]. Time series, such as the ones used in this work, can use similar approaches, where convolution occurs along the temporal dimension;

- Recurrent Neural Network (RNN): these make use of inputs as a sequence, processing one element at a time from the data. The hidden layers have neurons that have a connection with themselves, used for saving its value in successive iterations. This way, the output of a previous iteration is taken into consideration in the present iteration. These prove to deal well with dynamic information but struggle to retain information for long periods of time [59].

- Long Short-Term Memory based Neural Network (LSTM): these are an alternative architecture to RNNs, since they not only have similar hidden neurons capable of retaining information but also units that learn to manage the saved information, by deciding when the neuron should forget its saved value or not (often denominated as forget-gates) [59].

2.2.5 Model Overfitting and Evaluation

It was previously mentioned, in Chapter 2.2.1, that the dataset is usually split in multiple groups, namely the **training** and **testing** groups, when developing a ML model.

The reason for this splitting stems from the concept of **model overfitting**, and how to evaluate a model's performance when completing its task. More than being accurate in predicting the outcomes when using data it was trained with, it is important that a model is generalizable and provides good results with data it has never seen before, indicating that it has not been overfit to the particular training dataset [38].

Thus, the original dataset is split into the multiple smaller datasets, usually resulting in three different groups: the **training** group, **validation** group and evaluation or **testing** group. These are represented in Figure 2.4 and each has its purpose. The first two groups are used to develop the model, the training dataset being used to train the algorithm and the validation dataset to assess it and adapt its parameters while it is being designed. The evaluation dataset is used to test the performance of the final model [38].

K-Fold Cross Validation

A common technique used to avoid model overfitting during training consists of using K-Fold Cross Validation, which is also represented in Figure 2.4.

It achieves this by splitting the training data into different groups, called **folds**, with k being the parameter that defines the number of folds, and validates the model several times, with different training and validation datasets. Each group is used as a validation dataset once and as a training dataset $k-1$ times, with the overall evaluation metrics being the mean value of all the individual values with each validation dataset. When k is equal to the total number of examples in the dataset, the method is called **leave-one-out**, being each example left out once for evaluation of the model [38].

Performance Metrics

After developing the model and performing all the validation tests needed to tune its parameters, its performance is assessed with the test dataset, the partition of the data that has never been seen during the training of the model. Some important measures, associated with model evaluation, are the True Positive (TP) and True Negative (TN) predictions, which correspond to correctly predicted classifications, False Positive (FP) and False Negative (FN) predictions, which in turn correspond to falsely predicted classifications [38].

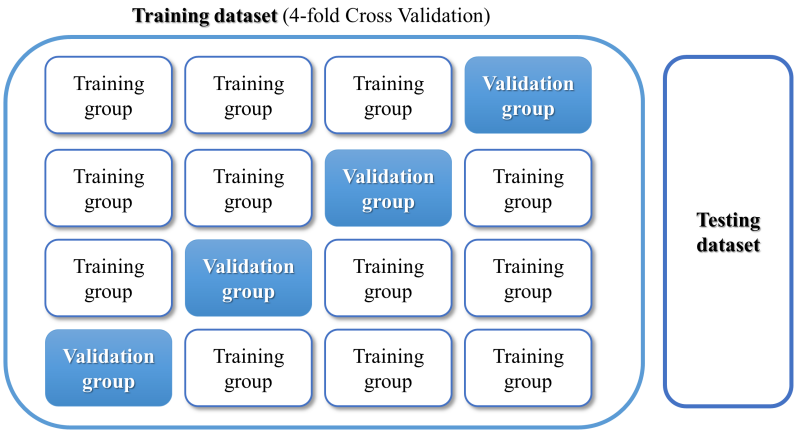


Figure 2.4: Representation of a split dataset for ML model development, with k-fold cross validation. In this example, 4-fold cross validation is used.

A model's **accuracy** can be summarized by the ratio between the correctly predicted outcomes, and the total predictions made by the model [38]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

A model's **precision** indicates how often the positive outcomes are correctly labelled as positive. This might be important for classification tasks that need to minimize FP, such as predicting the onset of sepsis when it is not occurring. It is described by the following equation [38]:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

A model's **sensitivity** or **recall**, on the other hand, is an important metric when trying to minimize FN, like failing to identify the onset of sepsis, since it measures the ratio of correctly labelled positive outcomes to the total number predicted positive outcomes. It is also called True Positive Rate (TPR) and is given by [38]:

$$TPR = \frac{TP}{TP + FN} \tag{2.5}$$

Confusion matrices are often used to represent the evaluation of classification problems, due to its interpretable visualization. It represents how often a sample belonging to a certain class has been correctly identified or not. Figure 2.5 represents an example of a confusion matrix for a binary classification problem. If the problem was multiclass classification, the number of rows and columns would increase according to the number of classes [38]. Ideally, one would want their confusion matrix values to be concentrated, as much as possible, along the main diagonal of the matrix. In fact, if all the values are disposed along the diagonal, it means that the model did not give any false predictions.

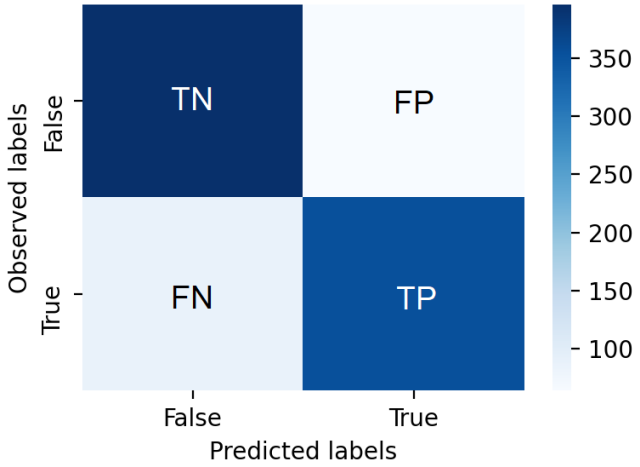


Figure 2.5: Representation of a confusion matrix for a binary classification task.

ROC Curve and AUROC

The Receiver Operating Characteristics curve (ROC curve), as the ones represented in Figure 2.6, shows the model's performance at a certain task as its discrimination threshold changes. The plot corresponds to the False Positive Rate (FPR) against the TPR. Similar to the TPR, FPR is given by the ratio between the FP and all the negative examples:

$$FPR = \frac{FP}{FP + TN} \tag{2.6}$$

The Area Under the ROC curve (AUROC) is extensively used to easily assess a model's performance and to compare several models. Ideally, a well-performing model will have a high TPR and a low FPR, represented by a ROC curve that rises and saturates very rapidly [38] and is close to 1. In Figure 2.6, two ML models, one based on RF and the other on KNN, are presented as an example, where the RF achieved a higher AUROC than the KNN model, and therefore has a ROC curve that saturates quicker.

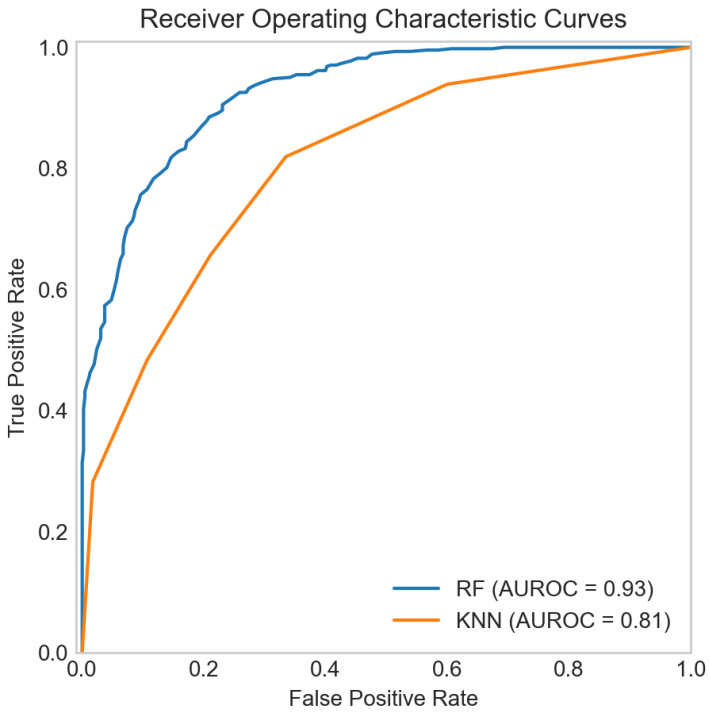


Figure 2.6: Example of two ROC curves.

Chapter 3

State of the Art

This chapter details a literature review regarding the state-of-the-art research on the use of AI to obtain a sepsis diagnosis. It begins by analysing and comparing the clinical detection criteria mentioned previously, for tasks like predicting in-hospital mortality. Then, the use of ML for sepsis detection is delved into, as well as methods for feature selection regarding the analysis of sepsis patients. Recent work within the DL area, developed in the context of this project, is also described.

3.1 Sepsis Detection Criteria

As previously mentioned, SOFA has recently been chosen as the sepsis detection criterion of choice during the Third International Consensus Definitions for Sepsis and Septic Shock, with the use of qSOFA being recommended when there is a need to quickly assess patients with a suspected infection [1]. Besides these two criteria, The NHS has recommended NEWS as the early warning score to detect clinical deterioration in patients with suspected sepsis, due to its easy assessing and to the fact that it evaluates vital signs that are already routinely monitored in the UK's healthcare system [35].

SOFA, qSOFA and SIRS have been studied by Raith et al. [64] in terms of their prognostic accuracy for in-hospital mortality for adult patients with suspected infection in the ICU. This retrospective analysis accounted for 184 875 patients between 2000 and 2015. They found that SOFA discriminated substantially better in-hospital mortality than qSOFA or SIRS for admission within 24 hours. The three criteria achieved AUROC scores of, respectively, 75.3%, 60.7% and 58.9% for in-hospital mortality, and 73.6%, 60.6% and 60.9% for in-hospital mortality or an ICU longer than three days.

Khwannimit et al. [65] also compared the same three criteria, for mortality and organ failure prediction in the ICU. This was done through a 10-year retrospective cohort study

of 2 350 patients with sepsis. It was concluded that SOFA had the best performance in discriminating hospital mortality, with an AUROC of 83.9%, while qSOFA and SIRS achieved, respectively, 81.4% and 58.7%. SOFA also performed better at predicting organ failure, with 99.4%. qSOFA and SIRS, on the other hand, achieved 84% and 66.9%, respectively. This allowed them to also conclude that qSOFA had a significantly better prognostic accuracy than SIRS.

Later, Khwannimit et al. [66] studied several early warning scores, including NEWS, as well as qSOFA and SOFA, to predict mortality among patients admitted, once again, to the ICU. For this, another retrospective study with a population of 1 589 sepsis patients was conducted. It was confirmed, once again, that the SOFA score had the best accuracy for predicting not only 30-day mortality but also multiple organ failures among septic patients, with the best AUROC achieving a value of 88%, followed by qSOFA with 84.7%, which outperformed NEWS, with 83.3%.

SIRS, NEWS and qSOFA were compared by Usman et al. [13] when it comes to sepsis screening, this time in an Emergency Department (ED) setting. Their study was done with a population of 130 595 adult visits, of whom 930 were sepsis patients and, unlike Khwannimit et al. [66], it was found that NEWS was not only more accurate than qSOFA, but also that it improves as the severity of the illness goes up. qSOFA was found to be a poor sepsis screening tool for early detection, with SIRS performing better but showing no statistically significant difference for predicting sepsis-related mortality. For detection of the syndrome, qSOFA showed to have the lowest sensitivity score, by far, with 28.5%, while NEWS and SIRS had 84.2% and 86.1%, respectively.

Song et al. [15] compared the predictive efficacy of qSOFA and SIRS but this time for predicting in-hospital mortality in patients with suspected or confirmed infection who were not in the ICU. This was done through literature research, which included 23 papers and a total of 146 551 patients. Their findings showed high specificity of qSOFA for early detection of the in-hospital mortality when compared to SIRS, with 83% versus 29% respectively, but they also concluded that it lacked sensitivity, achieving 51%, contrarily to the second criteria, that had a score of 86%. Askim et al. [14] had also already studied the clinical utility of these two criteria, before arriving at the ED, and found that qSOFA is not a reliable diagnostic tool for sepsis in this setting, having achieved a sensitivity of only 32% for detecting patients with severe sepsis at the time of admission.

Finally, Lim et al. [67] performed a retrospective study of the use of NEWS. This was done by analysing the patient's deterioration condition, and one of the three possible outcomes: 1) the patient had to be transferred to the Intensive Care Area, 2) the patient had to be transferred to the ICU, or 3) the patient had died within 24 hours of a vital signs observation set. This study counted with 11 300 patients between 2015 and 2017, and they concluded that the criteria correctly identifies patients in infection-related acute medical situations based on the risk of poor outcomes, since the AUROC over 24 hours, for all three scenarios, was 89.6%.

Conclusion

With the presented literature, it is hard to conclude which criterion is best. On one hand, specialists recommend SOFA as the criterion of choice, but it is a complex score to calculate, which may delay the diagnosis and therefore does not suit the goal of this dissertation project. On the other hand, qSOFA has been suggested as an early detection tool, but it has been shown to lack in sensitivity [13]–[15], which is important for a syndrome like sepsis, that needs an urgent diagnosis [1]. The analysed results regarding NEWS do not allow for a decisive conclusion, as it has had good performance [13], [67], but it has also been outperformed by other criteria [66]. All of this emphasizes the complexity of the syndrome and how difficult it is to detect, even with the available clinical screening tools.

3.2 Machine Learning

The use of ML for early diagnosis of sepsis is a widely researched topic. The amount of literature on this subject is substantial, increasing each year. A good example that supports this claim is the *2019 PhysioNet and Computing in Cardiology Challenge*, a yearly competition that aims to find solutions for problems in the medical field. This year's goal was focused precisely on the early diagnosis of sepsis through clinical data [68].

3.2.1 Sepsis Detection and/or Prediction

Taylor et al. [39] have developed a RF model to predict in-hospital mortality and compared it to a classification and regression tree (CART) model, a logistic regression model, and two previously developed prediction metrics based on the validation dataset, one of them being the AUROC. For this, a retrospective analysis was performed, of admitted patients to the ED during one year. From the 5 278 admitted patients, 4 676 were identified as sepsis patients according to the SIRS criteria. The model was developed based on more than 500 clinical features and the training dataset was split into 80%/20% training and validation data. The constructed model achieved an AUROC of 86%.

Mao et al. [22] used GBDT for detection and prediction of the three stages of sepsis (sepsis, severe sepsis, and septic shock). For the data preprocessing, missing values were replaced with the previous hour's value, and multiple values for the same hour were replaced by their mean. 10-fold cross-validation was used for training. Each tree was limited to split six times and each iteration of the gradient boosting had a maximum of 1000 trees. It showed a better performance than the several sepsis scores (not only but including SOFA), with AUROC values being 92% and 87% for detecting the onset of sepsis and severe sepsis respectively. In terms of predicting septic shock four hours in advance, it achieved values of 96%. It was the first screening tool at the time to have exceeded AUROC values of 90% using only vital sign data.

Mitra and Ashraf [40] have compared several ML models for detection and prediction of the same three sepsis stages from patients in the ICU, using only vital sign data, like Mao et al. [22]. The clinical data was also extracted from MIMIC-III, one of the used datasets in this dissertation project, taking into account a total of 1 785 sepsis patients. The trained models were based on Logistic Regression, RF, XGBoost and a shallow NN. Besides these, an ensemble of the best three performing models (the last three mentioned) was studied. For training, the data was split into 70% training, 10% validation, and 20% for testing. Out of the mentioned individual models, the best performing one was RF, followed by XGBoost and the NN, with LR having the worst results. The ensemble model, on the other hand, surpassed all the scores achieved by the individual models. Regarding the three stages, sepsis, severe sepsis, and septic shock, it reached AUROC values of 97%, 96%, and 91% for the detection task, and 90%, 91%, and 90% for the prediction task.

Delahanty et al. [23] used the same model as Mao et al. [22], GBDT, as a feature selector to create a new screening tool for early identification of patients at risk of developing sepsis. EHRs from 49 urban community hospital emergency departments during 22 months were used. Some features were selected according to specialists, while others were engineered from a combination of features, and when there were multiple observations of the same data point, several summarizing parameters were considered, such as first and last available, mean, and minimum values. The missing values were replaced with extreme values (-9.999). 5-fold cross-validation was used, and each tree was limited to three splits. Out of 217 features, the model retained 13, with lactic acid being the most important contribution for the model. Despite this, the model still surpassed the criteria when lactic acid was not considered. It performed better in terms of sensitivity and precision than SOFA and was the most discriminant across all the different time thresholds taken into consideration with an AUROC between 93% and 97%.

Le et al. [41] have studied a boosted ensemble of DTs for early detection of paediatric severe sepsis. The model was trained with a dataset containing 101 anonymized paediatric patients, labelled with severe sepsis, with 4-fold cross-validation. It outperformed the relevant clinical detection criteria, having reached an AUROC of 91.6% from classifying the paediatric sepsis patients and the control patients at the time of onset, and 71.8% at four hours before the onset of the syndrome.

Yuan et al. [24] have developed a XGBoost-based model to obtain a sepsis diagnosis, and compared its performance with the SOFA score. 106 features were selected by a responsible specialist and another five were derived from vital sign data. Two approaches were used to handle missing values: for vital signs, the data was excluded, for other point-of-care data (like patient information, medication, etc) the missing values were replaced with the corresponding median values. 5-cross validation was used during training. The model achieved an AUROC of 89%, while the SOFA score that was calculated for the same population achieved an AUROC of 59.6%.

The varying unregulated responses of patients during the onset of sepsis, as previously mentioned, is an obstacle to its diagnosis. This factor has been emphasized by

Ibrahim et al. [69] by demonstrating the improved performance in the specificity of their models when the feature selection takes into account subpopulations with different clinical manifestations of the syndrome. The missing values were substituted using k-nearest neighbours, with $k = 7$. To identify these subpopulations, clustering and self-organizing maps (an unsupervised ANN algorithm) were used, based on the SOFA score parameters, each subpopulation's features being identified through RF. Four subpopulations were detected, representing the following organic dysfunctions: liver system disease, cardiogenic and renal dysfunction, minimal organ dysfunction, and, finally, cardiogenic dysfunction with hypoxemia and altered mental state.

Burdick et al. [42] have developed an ensemble model based on GBDT and XGBoost to predict severe sepsis, up to 48 hours in advance. This was done with a retrospective study of EHR from more than 450 hospitals, accounting for 270 438 patients. Features were established from vital signs that are frequently assessed in clinical settings. Tree branching was limited to six levels, and the final number of trees in the ensemble was defined as 1000. The learning rate for the XGBoost was set to 0.1. The results showed that the model surpassed clinical detection criteria (like SOFA), having reached an AUROC of 93% at the time of onset. For the prediction task, 48 hours before onset, AUROC was 82.7%.

Darwiche et al. [43] created a model based on Adaboost ensemble, with a Cox regression model to add a risk factor to it, to predict septic shock in an ICU setting. The techniques for imputation of missing data were based on carry forward and carry backward methods, but because of the high number of missing values in the used dataset, the total population consisted of 720 patients. The training of the model was done with 10-fold cross-validation. It showed great performance, achieving an evaluation accuracy of 95.77%, sensitivity of 88.89%, and specificity of 98.11%.

3.2.2 Feature Importance

In the context of the secondary goal of this dissertation project, which focuses on understanding which clinical parameters are most important for the sepsis classification task, recent studies have employed ML for this very purpose.

With the construction of the previously mentioned predictive model for septic shock progression, Mitra and Ashraf [40] have also obtained a feature ranking for the six vital signs taken into account as features, not only for the sepsis detection task but also for septic shock prediction. In terms of the first case, temperature was the most important feature, followed by heart rate. Regarding prediction, systolic blood pressure and respiratory rate were found to be the most important.

Aushev et al. [70] have worked on feature selection methods for prediction of both septic and cardiogenic shock mortality in the ICU at different time-steps during the hospital stay, and found that, as shock progresses, different parameters are more helpful for the prediction task at different moments. They analysed different sets of the clinical

parameters resulting from the feature selections and trained several ML models with each one. The selection methods were based on univariate selection (a filter approach), recursive feature elimination (a wrapper-type algorithm), and RF feature selection. It was also studied if there were any relations between the chosen features. The selected features for the earlier time frame (less than 16 hours after measurements) included SOFA, respiratory rate and GCS. For the second time frame (either less than 16 or 48 hours), SOFA and GCS were once again among the important features. Taking all the measurements into account and ignoring the different time steps, the most important variables included systolic blood pressure, respiratory rate and heart rate.

Chicco and Oneto [71] have studied the variable importance for three prediction tasks, one of them also being septic shock. The other two were for SOFA score and survival predictions. The EHRs gathered clinical information from 364 patients from the ICU and considered 29 clinical features. Among the used models (RF, DTs, Logistic Regression, Support Vector Machines, NNs and more), RF and the NN outperformed the other algorithms, with the first one being the top-performing model. Despite this, no model obtained good precision scores, as it was difficult to correctly distinguish patients without septic shock. From this classification, through RF feature selection, the resulting feature ranking showed creatinine, GCS, mean blood pressure and PCT as the most important features.

Conclusion

With the mentioned works, it is possible to see that achieving a sepsis diagnosis is a challenge that can be approached in different ways, namely from a prediction and a detection standpoint, with different algorithms and parameters. Likewise, understanding the most important features for the syndrome's detection has also been done with different methods. The mentioned results of performance show values of AUROC that are generally between 85% and 95%, which emphasizes the impact that these models can have in a clinical environment.

For this project, the early diagnosis of sepsis can be seen as a blend between a detection and prediction task, since data from the earlier care settings is used to detect the syndrome before the patients are admitted in the ICU and give a prediction of sepsis, despite not explicitly following the timely evolution of the clinical parameters, as will be explained in Chapter 4.3.2.

3.3 Deep Learning

Some examples of work conducted with ML have been described. As previously mentioned, the specific field of DL, within ML, has been increasingly used for research related to EHR analysis in recent years [72], and due to the promising results for early sepsis diagnosis, an overview of some studies is presented.

Lin et al. [60] describe a DNN structure that uses LSTM and adds two components to improve the performance of the model, to predict septic shock: 1) a CNN that precedes the LSTM to extract specific dynamic characteristics (such as vital signs, treatments, and laboratory results); 2) a fully connected neuronal network to introduce static features (like age, gender, and comorbidities) into the LSTM network, which then has the function of dealing with dynamic information. 3-fold cross-validation was used in all models. This framework obtained good results taking into account 1 869 hospital admissions with septic shock, most notably with data from the first three hours of the EHRs with an AUROC of 92%.

Lauritsen et al. [61] had the same approach in creating a model for sepsis detection, composed of a CNN followed by an LSTM, based on data from 3 126 patients, and compared it with GBDT and a FFNN. For the GBDT, the selection and extraction of characteristics followed a conventional model, considering six features of the EHRs (including, for example, systolic blood pressure and heart rate), from which five more features were engineered. The second control model consists of a FFNN with two hidden layers, in which features were generated from retrospective time windows with different hour intervals. This resulted in a total of 30 000 features. The CNN+LSTM model achieved AUROC values between 85.6% (for three hours before the onset of sepsis) to 75.6% (to 24 hours before the onset), outperforming the GBDT.

Zhang et al. [62] have used an LSTM-based architecture with pooling, to propose an interpretable model that captures time information and predicts sepsis in the ED department, taking into account more than 52 800 sepsis patients. They have used a data split ratio of 80% for the train group, and 10% for both validation and test groups, having achieved an average AUROC of 89.2%, as well as surpassing four early-warning scores (including qSOFA, NEWS and SIRS), and three baseline ML models (including RF and GBDT). They have also evaluated the model's performance when specific subpopulations are considered for sepsis prediction, based on gender, race, and different age gaps. The model achieved AUROC values above 90% in all cases.

Conclusion

Despite this project's focus on ML and data preprocessing, the realm of DL has been showing consistent results when it comes to early sepsis detection and prediction, which is why they are mentioned in this work. Nevertheless, as mentioned in Chapter 2.2.4, there are drawbacks to the use of such architectures. These are ideal to deal with large quantities of data, although they tend to display a higher computational complexity and concomitant longer training times.

Regardless of these facts, experiments have been made with a CNN in the context of this project, but the results were not satisfactory due to the rapid overfitting of the network. The amount of data available for early detection in this work is limited, and small when compared to the mentioned studies, which could be one of the causes for the

poor performance. In Chapter 6, more details regarding this attempt will be presented.

Chapter 4

Methodology

This chapter starts by detailing the software used, not only to develop the classification models but also to implement the best models in the sepsis detecting platform, previously developed by Miguel [25]. This is then followed by an explanation of the methodology used to develop said models, including a description of the used datasets, as well as the data preprocessing methods.

4.1 Software

The classification models were developed with Python, version 3.8.2, including all the data preprocessing, training, and evaluation. All the necessary libraries were imported, with the most important ones being: **Sci-kit learn** [73], a ML dedicated library; **Pandas** [74], a library that focuses on data structures and their analysis and manipulation; **Numpy** [75], a library that works with multidimensional arrays and matrices, as well as high-level mathematical functions to operate on these.

The implementation of the best performing models in the sepsis detecting platform was done with MATLAB, version R2021b, the same software used by Miguel [25] to develop the platform.

4.2 Datasets

As mentioned in Chapter 1.2, the main goal of this project is to obtain an early diagnosis of sepsis. To achieve this, the detection is seen as a supervised classification task. This way, clinical data from two databases was used: MIMIC-III and CHULC.

4.2.1 Medical Information Mart of Intensive Care III database

Firstly, to initiate the construction of the classification models, clinical data from the database Medical Information Mart of Intensive Care III (MIMIC-III) was used. It is an open-source, anonymized, large database with information regarding 53 423 hospital admissions in the ICU of the Beth Israel Deaconess Medical Center, between 2001 and 2012. It includes many different parameters and data information, such as vital signs, laboratory measurements, diagnostic codes, hospital length of stay, and more [76].

Since these patients are already in the ICU, they will most likely have a more advanced stage of the syndrome and will, therefore, be easier to identify. This way, the results obtained from this database will be used as baseline scores, to compare the results obtained with the second database, which allows for the early classification of patients (as will be described further, in Section 4.2.2). Thus, only some vital parameters and laboratory measurements were used from these EHRs, as well as information regarding the age and gender of the admitted patients.

The data extraction was done using the software pgAdmin4 [77], which uses Structured Query Language (SQL) programming, as described by Miguel [25]. Clinical data of hospital admissions of adult patients (aged at least 18 years) was acquired, both for the sepsis population and control population. These two populations were distinguished by the diagnostic IDs of sepsis.

The extracted parameters were:

- Vital signs: Heart Rate (HR), Respiratory Rate (RR), Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Mean Blood Pressure (MBP), Body temperature (temp) and Blood oxygen saturation (SaO_2);
- Laboratory analysis work: Bilirubin (bilir), Creatinine (creat), Fraction of inspired oxygen (FiO_2), Partial pressure of oxygen (PaO_2), and Platelet count (plat);
- Patient monitoring: Glasgow Coma Scale (GCS) score;

This extraction resulted in 53 326 hospital admissions, from which **5 974** had a sepsis diagnosis. Each admission was treated as a separate patient, with its unique ID. Each clinical parameter was saved in a separate file. From this, 13 .csv files were created for each of the two populations, sepsis and control, resulting in a total of 26 files. The age and gender of the patients were saved in a separate file, with the corresponding IDs.

4.2.2 Centro Hospitalar Universitário de Lisboa Central database

The second clinical database used in this project is from Centro Hospitalar Universitário de Lisboa Central (CHULC), with whom this dissertation project was done in partnership, containing clinical information of about 9 000 patients admitted at the intermediate care and infirmary settings of CHULC, in the year 2018, from January 1st to December 31st.

This database has information regarding hospital circuit, diagnoses, infirmary records, laboratory and pharmacy data. The goal of using this database consists of developing and evaluating the performance of the algorithms regarding the **early diagnosis**, since being able to classify septic patients in the mentioned clinical settings would help with initiating the treatment before the patient is admitted to the ICU.

The following clinical information was extracted:

- Vital signs: Heart Rate (HR), Systolic Blood Pressure (SBP), Mean Blood Pressure (MBP), Diastolic Blood Pressure (DBP), Body temperature (temp) and Blood oxygen saturation (SaO_2);
- Laboratory analysis work: Bilirubin (bilir), Creatinine (creat), Platelet count (plat), Hemoglobin (hemo), Leukocytes (leuko), C-reactive protein (CRP) and Procalcitonin (PCT);
- Infirmary notes regarding speech, orientation, breathing and dyspnoea states of patients;

From the vital sign parameters and laboratory analysis work, 13 .csv files were created for the entire population within the database. Besides the mentioned clinical data, information regarding the diagnostic and antibiotic treatment were also obtained, to allow for the labelling of the two populations, sepsis and control, within the dataset. This process is described in Chapter 4.3.1.

Some parameters that were extracted from MIMIC-III weren't available because they are not parameters that are frequently assessed in the earlier care settings, like GCS and RR. To overcome this obstacle, the infirmary notes were used in an attempt to achieve an equivalent calculation for the missing parameters. Thus, from these, resulted two more .csv files. The age and gender of the patients were, once again, saved in a separate file, with the corresponding IDs.

4.3 Preprocessing

As seen previously, the first step to develop ML classification models consists in preprocessing the available data, arranging it in such a way that allows the models to learn from it.

The process described in this section was done for the extracted clinical information from both databases, MIMIC-III and CHULC. Slight differences in the implementation methods will be mentioned throughout this Chapter, but the approach was generally the same for both. The most significant difference was the fact that the data extracted from MIMIC-III was already anonymized and labelled according to the sepsis diagnosis ID (i.e. the septic patients were already separated from the control patients at the time of extraction). The data from CHULC, on the other hand, was not. Therefore, this dataset

was pseudo-anonymized, and the first subsection regarding the data extraction and labelling (Chapter 4.3.1) only concerns this specific dataset. The remaining subchapters refer to both datasets, which means that, in terms of feature engineering, scaling, data imputation, and dimensionality reduction, both were processed similarly.

4.3.1 Data Extraction and Labelling

As mentioned before, EHRs from the CHULC database were not labelled. Three different approaches were used to try and gather as many sepsis episodes as possible. Having the maximum amount of episodes, and therefore clinical data points, is important since the more data the models are trained with, the higher the chances of developing an accurate and generalizable model.

The approaches used to label the data were then to extract the following types of clinical episodes:

- Episodes with a direct diagnosis of some type of sepsis and/or condition that led to sepsis;
- Episodes in which the patient was, at some point, under antibiotic treatment for a sepsis-related infection;
- Episodes with a diagnosis for specific types of infection were analysed: if said patient had such a diagnosis and eventually was transferred to the ICU (information that was present in the database), then it was safe to assume that the probability of that patient having had sepsis is high, despite not having the diagnosis. The selection of the infection diagnostics was done by medical professionals from CHULC.

This process resulted in EHRs of **323 episodes of sepsis**, as well as 8 687 control episodes. Once again, as done for the MIMIC-III dataset, each episode was treated as a unique patient with its corresponding ID, to maximize the amount of data with which to train the models.

4.3.2 Feature Engineering

From the data extraction processes resulted several files, one for each clinical parameter with the patient's measurements recorded throughout the length of stay at the hospital, as mentioned previously. These were heterogeneous and complex, as some parameters were sometimes periodically measured, while others had missing values and were incomplete.

The feature engineering process, which was almost identical for both datasets, consisted of the following steps, which are also represented in Figure 4.1:

1. For each parameter file, the following IDs were excluded:
    - IDs with more than 30% of null or missing values;

- IDs with an age calculation higher than 130 years old (which only existed in the MIMIC-III database).

2. The remaining were grouped by ID, allowing for the calculation of three different features, per clinical parameter: **mode**, **mean** and **variance** values. For example, as shown in Figure 4.1 in the upper right corner, the file with data regarding body temperature has the mode, mean, and variance values of body temperature for each ID.

3. The files were then merged to create the respective population dataset (sepsis population and control population), also according to the ID, resulting in an isolated dataset for each, some features having missing values since not all parameters were measured for every patient. The age and gender of each patient were also added as features.

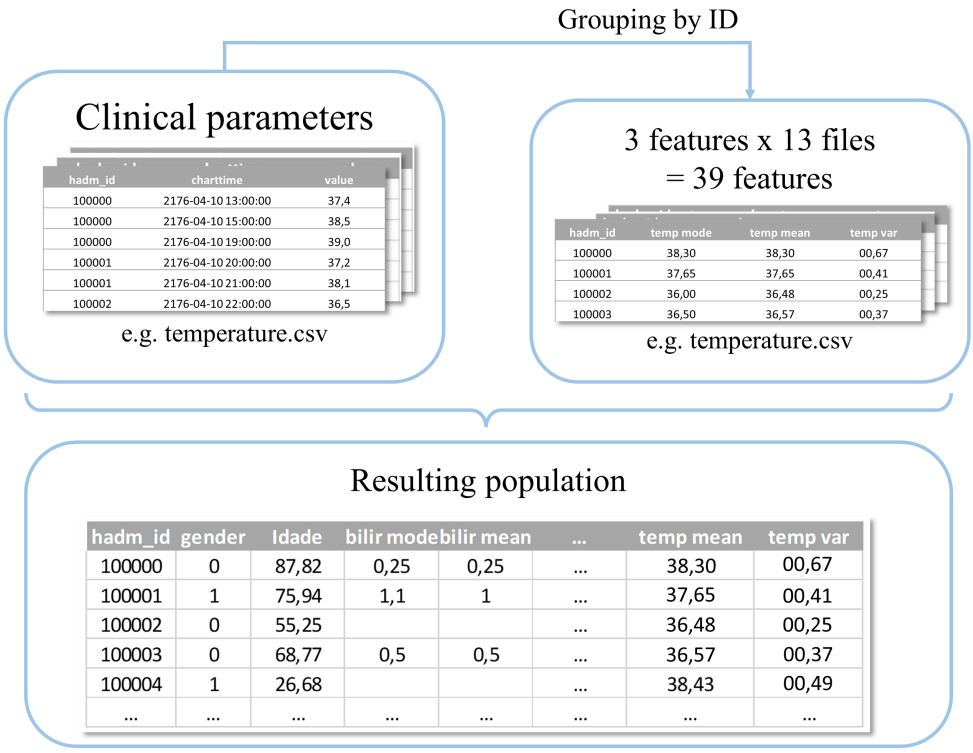


Figure 4.1: Flow chart of the feature engineering process.

Regarding MIMIC-III, this resulted in a dataset with **41 features** for each population, while the dataset from CHULC had **47 features**. For the sake of simplicity, the resulting datasets will be referred to MIMIC-III dataset and CHULC dataset, from this point forward. Figures 4.2 and 4.3, which will be explained further in Chapter 4.3.3, contain the list of features in the x axis, for each dataset respectively. Table 4.1 shows the dimension

of each dataset in regard to the number of patients with and without sepsis, that resulted from this process. It is already possible to see how imbalanced both datasets are, having a substantial difference between the control population and the septic population. This issue will be addressed later, in Chapter 4.3.4.

Table 4.1: Number of patients for each class, sepsis and control, in each dataset, MIMIC-III and CHULC.

Dataset	Sepsis patients	Control patients	Total patients
MIMIC-III	4 207 (11.53%)	32 288 (88.47%)	36 495
CHULC	323 (3.58%)	8 687 (96.42%)	9 010

4.3.3 Handling Missing Data

Regarding the MIMIC-III dataset, Figure 4.2 shows the percentage of missing values for each calculated feature. It is visible that the subset of features with higher percentage of missing values is roughly the same for both classes, with the following parameters being the most incomplete:

- FiO_2;
- bilir, for the control population in particular;
- GCS, for both populations, but specially for the septic population.

Besides these three groups of features, features related to plat and temp also have almost half of the total data points missing.

Regarding the CHULC dataset, the corresponding information is shown in Figure 4.3. For this dataset, the most incomplete features are GCS, PCT and RR.

Despite the effort to obtain values for GCS and RR from the infirmary assessments, the available data from these was still scarce. Because of this, the features were, as it is possible to see, extremely incomplete.

Comparing both Figures 4.2 and 4.3, it is possible to conclude that the CHULC dataset is more uniform than the MIMIC-III dataset. The second one, not only has more features with a substantial amount of missing values, but it also has varying degrees of missingness. The CHULC dataset, on the other hand, has the three mentioned groups of features which are very incomplete, with generally more than 80% of their values missing. Despite this, every other group of parameters has a missingness percentage lower than 20%. In this sense, it is possible to say that the CHULC dataset is more regular and less incomplete than the MIMIC-III dataset.

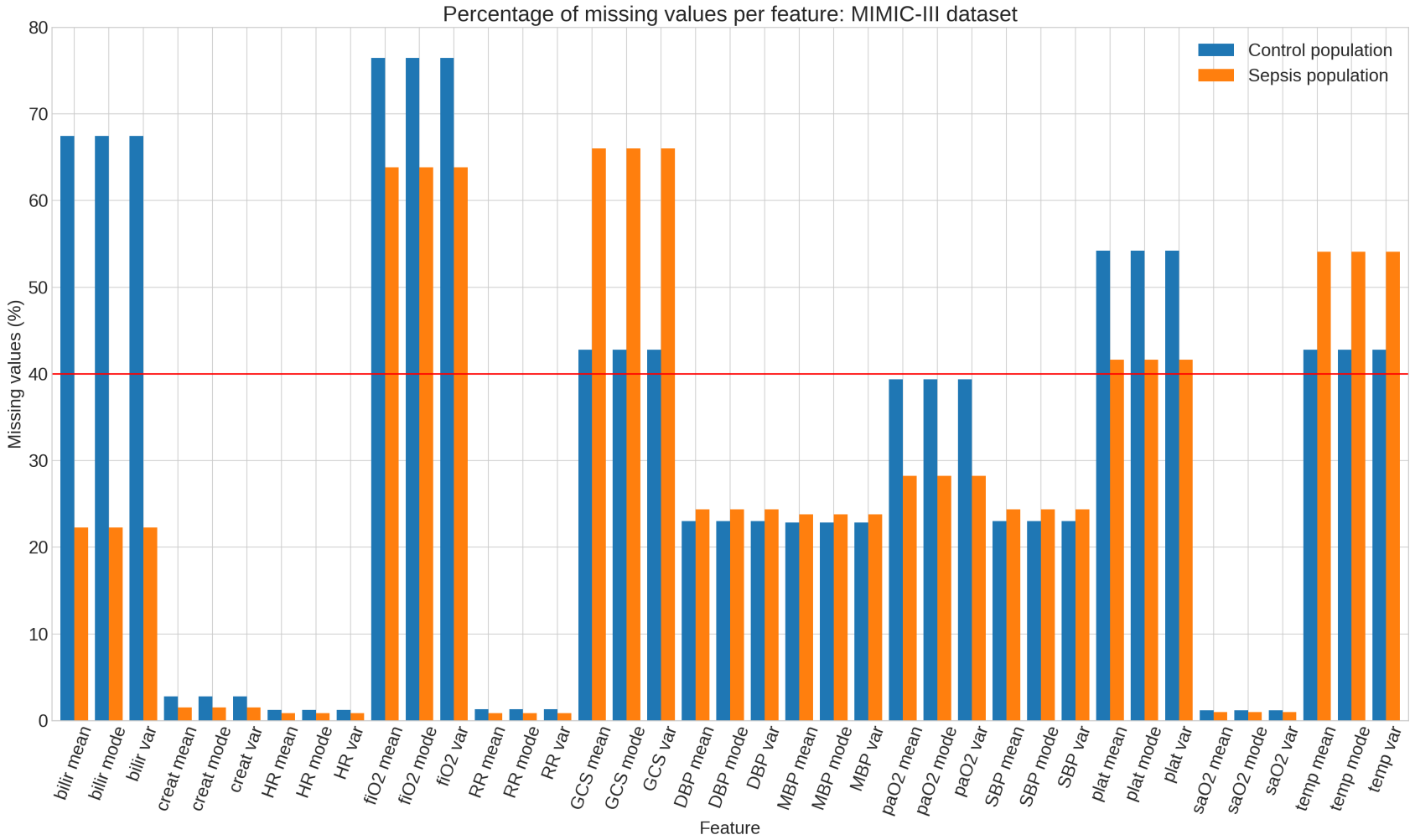


Figure 4.2: Percentage of missing values per feature, for the MIMIC-III database. The features with most missing values are roughly the same for both populations. *Note that age and gender were considered as features but these did not have any missing values and, therefore, are not represented in the plot.*

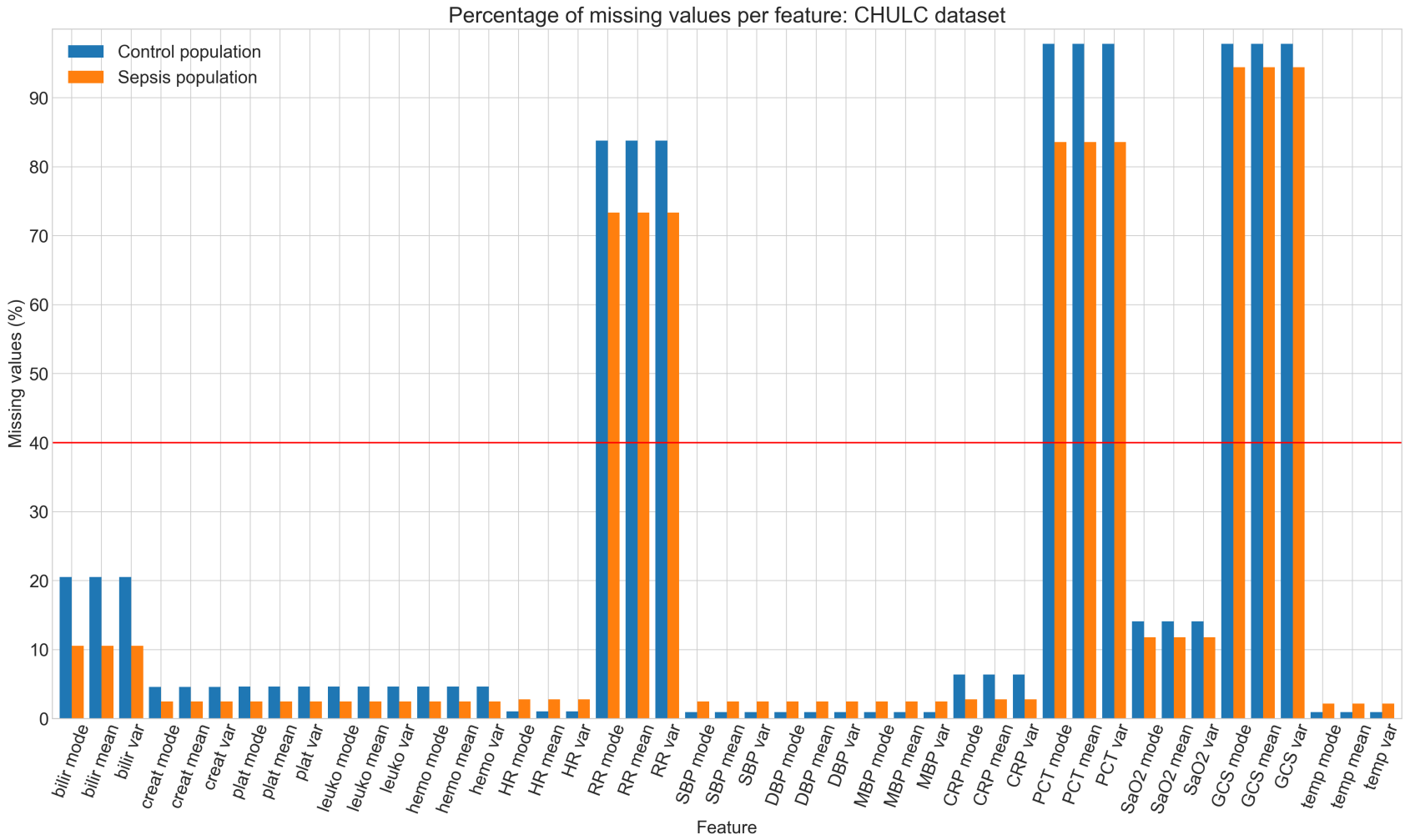


Figure 4.3: Percentage of missing values per feature, for the CHULC database. The features with most missing values are roughly the same for both populations. *Note that age and gender were considered as features but these did not have any missing values and, therefore, are not represented in the plot.*

To deal with these missing values, a data imputation method based on **K-means clustering**, the unsupervised ML model that was introduced in Chapter 2.2.2, was used. This method was used once for each population (septic and control populations) from each dataset. It consisted of training a clustering model and using it to assign each patient of the population to a cluster, to compute the mode value of each cluster, and use these to impute the missing values of each feature. The reason why the mode value was chosen, instead of the mean value, for example, is because the feature distributions did not always correspond to a normal distribution. Thus, in order to impute values that were the most representative of each feature as possible, the mode value was chosen as opposed to other values.

The training of each K-means clustering model needs to be performed with a complete dataset, i.e., without missing values. Regarding the MIMIC-III dataset, due to its' irregularity and varying degrees of missingness, in an attempt to obtain these complete datasets, features with more than 40% of missing values were temporarily excluded. This threshold is visible in Figure 4.2, being represented by the red line. Note that, besides the features represented in the plot, two more are present during all the analysis, that did not have any missing values and, therefore, are not represented in the plot: age and gender. This resulted in a subset of features that were less incomplete, with the sepsis clustering model being trained with 29 features, and the control model with 26 features. Patients that continued to have missing values after this feature selection were excluded, in order to get the most complete subset of patients, for the training. Had the feature selection not been performed, the resulting number of patients to train the models, i.e. that did not have missing values, would have been 8 sepsis patients and 11 control patients. This goes to show how heterogenous the dataset is, as well as how important was the temporary feature selection. The number of patients used to train each of the K-means clustering model was:

- 2 321 patients for the sepsis population, out of 4 207 (55.17%);
- 17 699 patients for the control population, out of 32 288 (54.82%).

This means that around 50% of the patients present in each of the population datasets were used to train the clustering models.

Since the CHULC dataset is more uniform, the training of the clustering models did not require features to be temporarily excluded from the dataset. Instead, the most incomplete ones were permanently removed, resulting in a total of **38 features**. The training of the clustering models only required the exclusion of patients with missing values. This way, the number of patients used to train each of the K-means clustering models, for this dataset, was:

- 257 patients for the septic population, out of 323 (79.57%);
- 5 948 patients for the control population, out of 8 687 (68.47%).

Elbow Method

To determine the optimal number of centroids, k, the elbow method was used (which was also introduced in Chapter 2.2.2). Figure 4.4 represents the obtained elbow plot for the sepsis population of the MIMIC-III dataset. By analysing the inflection point, k was defined as 15. All of the obtained elbow plots had a similar shape, not only for the control population of this dataset, but also both populations of the CHULC dataset (see Figures A.1, A.2 and A.3). For this reason, the same value was used in all situations, which means that each patient, in each of the respective population datasets, is assigned to one of 15 clusters. Table A.1 shows the size of the obtained clusters for each dataset. Finally, the mode value of each cluster was calculated, for each feature, and was used to impute the missing values in patients that were lacking the specific parameter.

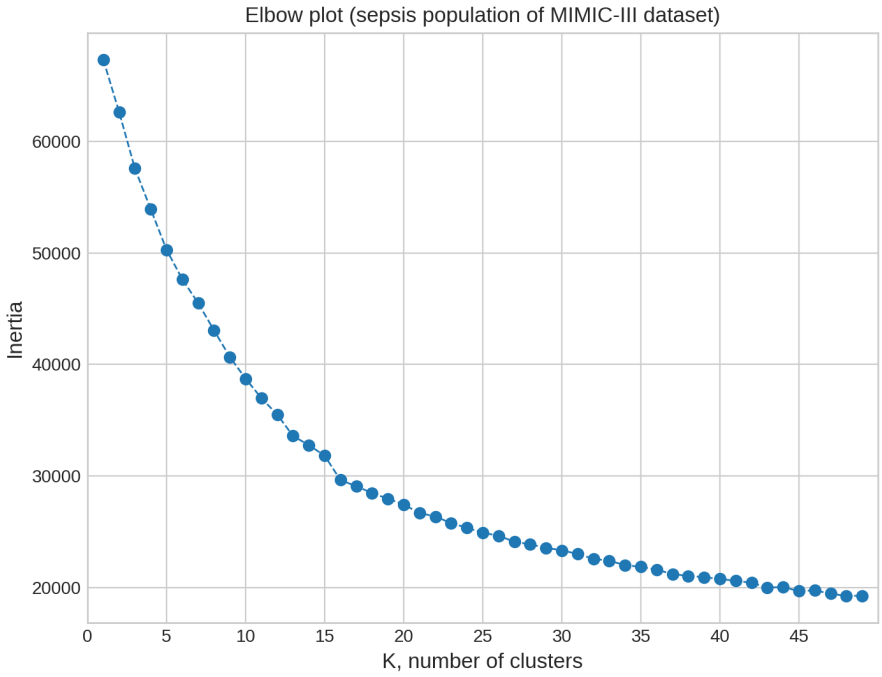


Figure 4.4: Elbow method for determining the optimal number of cluster centroids, for the sepsis population of the MIMIC-III dataset.

From this process resulted a complete feature subspace for the CHULC dataset, without any missing values. This means that the initial percentage of missing values also corresponds to the percentage of imputed data. For the MIMIC-III dataset, on the other hand, a residual number of patients (7 sepsis patients and 102 control patients) still had missing values for the mean, mode, and variance values of FiO_2 and plat. This is a result of the heterogeneity of the dataset, since these 6 features belonged to the group of the most incomplete parameters in the feature subspace, and were, therefore, not used for

the training of the K-means clustering models. This resulted in certain clusters being characterized precisely by the lack of these values. Since these patients represent less than 0.5% of both populations, they were simply excluded from the dataset.

To conclude, the features were standardized by removing the mean and scaling to unit variance. This is an important step, as seen previously in Chapter 2.2.3, to ensure that our models are influenced equally by all features, regardless of their value range.

4.3.4 Class Imbalance

As seen previously, both the MIMIC-III dataset, as well as the CHULC dataset, suffered from high class imbalance, which means that one population (in both cases, the control population) had a significantly higher number of patients than the other, as was to be expected. If the ML models were trained with such a proportion of patients, the models could be biased in their classifications. To overcome this and to balance both datasets, the control populations that resulted from the feature engineering and data imputation processes were randomly subsampled to around the respective number of sepsis patients in each dataset. This means that the fraction of subsampling was 15% for the MIMIC-III dataset and 4% for the CHULC dataset. The total number of sepsis and control patients of each dataset is represented in Table 4.2:

Table 4.2: Final number of patients, for each class and dataset, after random subsampling of the control population.

Dataset	Sepsis patients	Control patients	Total patients
MIMIC-III	4 200	4 828	9 028
CHULC	323	347	670

4.3.5 Dimensionality Reduction

After handling the missing data, each dataset had around 40 features or more. In order to study the impact of training the models with vital signs that are quick to measure, the feature subspaces of the two datasets were reduced to features with less than 50% of imputed data and features that correspond to vital signs, such as HR, MBP, etc. This not only reduces the complexity of the models, helping with generalization but is also extremely relevant when trying to achieve an early diagnosis of sepsis. If the model has a relatively small number of features that can be promptly assessed in an early care setting, it can quickly alert responsible clinicians and, in this sense, help achieve an even quicker diagnosis.

For the MIMIC-III dataset, the kept features were: gender, age, the mean, mode and variance values of HR, RR, DBP, MBP, SBP, SaO_2 and temp. The last group of features (temp) were considered, despite having more than 50% of imputed values, because its monitoring is easy.

For the CHULC dataset, the considered features were the same, excluding the mean, mode, and variance values of RR, since these had already been excluded during the preprocessing stages, due to the high percentage of missing values.

Thus, from the preprocessing approach described throughout this Chapter, four datasets resulted:

1. A *complete MIMIC-III* dataset, with 41 features (**training 1**);
2. A *reduced MIMIC-III* dataset, with 23 features (**training 2**);
3. A *complete CHULC* dataset, with 47 features (**training 3**);
4. A *reduced CHULC* dataset, with 20 features (**training 4**).

Each one of these datasets was then used for training and evaluation of the ML models, in an attempt to:

1. Compare the obtained classification capabilities in the ICU (referring to the MIMIC-III datasets, which was considered to be the *baseline* performance due to the higher chances of the syndrome having progressed to more severe stages) with their *early* classification in early care settings (related to the CHULC datasets, where its diagnosis is harder to achieve because of the probably less pronounced symptoms of the syndrome);
2. Determine which of these classification tasks is possible to be performed with few clinical parameters that are easy to assess.

4.4 Classification models

Six ML classification models were trained, five being based on DTs and ensembles of these (namely RF, GBDT, DT, Adaboost and XGBoost), and the sixth being KNN (which have all been introduced in Chapter 2.2).

4.4.1 Dataset Split

From the four different datasets that resulted from the data preprocessing, four different training tasks were performed and evaluated in terms of performance. Before the training tasks began, the datasets were split into the **training** and **testing** groups, with a ratio of 90% for the training group and 10% for the testing group. This split ratio is important, in particular for the CHULC dataset that has a smaller dimension, to maximize the amount of data with which the classifiers are trained. The patient distribution of each group was the same across the datasets from the same database, i.e., the training and testing groups from the CHULC datasets, reduced and complete, had the same patients within each of the groups (the same applies to the MIMIC-III datasets). This was done to avoid possible

variations in the obtained scores due to the different class distribution, making the results more stable and comparable among training tasks.

During each training, 10-fold cross-validation was used to avoid overfitting, meaning that each training group was also split into ten groups, each one being used as a validation dataset once and as a training dataset nine times, as described in Chapter 2.2.5.

Table 4.3 shows the number of patients from each class (sepsis and control) according to this data split. The difference between the overall dimension of each dataset, MIMIC-III and CHULC, is clear, and thus, one might anticipate differences in the performance of the models, from that fact alone.

Table 4.3: Number of patients from each class, according to the mentioned dataset spit ratio.

	MIMIC III dataset		**CHULC dataset**	
Population	**Training (90%)**	**Testing (10%)**	**Training (90%)**	**Testing (10%)**
Sepsis	3780	420	284	39
Control	4347	481	319	28
Total	8125	903	603	67

4.4.2 Parameter Tuning

Each classifier's parameters were first chosen during training 1, with the MIMIC-III dataset. Since the obtained results were already quite good, for both training 1 and 2, no extensive optimization task was applied to training 2. Regarding the trainings related to the CHULC dataset, a Randomized Search Cross-Validation [78] task was performed for both trainings 3 and 4, in an attempt to optimize each classifier's parameters, taking the new dataset into account. This process works by randomly choosing the values for each parameter of each classifier, from pre-determined intervals of values that are provided, for several iterations. For each iteration, the classifiers are trained with k-fold cross validation. In the end, the combination of parameters that provided the best average accuracy is chosen. The selected number of iterations number was 60, and 10-fold cross validation was used. No improvements on the evaluation metrics were seen after performing these optimizations, when comparing with the use of the original parameters chosen during training tasks 1 and 2, and for this reason, the chosen parameters were the same across all training tasks. These include the following specifications:

- The RF has a forest of 100 trees, each tree is limited to a maximum depth of 45 nodes;
- The GBDT also has 100 estimators, and a learning rate of 0.1, as well as a maximum depth of 3 nodes in each tree;
- The XGBoost has a higher learning rate of 0.5, but also has 100 trees in its forest and a maximum depth of 4 nodes, using L2 regularization on the weights of the trees;

- The Adaboost has 75 trees and a learning rate of 1, each tree being limited to a maximum depth of 1;
- The DT has also been limited to a maximum depth of 45 nodes for the tree, but contrarily to the previous models, its impurity criteria is not Gini, but entropy. The minimum number of data points to split a node is 10, as well as the minimum number of samples to create a leaf;
- The KNN uses 10 neighbours, and the weight function that measures the influence of each neighbour gives a higher weight to closer neighbours. The distance between data points that is taken into account is the Manhattan distance.

The next Chapter presents the results for the mentioned training tasks, as well as the analysis that was done to determine which features were most important for these classification tasks.

Chapter 5

Experimental Results

This chapter analyses the obtained experimental results throughout this work.

It starts by presenting the obtained performance metrics for each training task, as well as the most important features for the early detection cases. In the end, it describes the implementation of the best algorithms in the sepsis detecting platform [25].

5.1 Performance Evaluation

After training each model, with each one of the four datasets, the models' performance was evaluated. The confusion matrices and ROC curves resultant from each evaluation can be seen in Appendix B, when not presented in this Chapter.

Training 1

The performance results of the models for training 1 are presented in Table 5.1.

Table 5.1: Performance results of the models for training 1 (MIMIC-III dataset with the complete feature subset). All the scores are presented in percentage (%).

Model	Train Acc ± SD	Test Acc	AUROC	Sensitivity	Precision
XGBoost	99.58 ± 0.17	99.34	99.98	98.87	99.77
GBDT	98.95 ± 0.36	98.45	99.84	98.19	98.64
RF	98.35 ± 0.50	98.01	99.83	98.42	97.54
Adaboost	97.26 ± 0.46	96.46	98.45	95.26	97.46
DT	96.73 ± 0.78	96.46	99.42	97.29	95.57
KNN	92.46 ± 0.67	91.47	97.91	85.10	97.16

These models are performing extremely well, so much so that one might jump to the conclusion of them being overfit to the training data. But, as seen in Chapter 2.2.5, an overfit model is characterized by giving great prediction results on the training data while

performing badly on the test data. This does not happen, as all the models perform just as well on the testing group and achieve AUROC values of almost 100%.

An explanation for this is the degree of regularization that was forced on the dataset, with the imputation technique that was used on the missing values. The original dataset was very heterogeneous and incomplete, with almost half the features having more than 40% of their data imputed (see Figure 4.2). This not only meant that the clustering models could not have been trained with the complete subset of patients, which alone increases the chances of patients being wrongly assigned to their clusters (and therefore, assuming that different patients belong to the same cluster when, in fact, they do not), but they were also not trained with the complete subset of features, increasing these chances even more. The incorrect assignment of patients to clusters has forced the imputation of missing values that are not as representative of them, leading therefore to uniformization of the dataset, reducing its complexity and making it easier for the models to distinguish the patients. We can then conclude with this training, that the developed data imputation method should not be used in datasets like MIMIC-III, i.e., datasets with, not only with many incomplete features, but also with highly varying degrees of missingness.

Training 2

Regarding training 2, which uses the reduced MIMIC-III dataset, the performance results are given in Table 5.2. The best performing model was the RF classifier. Even though the XGBoost classifier obtained the same training accuracy score and AUROC, the RF surpassed it in terms of testing accuracy, sensitivity, and precision. As mentioned in previous Chapters, more important than achieving good results on the training data, it is crucial that a model performs well on data it has not been presented with, i.e., the testing data. Thus, the best classifier for this training task was the RF classifier.

Table 5.2: Performance results of the models for training 2 (MIMIC-III dataset with the reduced feature subset). All the scores are presented in percentage (%).

Model	Train Acc ± SD	Test Acc	AUROC	Sensitivity	Precision
RF	**83.9** ± 1.0	**82.9**	**92.0**	**81.0**	**83.7**
XGBoost	83.9 ± 1.5	81.5	92.0	79.0	82.6
GBDT	83.2 ± 1.4	80.0	91.1	78.6	80.2
Adaboost	81.4 ± 1.5	79.5	89.6	77.0	80.4
KNN	77.8 ± 1.4	76.5	86.6	68.2	81.0
DT	77.4 ± 1.0	76.4	84.3	77.4	75.2

Figure 5.1 represents the obtained ROC curve for each of the models, where we can see that all of them have curves with a similar morphology. The best performing classifiers have achieved extremely close scores for the AUROC score, which is reflected by their almost overlapping ROC curves. The RF classifier obtained the highest value for sensitivity, i.e., for the TPR, which is visible in the plot by having the curve that maximizes it and is, therefore, closest to the top left corner of the graph.

It should be noted that, despite having a reduced number of features, in particular features that are easily measured and assessed, the models still performed well. This dimensionality reduction has, in a way, compensated the uniformization resultant from the missing data imputation method. With this, it is plausible to say that the clinical parameters resultant from the feature reduction could be used to achieve a faster diagnosis of sepsis in ICU settings.

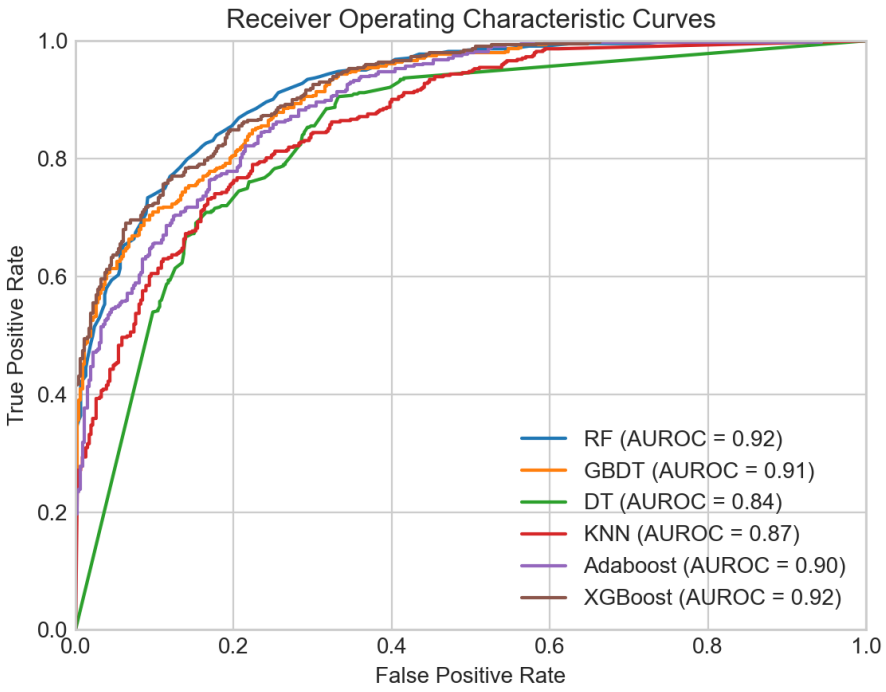


Figure 5.1: ROC curves for the tested models, during training 2.

Training 3

The results for the training with the complete CHULC dataset can be seen in Table 5.3.

It was expected that most scores would be lower since, not only is the dataset considerably smaller, but it is also important to remember that it contains clinical information of patients in intermediate care and infirmary settings, which means that the sepsis patients are most likely in an early stage of the syndrome. This inevitably hinders the classification process. Despite this, the results obtained with this training are not far from the value range of the previous training, with AUROC scores higher than 85% and sensitivity scores reaching almost 77%, for the best two models. Thus, good results have been obtained, which achieve the main goal of this dissertation project: to classify sepsis patients before they are admitted to the ICU, i.e., in earlier care settings, and thus achieving an earlier diagnosis.

Table 5.3: Performance results of the models for training 3 (CHULC dataset with the complete feature subset). All the scores are presented in percentage (%).

Model	Train Acc ± SD	Test Acc	AUROC	Sensitivity	Precision
XGBoost	80.8 ± 5.1	79.1	86.4	76.9	85.7
RF	**80.3** ± 4.1	**77.6**	**87.3**	**76.9**	**83.3**
GBDT	80.8 ± 3.9	77.6	84.8	74.4	85.3
DT	71.1 ± 5.7	73.1	78.3	69.2	81.8
Adaboost	79.9 ± 4.4	68.7	81.1	64.1	78.1
KNN	75.8 ± 3.3	68.7	73.9	59.0	82.1

Despite not having the highest scores for every metric calculated, the best model was concluded to be, once again, the RF classifier. When it comes to getting a sepsis diagnosis, minimizing the FN is of greater priority when compared to minimizing FP, i.e., sensitivity is more important than precision. Despite having the third best precision score, not only does the RF classifier have the same sensitivity as XGBoost, meaning that both minimize FN in the same way, but it also achieved the best AUROC score of 87%, which lead to it being chosen as the best model for this training task. Figure B.2 shows the corresponding ROC curves.

Training 4

With training 3, it was concluded that an early classification of sepsis is possible. Training 4 now assesses if, on top of having data from early care settings, this classification is also possible to achieve with parameters that are easy to measure. Therefore, the results shown in Table 5.4 consider the reduced subset of features for the CHULC dataset.

Table 5.4: Performance results of the models for training 4 (CHULC dataset with the complete feature subset). All the scores are presented in percentage (%).

Model	Train Acc ± SD	Test Acc	AUROC	Sensitivity	Precision
RF	**70.3** ± 6.5	**70.2**	**77.7**	**61.5**	**82.8**
GBDT	70.0 ± 4.8	67.2	79.2	56.4	81.5
XGBoost	67.7 ± 7.4	64.2	75.8	59.0	74.2
Adaboost	68.5 ± 6.8	62.7	75.1	53.9	75.0
KNN	66.0 ± 7.4	62.7	71.4	61.5	70.6
DT	61.5 ± 6.8	59.7	65.5	53.9	70.0

The obtained performance scores are lower than all the other analysed circumstances. This can be explained, once again, by the same two factors mentioned previously (the small dataset and the earlier stage of the syndrome), but this time, the reduced number of features also hinders the classification process and does not offer the best results, as expected. Despite this, the best ML classifier, which was determined to be, once again, the RF classifier, achieved good values for the accuracy scores (both for training and testing), sensitivity, and precision. The AUROC is the only value that is not the highest, but it still is almost 78%. GBDT, the second-best model, obtained an AUROC of 79% but was

not chosen to be the best model due to the lower sensitivity, which is 56%, compared to almost 62% for the RF classifier.

In conclusion, the RF classifier was chosen as the best model for training tasks 3 and 4, and was, therefore, the model that was integrated in the monitoring platform, to achieve an early sepsis diagnosis.

5.2 Feature Importance

In the context of the secondary goal of this project, which is to determine the most important clinical parameters for the early classification tasks, the RF, XGBoost and GBDT classifiers, the best performing models that were trained with the CHULC datasets (training tasks 3 and 4), were selected, and its corresponding feature importance, for each of the training tasks, was determined. For this, the six feature selection techniques that were introduced in Chapter 2.2.3 were used: feature importance by MDI, feature importance by MDA, SFS, SBS, SFFS and SBFS. This means that, for each of the best two models of each dataset, complete and reduced, the six feature selection techniques were performed.

As previously mentioned, MDI feature importance is obtained directly from the training of the models, since they are all based on ensembles of DTs. Regarding the MDA feature importance, it was determined with 30 permutation iterations.

Regarding the wrapper methods, the order with which the features were added/removed from the feature subset was used to determine the feature importance, as explained previously. For the SFFS and SBFS, the respective logic was applied, but the best ranking obtained for each feature was the one taken into account. Taking SFFS as an example, if a feature was added during the 4^{th} iteration, and was later removed, it was still considered to be the 4^{th} most important feature.

The goal with using more than one method, as well as analysing not only the best but also the second-best models, was to assess if the selected features were common among the different techniques, as well as the different models. By using more than one selection method, it was also intended to overcome each method's drawbacks. The order by which the features were ranked in terms of importance was assessed for all approaches and used as a score, and the average value of this score was calculated and presented. The lower the score, the higher the importance of the feature, since it obtained a higher ranking in the mentioned methods. With information regarding the most important features, it was possible to determine which clinical parameters are more relevant for achieving the diagnosis of sepsis. Besides the presented tables that summarize the ranking information, plots resultant from the embedded feature selection methods were also obtained and can be seen in Appendix B.

Training 3

Regarding the complete CHULC dataset, the best performing models were the RF and XGBoost classifiers. The feature importance results for the best model, the RF classifier, are shown in Table 5.5, which shows the rank of the 5 most important clinical parameters.

Table 5.5: Feature importance scores of the 5 most important parameters for RF classifier, with the complete CHULC dataset. The complete version of this table can be seen in Table B.1.

Feature Importance										
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
SaO_2	mean	18	11	33	19	11	25	23.4	8.5	4.6
	mode	34	18	30	27	14	14	27.4	8.6	
	var	5	6	3	3	3	3	4.6	1.3	
creat	mean	15	25	20	10	12	12	18.8	5.8	6.2
	mode	23	22	14	21	23	17	24.0	3.7	
	var	6	1	13	2	6	3	6.2	4.4	
leuko	mean	11	33	8	14	22	21	21.8	9.1	8.6
	mode	12	27	19	20	27	8	22.6	7.7	
	var	3	2	2	32	2	2	8.6	12.2	
CRP	mean	1	20	1	1	1	29	10.6	12.5	11.9
	mode	2	38	36	33	30	1	28.0	17.1	
	var	4	29	18	30	7	5	18.6	11.9	
plat	mean	21	23	29	35	32	17	31.4	6.9	14
	mode	31	34	16	8	25	37	30.2	11.2	
	var	7	17	6	4	32	4	14.0	11.1	

The parameter that achieved the best average ranking is the SaO_2. In fact, its variance value has been chosen as one of the five most important features for almost every selection technique. It was then followed by creat, leuko, CRP and plat. The variance value of creat was also chosen among the most relevant features for most methods, except for the SFS model, which ranked it 13^{th}, right before its mode value. Nonetheless, it obtained the second-best average ranking score. Analysing leuko, its variance value was chosen as 2^{nd} best for most methods, except for SBS, in which it ranked 32^{nd}. This inevitably lowered its average ranking score, but the rest of the results indicate that it is a relevant feature for this model.

Despite having achieved the 4^{th} highest ranking, it is also relevant to analyse CRP, in particular the mean value, since it was the most important feature for every selection method, except for the permutation importance (MDA) and the SBFS model. In Table 5.5, it is possible to see that the feature that was determined as most important by SBFS was actually this parameter's mode value. Not only that, but in terms of the permutation importance, when looking at Figure 5.2, it is possible to see that CRP mode achieved a high absolute score, but negative, meaning that the changing of the feature's values actually helped the classification task. This could have happened by chance, but given the feature's results with the other selection techniques, it is likely that in this case too,

CRP mean is an important feature. In Table 5.5, it is also possible to see that all three features related to CRP are of high importance, when it comes to MDI. Thus, given these facts, one can conclude that CRP might be one of the most important clinical parameters for the RF classifier, for this early classification of sepsis.

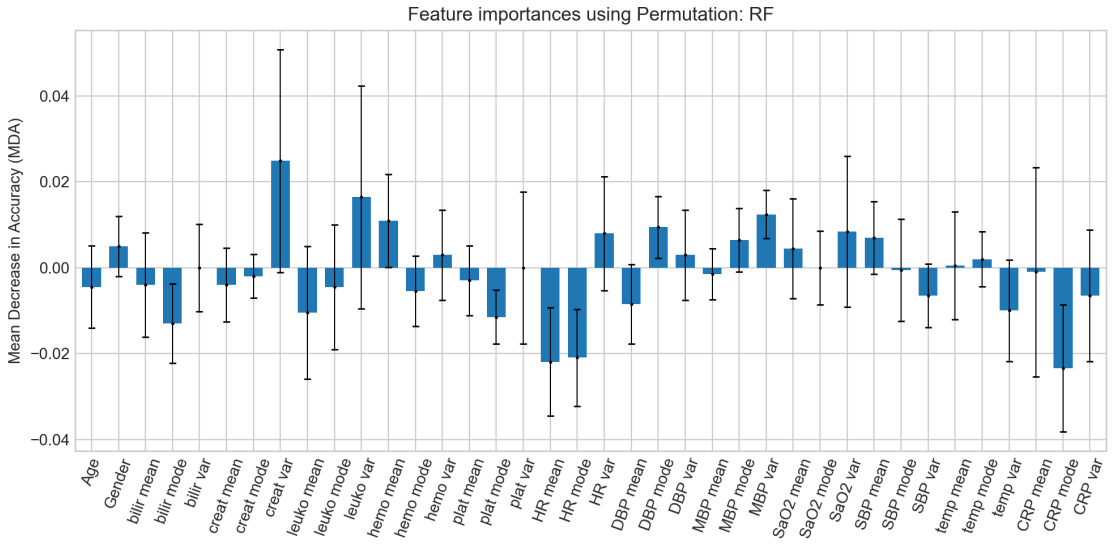


Figure 5.2: Permutation feature importance of the RF classifier, the best performing model, for training 3.

The corresponding results, this time, for the XGBoost model, the second-best model, are presented in Table 5.6. At first glance, it is possible to see that, out of the best five features in terms of the average ranking score for this model, four correspond to results obtained with RF, namely SaO_2, leuko, CRP and creat, which means there is consistency across the best performing models, in terms of feature importance.

Once again, the SaO_2 stands out, having the best ranking score for this dataset, with its variance value being chosen as one of the top four features for all selection tasks. The parameter leuko also was chosen as extremely relevant, namely since its variance value was the second best feature in most cases for this model as well, with the same selector, the SBS model, being the only one that discarded it. Given the similarity in the results, both between selection models and classifiers, the clinical parameter appears to be of considerable importance for the classification task. Regarding CRP, its mean value was, again, discarded by two selection techniques. Despite not being the same techniques as before, these have also determined CRP mode as the most important feature, similar to what happened with RF classifier. This corroborates the hypothesis presented during the previous analysis regarding the importance of CRP. For both models, creat is among the most important five clinical variables, despite the different average ranking scores.

It is interesting to note that, for all 5 clinical parameters, for both classification models, one feature, out of the three belonging to the same variable, often appears to stand out

Table 5.6: Feature importance scores of the 5 most important parameters for the XGBoost classifier, with the complete CHULC dataset. The complete version of this table can be seen in Table B.2.

Feature Importance										
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
SaO_2	mean	8	35	25	11	19	3	20.2	11.9	3.4
	mode	15	13	24	6	21	21	20.0	6.7	
	var	4	1	3	2	3	4	3.4	1.2	
leuko	mean	7	22	18	25	12	11	19.0	7.0	6
	mode	38	34	22	34	27	9	32.8	10.7	
	var	2	3	2	19	2	2	6.0	6.9	
CRP	mean	1	2	28	1	33	1	13.2	15.2	13.2
	mode	33	26	1	29	1	17	21.4	14.1	
	var	19	27	6	7	6	8	14.6	8.8	
bilir	mean	9	11	14	10	21	5	14.0	5.4	14.0
	mode	34	24	4	20	4	8	18.8	12.3	
	var	29	12	5	26	5	24	20.2	10.8	
creat	mean	24	16	9	31	9	15	20.8	8.7	14.6
	mode	10	23	11	22	11	12	17.8	6.0	
	var	3	4	26	3	32	5	14.6	13.2	

as most relevant. This is reflected by the disparity in the average ranking scores of each group of features, in which more often than not, the variance value seems to provide the most information, with the mean and mode values not being as relevant, and often being among the least important.

From all of the analysed features, it is possible to conclude that among the most important clinical parameters for the early classification of sepsis, are SaO_2, leuko, CRP, and creat. Despite having considered patients in the ICU and not in earlier care settings, Chicco and Oneto [71] also studied feature importance with RF selection, based on MDI, MDA and a third method that was not considered in this project. Nonetheless, it is interesting to note that they found creat as the most important parameter, and among those was also PCT, instead of CRP. As mentioned previously, it was not possible to analyse PCT during this project, due to its high percentage of missing values. But having been compared with CRP in the past [6], [32], the similar results regarding the importance of a molecular biomarker between this dissertation and the work of Chicco and Oneto [71] are a good indicator of its importance.

From all the mentioned parameters, SaO_2 is the only parameter that is a vital sign and can, therefore, be promptly assessed. This is not ideal when trying to achieve an early diagnosis of sepsis, in particular, if one wants to extend these early warnings to monitoring outside a clinical setting. Nonetheless, the two proposed goals for this project have been accomplished, which consisted in achieving an early diagnosis of the syndrome and determining the most contributing parameters for this task.

Training 4

The best performing models for the reduced training task, RF and GBDT classifiers were also analysed, to determine the most important vital signs for the earliest sepsis detection task. The feature importance ranking for the RF classifier can be seen in Table 5.7.

Table 5.7: Feature importance scores for the RF classifier, with the reduced CHULC dataset.

		Feature Importance								
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
SaO_2	mean	7	8	19	6	14	10	12.8	5.0	1.6
	mode	18	17	9	18	10	12	16.8	4.1	
	var	1	1	1	1	1	3	1.6	0.8	
temp	mean	5	7	12	3	18	6	10.2	5.5	6.8
	mode	19	19	13	12	5	5	14.6	6.3	
	var	3	4	4	17	4	2	6.8	5.6	
HR	mean	2	2	7	7	15	6	7.8	4.8	7.4
	mode	6	3	20	19	6	1	11.0	8.2	
	var	4	6	3	15	3	6	7.4	4.5	
SBP	mean	14	11	2	2	2	9	8.0	5.4	8.0
	mode	15	10	18	16	16	19	18.8	3.1	
	var	13	14	17	5	18	11	15.6	4.7	
DBP	mean	8	5	8	9	11	5	9.2	2.3	9.2
	mode	16	9	5	13	4	4	10.2	5.1	
	var	9	18	16	4	15	8	14.0	5.5	
age	-	11	12	6	14	6	7	11.2	3.4	11.2
MBP	mean	17	16	10	8	9	15	15.0	3.9	15.0
	mode	10	13	15	10	13	14	15.0	2.1	
	var	12	15	11	11	18	14	16.2	2.7	
gender	-	20	20	14	20	7	15	19.2	5.2	19.2

For this case, the most important parameter is clearly SaO_2, whose variance not only achieved the best average ranking score out of all the analysed cases but also was indicated as the best feature for most techniques, with the lowest rank that it obtained being 3^{rd}, a high rank nonetheless, for the SBFS model.

Regarding the temp, a similar situation to what was seen with CRP mean happened with its variance value, where, even though the feature was not the most important for any selection method, it is within the four most relevant features for all approaches, except for the SBS model. When inspecting Table 5.7, it is possible to see that among the best three features for this particular selector, is the mean value of temp, reflecting the parameter's influence for the task.

Thus far, there were no two features from the same clinical parameter that have obtained close average ranking scores to each other. It seemed, up until this point, that one feature would provide enough information from the respective clinical parameter. Despite this, both the variance and mean values of HR have reached similar scores, and have a similar range of importances throughout the selection methods. This indicates

that, for the HR, both mean and variance values are equally relevant.

The corresponding results for the GBDT model, the second-best classifier, can be seen in Table 5.8. Once again, consistency across models is present, since the three parameters with the best scores correspond to the ones obtained previously.

Table 5.8: Feature importance scores for the GBDT classifier, with the reduced CHULC dataset.

Feature Importance										
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
temp	mean	6	7	20	16	12	6	13.4	5.9	3.4
	mode	13	19	14	14	12	8	16.0	3.6	
	var	3	4	3	1	3	3	3.4	1.0	
SaO_2	mean	7	8	13	8	19	17	14.4	5.1	5
	mode	19	17	12	9	7	15	15.8	4.7	
	var	1	1	1	20	1	1	5.0	7.8	
HR	mean	2	2	8	17	7	4	8.0	5.6	7
	mode	11	3	18	3	10	18	12.6	6.7	
	var	4	6	2	19	2	2	7.0	6.6	
DBP	mean	5	5	11	15	7	6	9.8	4.0	9.8
	mode	8	9	15	18	16	4	14.0	5.5	
	var	18	18	17	10	14	13	18.0	3.2	
age	-	10	12	5	12	5	9	10.6	3.2	10.6
MBP	mean	9	16	7	13	7	5	11.4	4.2	11.4
	mode	14	13	16	5	12	7	13.4	4.3	
	var	12	15	4	4	4	18	11.4	6.3	
gender	-	20	20	6	2	6	11	13.0	7.7	13
SBP	mean	16	11	9	11	11	10	13.6	2.4	13.6
	mode	15	10	19	6	16	14	16.0	4.6	
	var	17	14	10	7	7	15	14.0	4.3	

The most important parameters were, once again, the temp and SaO_2. Regarding the first, its variance value was one of the four most relevant for all selection methods. In terms of the SaO_2, in particular, its variance value as well, it was consistently selected as the most relevant feature across the many methods, except for SBS. Given that SaO_2 also obtained the highest score for the previously mentioned trainings, that consider the complete dataset, it can be concluded that this clinical parameter is of great importance for early sepsis classification, not only considering just vital signs but also in cases where more complex parameters, like laboratory work, are available. The significance of the HR as a clinical parameter, for the reduced dataset, has also been confirmed, as the same two features, the mean and variance values, have achieved similar average ranking scores.

In conclusion, for the early sepsis classification that uses vital sign parameters for the diagnosis, the most important clinical parameters were found to be SaO_2, temp and HR. The last two were also determined as most important by Mitra and Ashraf [40] for sepsis detection, which corroborates the obtained results.

5.3 Sepsis Detecting Platform

As previously mentioned, a sepsis detecting platform has been developed in the context of a previous work by Miguel [25]. In this platform, the focus was on the extraction of the patient's breathing signal from their electrocardiogram, to detect sepsis. In the context of the present project, the classification of sepsis is achieved through the use of ML models, without resorting to any sepsis detection criteria. Thus, some functional changes had to be implemented in the platform.

The RF classifier, the best performing model, was the one chosen for the implementation, since it was the best for both early sepsis classification tasks, i.e., both when all the clinical parameters are available, as well as when only vital sign parameters are available.

The interface, which can be seen in Figure 5.3, as well as the changes to the platform's functioning, were both implemented with App Designer [79], which is MATLAB's development tool for creating apps, designing their layout, and programming their behavior.

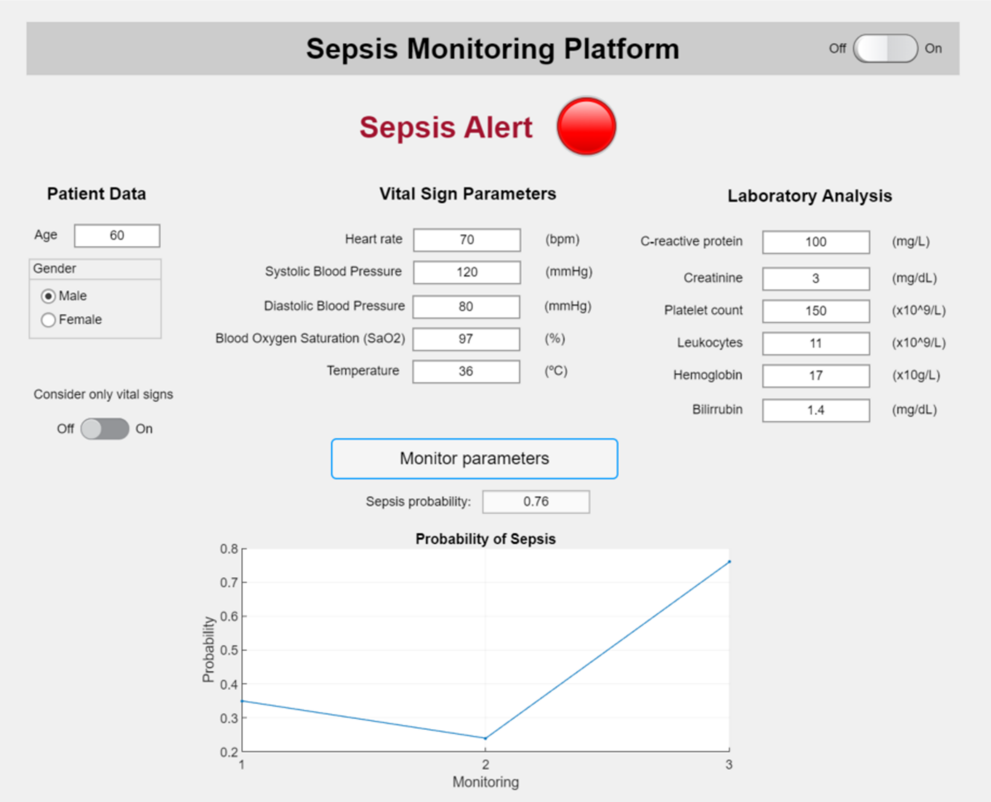


Figure 5.3: Sepsis monitoring platform interface, in which the RF classifier was implemented, in order to monitor the disease's progression and help its alert system.

In the interface, there are fields to enter each of the patient's clinical parameters, as well as personal information. After inserting these, the responsible clinician presses the

button to monitor them, which results in the trained RF model performing the classification task, by determining the probability of the patient having sepsis. If the probability is higher than 50%, and therefore the patient is classified as having sepsis, the red LED lights up, which is what is seen in Figure 5.3. Figure 5.4 shows the interface when the patient is not classified as having sepsis, and therefore, has the alert light off. The parameters can and should be inserted in the platform throughout time since it automatically calculates their mean, mode, and variance values, to use as features for the model to perform the classification. There is also a plot showing the evolution of the probability throughout the monitoring session. If the responsible clinician chooses to consider only vital signs, the laboratory analysis work fields are disabled, and the model only uses these as features for the detection, as represented in Figure B.14. It is possible to turn this option on and off multiple times during the monitoring of the same patient, as long as all enabled fields are provided. This means that, if at the beginning of the monitoring, the blood work analysis is not available, it is possible to initiate the monitoring with only vital signs and add the remaining parameters later.

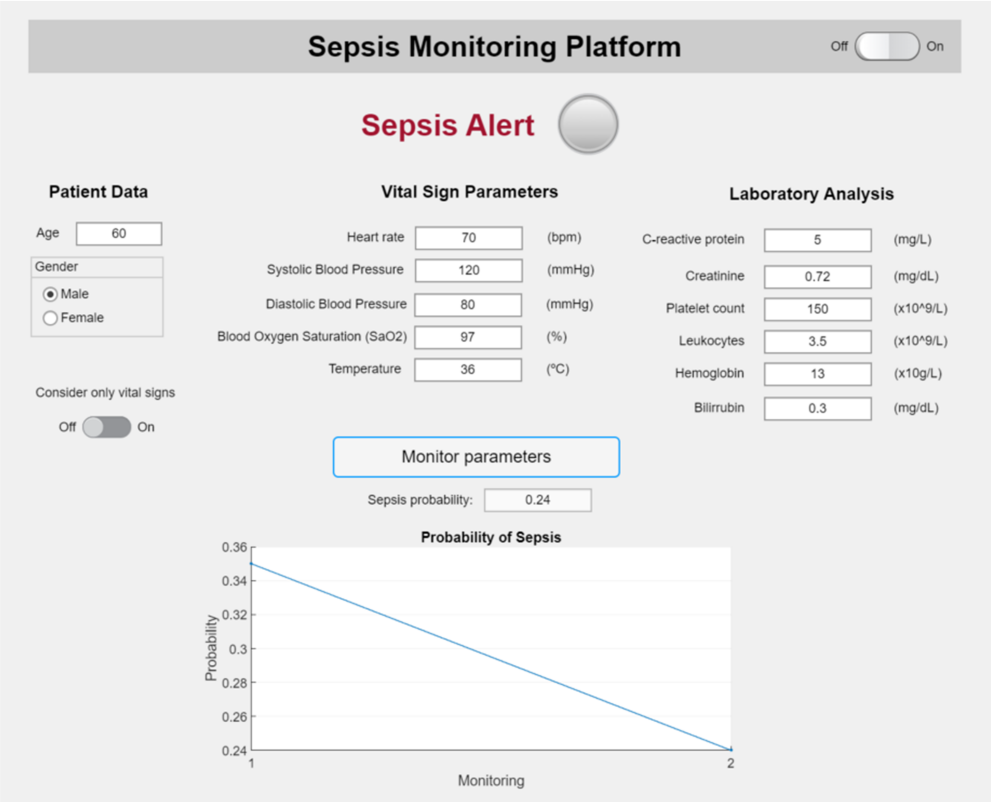


Figure 5.4: Sepsis monitoring platform interface, when the patient is classified as not having sepsis and, therefore, the alert LED light is off.

Chapter 6

Conclusions and Future Work

This Chapter summarizes the work developed throughout this dissertation project, as well as the main achieved contributions and limitations. It ends by analysing the prospects for future work.

6.1 Main Conclusions

Sepsis calls for an urgent diagnosis, since it is the primary cause of death by infection, and early treatment can drastically improve the chances of a good recovery for the patients, as well as minimize potential long-term effects. The available clinical criteria often delay the diagnosis due to either their complexity or lack of sensitivity. Not only that but the syndrome's intricate pathobiology, as well as the lack of a specific set of molecular biomarkers to detect it, make the early detection task difficult to achieve. As technology advances and the amount of clinical data stored in digital records increases, clinicians are faced with large, heterogenous quantities of information from an already complex syndrome, which they must make use of, to make clinical decisions. Thus, the obstacles to achieving an early sepsis diagnose are many, and transversal to many areas of clinical practice.

With this in mind, throughout this dissertation project, two main goals were achieved. The first focused on using AI algorithms to achieve an early diagnosis of sepsis. Two clinical databases were used, one with information regarding patients within ICU settings, which reflect a more severe stage of the disease and was used as baseline to compare results, and another one from the intermediate care and infirmary settings at CHULC, which includes information from patients in early care settings and, therefore, in an earlier stage of the syndrome, which allowed for the early detection task. This involved a meticulous data preprocessing approach, that used clustering algorithms to impute

the missing values for 15 subpopulations within each population, sepsis, and control, for each dataset. The second goal aimed to better understand the classification task at hand, namely which of the 38 clinical features were more relevant and had a greater influence in separating the two populations, regarding the early diagnosis. For this, six feature selection methods were deployed, which included two embedded feature selection techniques and the training of four greedy selection algorithms, to determine each parameter's influence in the respective detection task.

Regarding the clinical data from the ICU, it was concluded that datasets with a level of missingness as high as MIMIC-III's are not ideal candidates for data imputation methods such as the one used in this project, as their heterogeneity will lead to a high level of regularization and a situation resemblant of overfitting, where the dataset used in theory to develop the classification models might not be representative of the real clinical information. This limitation was, nonetheless, combated by eliminating most features with high percentages of imputation, as well as including only vital sign data. The performance of the developed models achieved results that were not only more realistic, but also very good, with AUROC values as high as 92% and 81% for sensitivity.

This obstacle was not present for the CHULC dataset, as the missingness of the data varied much less and was more extreme, which resulted in simply removing the incomplete features and imputing the rest of the values, not showing signs of high uniformization. The early diagnosis was achieved with success, and without using information from any clinical detection criteria, such as SOFA and qSOFA. The best performing models achieved scores of over 87% for AUROC, as well as 77% for sensitivity and 83% for precision. An even earlier detection was achieved by reducing the dimensionality of the feature subspace to as little as 20 features, using only vital sign data that is easy to measure. For this task, a value of almost 78% for AUROC was achieved, as well as close to 62% for sensitivity, with the same precision value as before. In both situations, good results are obtained according to the feature subspace, taking into account, not only the already harder classification due to the less pronounced symptoms of the syndrome but also the smaller dataset, which accounted for less than 350 patients diagnosed with sepsis. Another limitation regarding the use of this dataset relates to the lack of information related to the moment in time when the specific sepsis diagnosis was obtained for each patient, which could have provided more information regarding the evolution of the clinical parameters, possibly allowing the algorithm to consider features within the time domain.

Regarding the most important clinical parameters, laboratory analysis work was shown to have a high influence in the first early detection case, with variables like leukocyte count, levels of creatinine, and C-reactive protein being among the most important clinical parameters. The most relevant, though, was shown to be the blood oxygen saturation level. In fact, when taking only vital sign data into account, this parameter showed, again, to be the most important, and was followed by body temperature and heart rate.

Related to this, it was also seen that, when considering the complete subset of features, generated from each parameter, there is usually one that provides the most information.

This value was shown to be the variance value, in most cases, with the mean and mode values consequently being ranked as less important.

In the end, the Random Forest classifier, the best performing model for both early detection tasks, was implemented in the sepsis monitoring platform. Not only that but the platform was adapted to consider the features studied in this project, taking more information into account for the detection task.

6.2 Future Work

Despite the results achieved with this project, there is still room for future work. Even though the Random Forest algorithm was implemented in the platform's alert system, its functioning is still limited to discrete moments in time, since the values need to be manually inserted in the interface. To overcome this, and to create a continuous monitoring platform, the development of sensors that are specific to the syndrome's clinical parameters is suggested, namely for the vital signs. This way, the patient could be monitored in real-time, and the health care professionals could be alerted earlier, as the need to manually provide the values only applies to parameters related to laboratory analysis.

Even so, two limitations arise from the study presented. Firstly, it has not been tested in a real clinical environment. Secondly, the trained models classify patients based on statistical features (mean, mode and variance) to characterise the population in early care settings. If data is clearly non-stationary, these characteristics may vary significantly over time, and require monitoring over long periods to identify clear dynamical patterns. That may render syndrome detection rather lengthy. It is possible to overcome that issue, though, by using small timeframes for monitoring, allowing the system to better capture abrupt physiological changes in patient's data, and alert the responsible clinicians accordingly.

In addition to those limitations, another element of interest for further work is to develop a method to determine the most relevant statistical feature for each parameter. We observed that, for all classifiers, a given parameter was picked as a highly relevant one. Yet, the feature was not always the same for all classifiers. Hence, it may be useful to check if it is possible to design an approach to detect, for each parameter and classifier, which statistical feature to retain for processing.

Nonetheless, when paired with the obtained results, the monitoring platform shows potential to help in possibly creating an ambulatory-based monitoring device for sepsis, namely through the vital signs. The recent advances in technology regarding smart-watches could also be of interest in the context of this suggestion, namely for the creation of a small, portable device, that includes the mentioned sensors and monitoring windows.

Regarding the most important clinical parameters for early detection of the syndrome, as mentioned before, laboratory analysis work was shown to have a high influence, when these variables are present. It is important to emphasize, once more, that in general,

clinical parameters dependant on laboratory measurements may delay the early diagnosis of sepsis. This represents a limitation, not particularly of the developed model, but of working with medical data. In the case of the mentioned variables, generally, in the early clinical settings considered for this project, these parameters are regularly measured, but they are not very frequent, sometimes having intervals of 2-3 days between the measurements. This represents an obstacle regarding the previous suggestion, related to the ambulatory monitoring device. This is particularly relevant for future works, and as a suggestion for clinical practice in general. All recently developed technology related to the use of AI within clinical context would benefit tremendously from a more structured and cohesive data acquisition system in the clinical facilities. This helps to prevent information loss via missing values, to better follow the evolution of these parameters throughout the length of stay at the hospital, and consequently, improve the system's capabilities.

To further improve the model's performance, two approaches might be relevant for future work. The first is to study the used sepsis population in depth, to see if there is potential to identify and account for features that might be characteristic of a specific early clinical manifestation of the syndrome, similarly to the work done by Ibrahim et al. [69]. Secondly, the use of DL in this context might also help. As previously mentioned, this area has demonstrated remarkable performance when it comes to both sepsis detection and sepsis prediction, taking into account the temporal characteristics of the syndrome. In fact, during this project, experiments have been made with a CNN classifier based on 1D-convolution for the early classification task, namely with the reduced dataset from CHULC (used for training 4), though the results were not satisfactory due to the rapid overfitting of the network. Since DNNs are ideal to work with extremely large quantities of information, this possibly happened due to the reduced size of the dataset, which constitutes a limitation of this project, as it was composed of 670 patients. Nonetheless, in an attempt to overcome this overfitting, early stopping for the training was used (preventing the network to learn too much from the training data) and experiments with dropout layers were done (a dropout layer consists of a layer in which a determined fraction of random neuron weights are excluded in each iteration). Similar AUROC values were achieved, as well as higher sensitivity scores of around 73%. This shows that there is room for work to be done in this context, in particular, if there is access to large databases. In fact, a particular technique within DL is data augmentation. A famous architecture within this realm is Generative Adversarial Networks, also known as GANs [80], which could be of interest to increase the early detection dataset and improve the model's performance and sensitivity.

Bibliography

[1] M. Singer, C. S. Deutschman, C. W. Seymour, *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)", *JAMA*, vol. 315, no. 8, p. 801, Feb. 2016. DOI: 10.1001/jama.2016.0287.

[2] T. Evans, "Diagnosis and management of sepsis", *Clinical Medicine*, vol. 18, no. 2, pp. 146–149, Apr. 2018. DOI: 10.7861/clinmedicine.18-2-146.

[3] J. Arwyn-Jones and A. J. Brent, "Sepsis", *Surgery (Oxford)*, vol. 37, no. 1, pp. 1–8, Jan. 2019. DOI: 10.1016/j.mpsur.2018.11.007.

[4] K. E. Rudd, S. C. Johnson, K. M. Agesa, *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: Analysis for the global burden of disease study", *The Lancet*, vol. 395, no. 10219, pp. 200–211, Jan. 2020. DOI: 10.1016/s0140-6736(19)32989-7.

[5] H. Nishie, "Guidelines for management of severe sepsis and septic shock", *Okayama Igakkai Zasshi (Journal of Okayama Medical Association)*, vol. 125, no. 2, pp. 153–157, 2013. DOI: 10.4044/joma.125.153.

[6] J.-L. Vincent, "The clinical challenge of sepsis identification and monitoring", *PLOS Medicine*, vol. 13, no. 5, e1002022, May 2016. DOI: 10.1371/journal.pmed.1002022.

[7] P. E. Marik, "The demise of early goal-directed therapy for severe sepsis and septic shock", *Acta Anaesthesiologica Scandinavica*, vol. 59, no. 5, pp. 561–567, Feb. 2015. DOI: 10.1111/aas.12479.

[8] G. A. Westphal, Á. Koenig, M. C. Filho, *et al.*, "Reduced mortality after the implementation of a protocol for the early detection of severe sepsis", *Journal of Critical Care*, vol. 26, no. 1, pp. 76–81, Feb. 2011. DOI: 10.1016/j.jcrc.2010.08.001.

[9] R. P. Dellinger, M. M. Levy, A. Rhodes, *et al.*, "Surviving sepsis campaign", *Critical Care Medicine*, vol. 41, no. 2, pp. 580–637, Feb. 2013. DOI: 10.1097/ccm.0b013e31827e83af.

[10] X. Bai, W. Yu, W. Ji, *et al.*, "Early versus delayed administration of norepinephrine in patients with septic shock", *Critical Care*, vol. 18, no. 5, Oct. 2014. DOI: 10.1186/s13054-014-0532-y.

[11] P. Marik and R. Bellomo, "A rational approach to fluid therapy in sepsis", *British Journal of Anaesthesia*, vol. 116, no. 3, pp. 339–349, Mar. 2016. DOI: 10.1093/bja/aev349.

[12] H. C. Prescott and T. J. Iwashyna, "Improving sepsis treatment by embracing diagnostic uncertainty", *Annals of the American Thoracic Society*, vol. 16, no. 4, pp. 426–429, Apr. 2019. DOI: 10.1513/annalsats.201809-646ps. [Online]. Available: https://doi.org/10.1513/annalsats.201809-646ps.

[13] O. A. Usman, A. A. Usman, and M. A. Ward, "Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the emergency department", *The American Journal of Emergency Medicine*, vol. 37, no. 8, pp. 1490–1497, Aug. 2019. DOI: 10.1016/j.ajem.2018.10.058.

[14] Å. Askim, F. Moser, L. T. Gustad, *et al.*, "Poor performance of quick-SOFA (qSOFA) score in predicting severe sepsis and mortality – a prospective study of patients admitted with infection to the emergency department", *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, vol. 25, no. 1, Jun. 2017. DOI: 10.1186/s13049-017-0399-4. [Online]. Available: https://doi.org/10.1186/s13049-017-0399-4.

[15] J. U. Song, C. K. Sin, H. K. Park, S. R. Shim, and J. Lee, "Performance of the quick sequential (sepsis-related) organ failure assessment score as a prognostic tool in infected patients outside the intensive care unit: A systematic review and meta-analysis", *Critical Care*, vol. 22, no. 1, Feb. 2018. DOI: 10.1186/s13054-018-1952-x.

[16] M. Schinkel, K. Paranjape, R. N. Panday, N. Skyttberg, and P. Nanayakkara, "Clinical applications of artificial intelligence in sepsis: A narrative review", *Computers in Biology and Medicine*, vol. 115, p. 103 488, Dec. 2019. DOI: 10.1016/j.compbiomed.2019.103488.

[17] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence", *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019. DOI: 10.1038/s41591-018-0300-7.

[18] F. Khoshnevisan, J. Ivy, M. Capan, R. Arnold, J. Huddleston, and M. Chi, "Recent temporal pattern mining for septic shock early prediction", in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, Jun. 2018. DOI: 10.1109/ichi.2018.00033.

[19] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care", *Nature Medicine*, vol. 24, no. 11, pp. 1716–1720, Oct. 2018. DOI: 10.1038/s41591-018-0213-5.

[20] G. Gutierrez, "Artificial intelligence in the intensive care unit", *Critical Care*, vol. 24, no. 1, Mar. 2020. DOI: 10.1186/s13054-020-2785-y.

[21] M. Komorowski, "Clinical management of sepsis can be improved by artificial intelligence: Yes", *Intensive Care Medicine*, vol. 46, no. 2, pp. 375–377, Dec. 2019. DOI: 10.1007/s00134-019-05898-2.

[22] Q. Mao, M. Jay, J. L. Hoffman, *et al.*, "Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU", *BMJ Open*, vol. 8, no. 1, e017833, Jan. 2018. DOI: 10.1136/bmjopen-2017-017833.

[23] R. J. Delahanty, J. Alvarez, L. M. Flynn, R. L. Sherwin, and S. S. Jones, "Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis", *Annals of Emergency Medicine*, vol. 73, no. 4, pp. 334–344, Apr. 2019. DOI: 10.1016/j.annemergmed.2018.11.036.

[24] K. C. Yuan, L. W. Tsai, K. H. Lee, *et al.*, "The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit", *International Journal of Medical Informatics*, vol. 141, p. 104 176, Sep. 2020. DOI: 10.1016/j.ijmedinf.2020.104176.

[25] J. Miguel, "Desenvolvimento de uma plataforma de deteção precoce da sépsis através da extração do sinal da respiração do ECG [Development of an early sepsis detection platform by extracting the respiration signal from the ECG]", M.S. thesis, Document delievered for the purpose of obtaining the Master Degree in Biomedical Engineering at NOVA School of Science and Technology, 2020.

[26] J. L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: Time for change", *The Lancet*, vol. 381, no. 9868, pp. 774–775, Mar. 2013. DOI: 10.1016/s0140-6736(12)61815-7.

[27] W. H. Organization, *Sepsis*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/sepsis (visited on 11/22/2021).

[28] P. Póvoa, E. Almeida, P. Moreira, *et al.*, "C-reactive protein as an indicator of sepsis", *Intensive Care Medicine*, vol. 24, no. 10, pp. 1052–1056, Oct. 1998. DOI: 10.1007/s001340050715.

[29] C. Santonocito, I. D. Loecker, K. Donadello, *et al.*, "C-reactive protein kinetics after major surgery", *Anesthesia & Analgesia*, vol. 119, no. 3, pp. 624–629, Sep. 2014. DOI: 10.1213/ane.0000000000000263.

[30] A. L. Vijayan, Vanimaya, S. Ravindran, *et al.*, "Procalcitonin: A promising diagnostic marker for sepsis and antibiotic therapy", *Journal of Intensive Care*, vol. 5, no. 1, Aug. 2017. DOI: 10.1186/s40560-017-0246-8.

[31] F. Brunkhorst, K Wegscheider, Z. Forycki, and R Brunkhorst, "Procalcitonin for early diagnosis and differentiation of SIRS, sepsis, severe sepsis and septic shock", *Critical Care*, vol. 3, no. Suppl 1, P095, 1999. DOI: 10.1186/cc469.

[32] A. Luzzani, E. Polati, R. Dorizzi, A. Rungatscher, R. Pavan, and A. Merlini, "Comparison of procalcitonin and c-reactive protein as markers of sepsis", *Critical Care Medicine*, vol. 31, no. 6, pp. 1737–1741, Jun. 2003. DOI: 10.1097/01.ccm.0000063440.19188.ed.

[33] J. L. Vincent, R. Moreno, J. Takala, *et al.*, "The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure", *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, Jul. 1996. DOI: 10.1007/bf01709751.

[34] P. Silva, "Clinical deterioration detection for continuous vital signs monitoring using wearable sensors", M.S. thesis, Document delievered for the purpose of obtaining the Master Degree in Biomedical Engineering at NOVA School of Science and Technology, 2021.

[35] R. C. of Physicians, *National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS*. Updated report of a working party, 2017.

[36] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems", *Journal of Global Health*, vol. 8, no. 2, Oct. 2018. DOI: 10.7189/jogh.08.020303.

[37] X.-D. Zhang, "Machine learning", in *A Matrix Algebra Approach to Artificial Intelligence*, Springer Singapore, 2020, pp. 223–440. DOI: 10.1007/978-981-15-2770-8_6.

[38] A. C. Müller and S. Guido, *Introduction to machine learning with Python: a guide for data scientists*. "O'Reilly Media, Inc.", 2016.

[39] R. A. Taylor, J. R. Pare, A. K. Venkatesh, *et al.*, "Prediction of in-hospital mortality in emergency department patients with sepsis: A local big data-driven, machine learning approach", *Academic Emergency Medicine*, vol. 23, no. 3, A. Jones, Ed., pp. 269–278, Feb. 2016. DOI: 10.1111/acem.12876.

[40] A. Mitra and K. Ashraf, *Sepsis prediction and vital signs ranking in intensive care unit patients*, 2019. arXiv: 1812.06686 [cs.LG].

[41] S. Le, J. Hoffman, C. Barton, *et al.*, "Pediatric severe sepsis prediction using machine learning", *Frontiers in Pediatrics*, vol. 7, Oct. 2019. DOI: 10.3389/fped.2019.00413.

[42] H. Burdick, E. Pino, D. Gabel-Comeau, *et al.*, "Validation of a machine learning algorithm for early severe sepsis prediction: A retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals", *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, Oct. 2020. DOI: 10.1186/s12911-020-01284-x.

[43] A. Darwiche, A. EL-Geneidy, and S. Mukherjee, "Improving septic shock prediction with AdaBoost and cox regression model", in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, IEEE, Jan. 2021. DOI: 10.1109/iccece51280.2021.9342457.

[44] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning", in *2011 IEEE Control and System Graduate Research Colloquium*, IEEE, Jun. 2011. DOI: 10.1109/icsgrc.2011.5991826.

[45] Z. Zhou and G. Hooker, "Unbiased measurement of feature importance in tree-based methods", *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 2, pp. 1–21, Apr. 2021. DOI: 10.1145/3429445.

[46] L. Breiman, *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/a:1010933404324.

[47] T. Chen and C. Guestrin, "XGBoost", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2016. DOI: 10.1145/2939672.2939785.

[48] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting", *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.

[49] D. Steinley, "K-means clustering: A half-century synthesis", *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, May 2006. DOI: 10.1348/000711005x48266.

[50] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm", *J*, vol. 2, no. 2, pp. 226–235, Jun. 2019. DOI: 10.3390/j2020016.

[51] X. Yang, Y. J. Kim, F. Khoshnevisan, Y. Zhang, and M. Chi, "Missing data imputation for MIMIC-III using matrix decomposition", in *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, Jun. 2019. DOI: 10.1109/ichi.2019.8904824.

[52] Z. Hu, G. B. Melton, E. G. Arsoniadis, Y. Wang, M. R. Kwaan, and G. J. Simon, "Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record", *Journal of Biomedical Informatics*, vol. 68, pp. 112–120, Apr. 2017. DOI: 10.1016/j.jbi.2017.03.009.

[53] X. Shi, C. Prins, G. V. Pottelbergh, P. Mamouris, B. Vaes, and B. D. Moor, "An automated data cleaning method for electronic health records by incorporating clinical knowledge", *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Sep. 2021. DOI: 10.1186/s12911-021-01630-7.

[54] R. Li, Y. Chen, and J. H. Moore, "Integration of genetic and clinical information to improve imputation of data missing from electronic health records", *Journal of the American Medical Informatics Association*, vol. 26, no. 10, pp. 1056–1063, Apr. 2019. DOI: 10.1093/jamia/ocz041.

[55] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning", in *2014 Science and Information Conference*, IEEE, Aug. 2014. DOI: 10.1109/sai.2014.6918213.

[56] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Aug. 2007. DOI: 10.1093/bioinformatics/btm344.

[57] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection", *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, Nov. 1994. DOI: 10.1016/0167-8655(94)90127-9.

[58] D. Svozil, V. Kvasnicka, and J. Pospichal, "Introduction to multi-layer feed-forward neural networks", *Chemometrics and Intelligent Laboratory Systems*, vol. 39, no. 1, pp. 43–62, Nov. 1997. DOI: 10.1016/s0169-7439(97)00061-0.

[59] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. DOI: 10.1038/nature14539.

[60] C. Lin, Y. Zhang, J. Ivy, *et al.*, "Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM", in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, IEEE, Jun. 2018. DOI: 10.1109/ichi.2018.00032.

[61] S. M. Lauritsen, M. E. Kalør, E. L. Kongsgaard, *et al.*, "Early detection of sepsis utilizing deep learning on electronic health record event sequences", *Artificial Intelligence in Medicine*, vol. 104, p. 101 820, Apr. 2020. DOI: 10.1016/j.artmed.2020.101820.

[62] D. Zhang, C. Yin, K. M. Hunold, X. Jiang, J. M. Caterino, and P. Zhang, "An interpretable deep-learning model for early prediction of sepsis in the emergency department", *Patterns*, vol. 2, no. 2, p. 100 196, Feb. 2021. DOI: 10.1016/j.patter.2020.100196.

[63] S. Indolia, A. K. Goswami, S. Mishra, and P. Asopa, "Conceptual understanding of convolutional neural network- a deep learning approach", *Procedia Computer Science*, vol. 132, pp. 679–688, 2018. DOI: 10.1016/j.procs.2018.05.069.

[64] E. P. Raith, A. A. Udy, M. Bailey, *et al.*, "Prognostic accuracy of the SOFA score, SIRS criteria, and qSOFA score for in-hospital mortality among adults with suspected infection admitted to the intensive care unit", *JAMA*, vol. 317, no. 3, p. 290, Jan. 2017. DOI: 10.1001/jama.2016.20328.

[65] B. Khwannimit, R. Bhurayanontachai, and V. Vattanavanit, "Comparison of the performance of SOFA, qSOFA and SIRS for predicting mortality and organ failure among sepsis patients admitted to the intensive care unit in a middle-income country", *Journal of Critical Care*, vol. 44, pp. 156–160, Apr. 2018. DOI: 10.1016/j.jcrc.2017.10.023.

[66] B. Khwannimit, R. Bhurayanontachai, and V. Vattanavanit, "Comparison of the accuracy of three early warning scores with SOFA score for predicting mortality in adult sepsis and septic shock patients admitted to intensive care unit", *Heart & Lung*, vol. 48, no. 3, pp. 240–244, May 2019. DOI: 10.1016/j.hrtlng.2019.02.005.

[67] W. T. Lim, A. H. Fang, C. M. Loo, K. S. Wong, and T. Balakrishnan, "Use of the national early warning score (news) to identify acutely deteriorating patients with sepsis in acute medical ward", *Ann Acad Med Singapore*, vol. 48, no. 5, pp. 145–149, 2019.

[68] M. A. Reyna, C. S. Josef, R. Jeter, *et al.*, "Early prediction of sepsis from clinical data", *Critical Care Medicine*, vol. 48, no. 2, pp. 210–217, Feb. 2020. DOI: 10.1097/ccm.0000000000004145.

[69] Z. M. Ibrahim, H. Wu, A. Hamoud, L. Stappen, R. J. B. Dobson, and A. Agarossi, "On classifying sepsis heterogeneity in the ICU: Insight using machine learning", *Journal of the American Medical Informatics Association*, vol. 27, no. 3, pp. 437–443, Jan. 2020. DOI: 10.1093/jamia/ocz211.

[70] A. Aushev, V. R. Ripoll, A. Vellido, *et al.*, "Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase", *PLOS ONE*, vol. 13, no. 11, B. Mortazavi, Ed., e0199089, Nov. 2018. DOI: 10.1371/journal.pone.0199089.

[71] D. Chicco and L. Oneto, "Data analytics and clinical feature ranking of medical records of patients with sepsis", *BioData Mining*, vol. 14, no. 1, Feb. 2021. DOI: 10.1186/s13040-021-00235-0.

[72] J. R. A. Solares, F. E. D. Raimondi, Y. Zhu, *et al.*, "Deep learning for electronic health records: A comparative review of multiple deep neural architectures", *Journal of Biomedical Informatics*, vol. 101, p. 103 337, Jan. 2020. DOI: 10.1016/j.jbi.2019.103337.

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

[74] Wes McKinney, "Data Structures for Statistical Computing in Python", in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

[75] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[76] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database", *Scientific Data*, vol. 3, no. 1, May 2016. DOI: 10.1038/sdata.2016.35.

[77] "pgAdmin - PostgreSQL Tools". (), [Online]. Available: https://www.pgadmin.org/ (visited on 11/22/2021).

[78] *sklearn.model_selection.RandomizedSearchCV — scikit-learn 1.0.1 documentation*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model{_}selection.RandomizedSearchCV.html (visited on 11/22/2021).

[79] *MATLAB App Designer - MATLAB & Simulink*. [Online]. Available: https://www.mathworks.com/products/matlab/app-designer.html (visited on 11/22/2021).

[80] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks", *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020. DOI: 10.1145/3422622.

Appendix A

Methods' Appendix

This appendix contains figures, tables and graphs regarding the data preprocessing approach described throughout Chapter 4.

Table A.1: Number of patients in the assigned clusters, for each population and each dataset, during the missing data imputation.

	MIMIC-III		CHULC	
Cluster	**Sepsis**	**Control**	**Sepsis**	**Control**
1	107	1949	27	570
2	436	2843	25	909
3	820	2744	39	388
4	657	2195	38	223
5	172	102	17	1029
6	137	116	34	1446
7	583	4594	19	770
8	112	12	24	446
9	188	3104	8	3
10	465	949	31	326
11	39	23	6	675
12	157	423	26	724
13	10	3344	4	37
14	317	5331	22	30
15	7	4586	3	1111

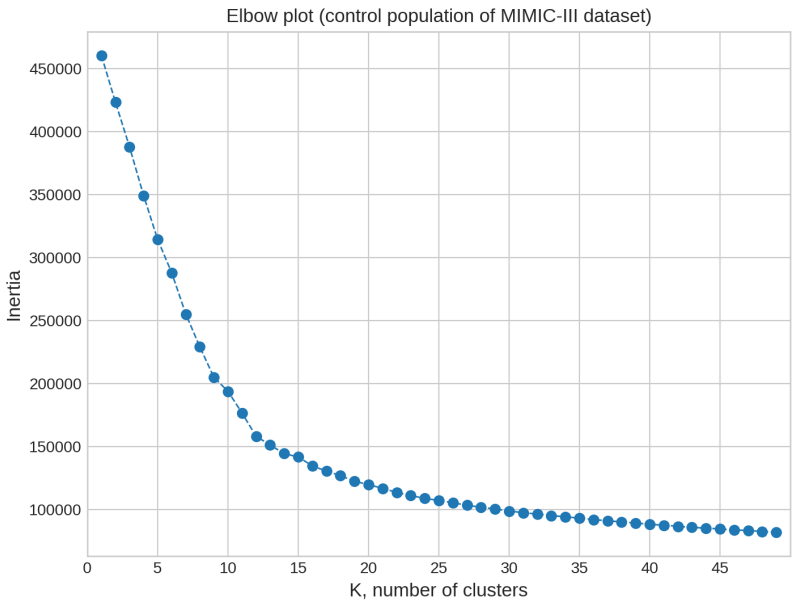


Figure A.1: Elbow method for determining the optimal number of cluster centroids, for the control population (MIMIC-III).

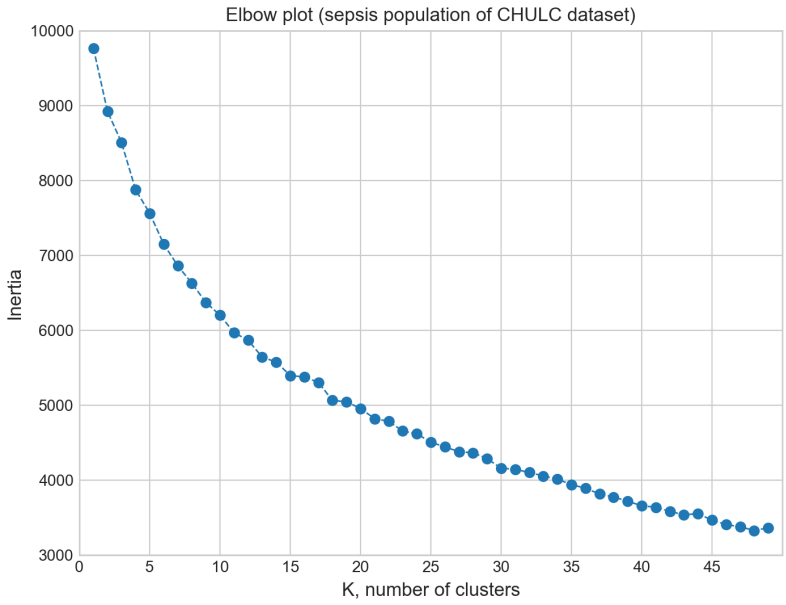


Figure A.2: Elbow method for determining the optimal number of cluster centroids, for the sepsis population (CHULC).

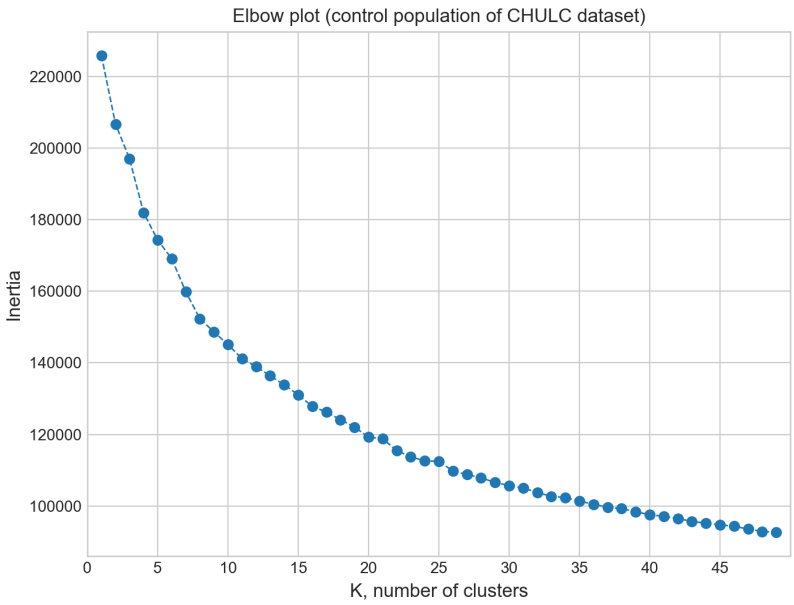


Figure A.3: Elbow method for determining the optimal number of cluster centroids, for the control population (CHULC).

Appendix B

Results Appendix

This appendix contains figures, tables and graphs regarding the training and testing of the ML models, described throughout Chapter 5, and the developed sepsis monitoring platform.

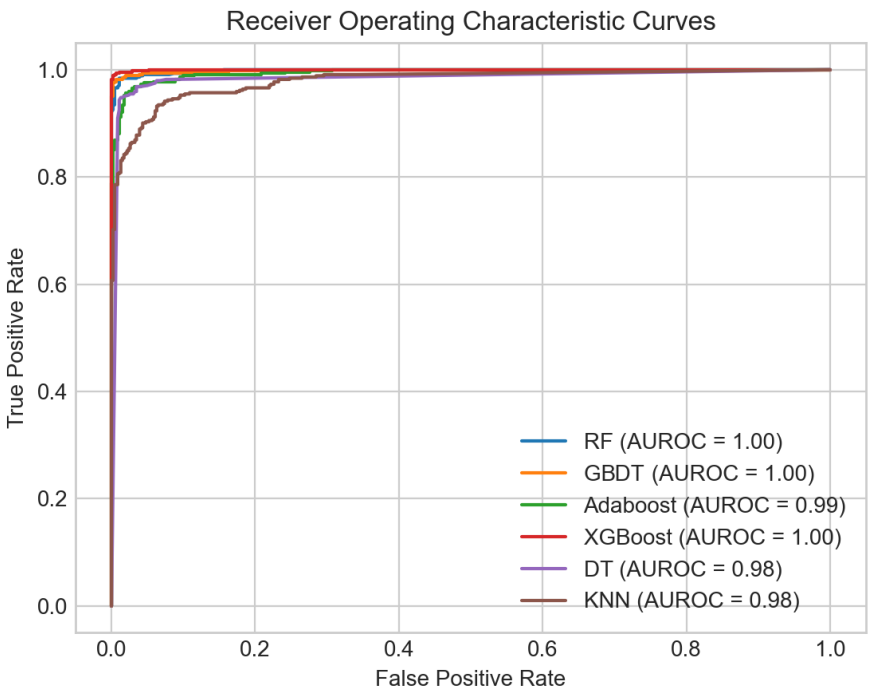


Figure B.1: ROC curves for training 1.

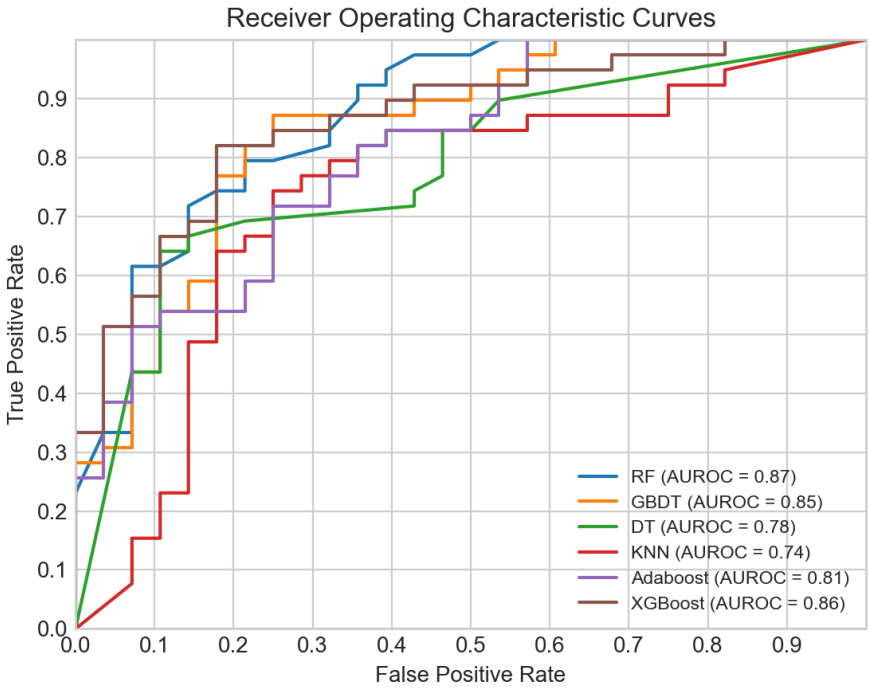


Figure B.2: ROC curves for training 3.

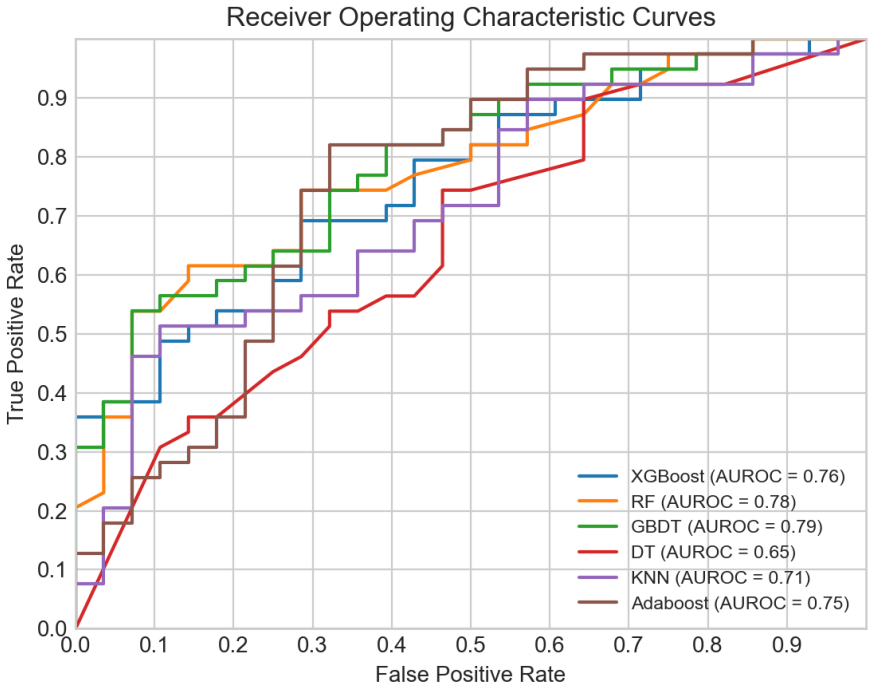
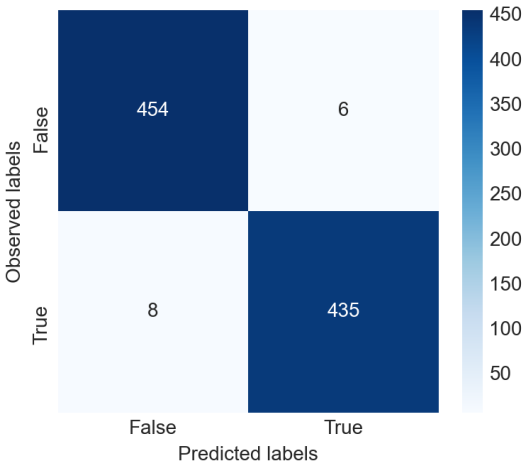
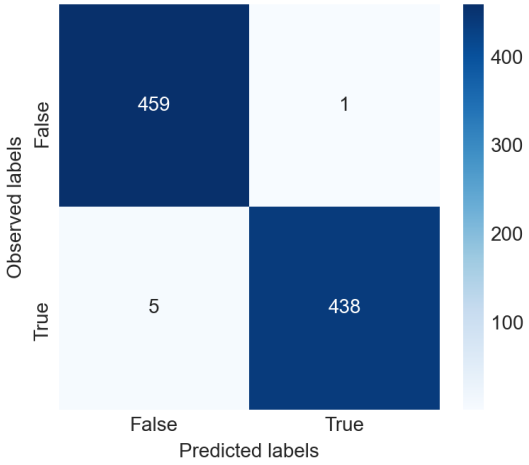


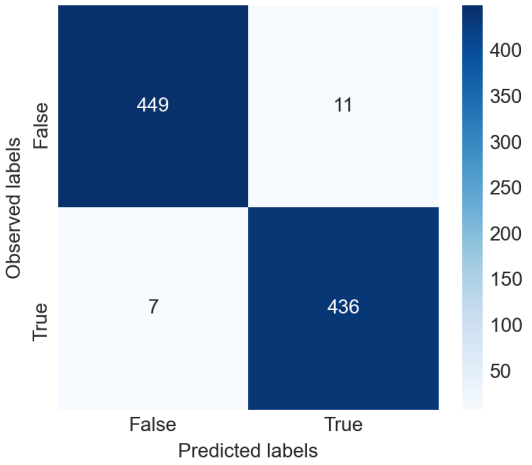
Figure B.3: ROC curves for training 4.



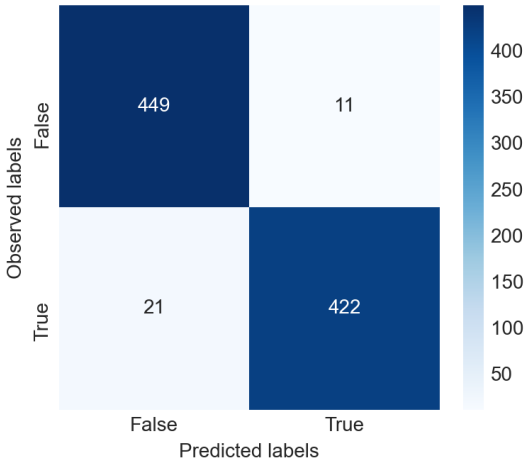
(a) Confusion Matrix: GBDT.



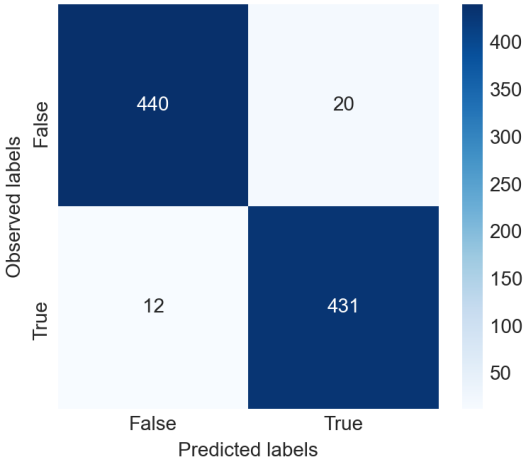
(b) Confusion Matrix: XGBoost.



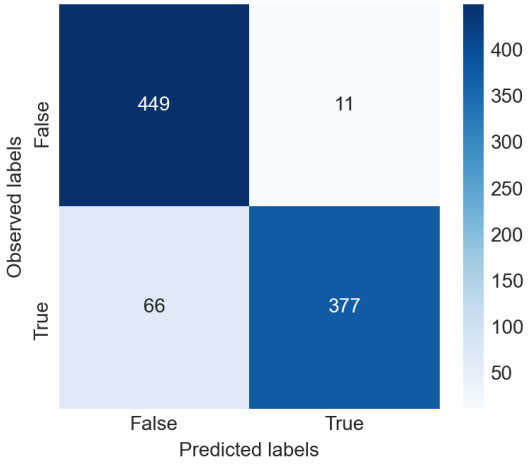
(c) Confusion Matrix: RF.



(d) Confusion Matrix: DT.

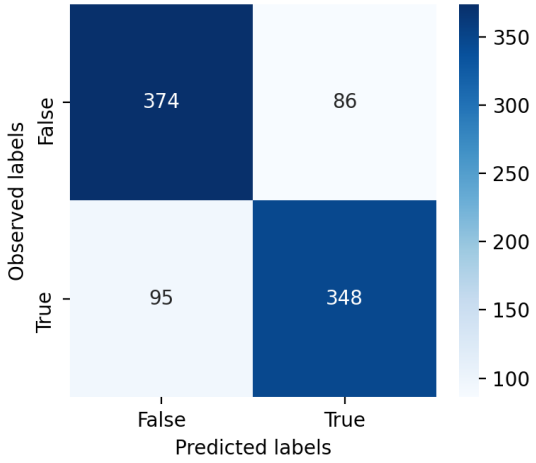


(e) Confusion Matrix: Adaboost.

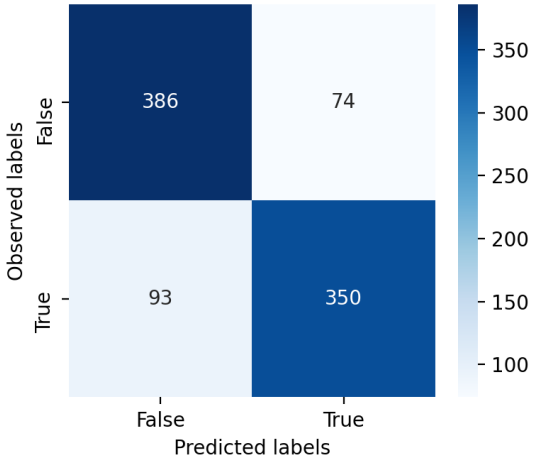


(f) Confusion Matrix: KNN.

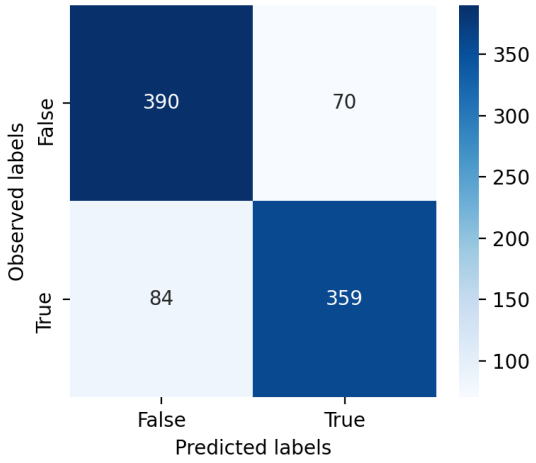
Figure B.4: Confusion matrices for training 1.



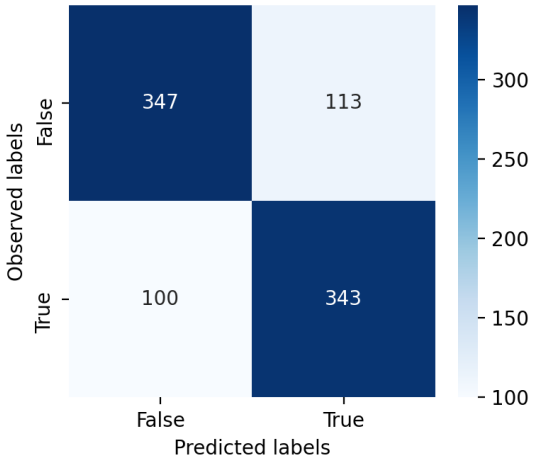
(a) Confusion Matrix: GBDT.



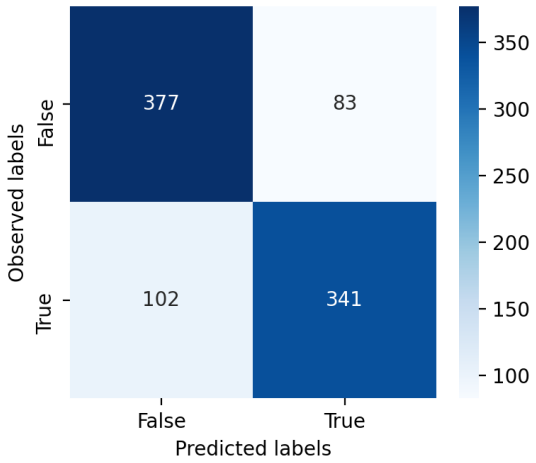
(b) Confusion Matrix: XGBoost.



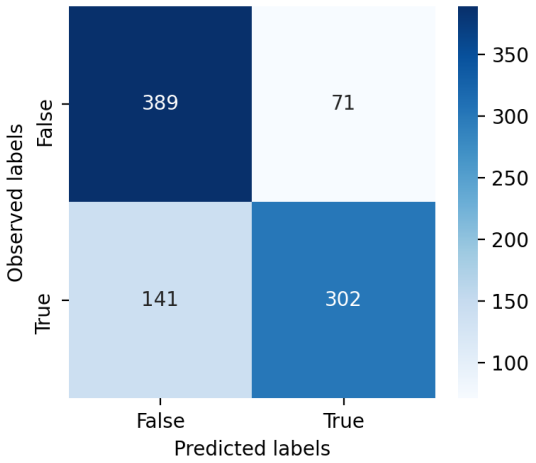
(c) Confusion Matrix: RF.



(d) Confusion Matrix: DT.

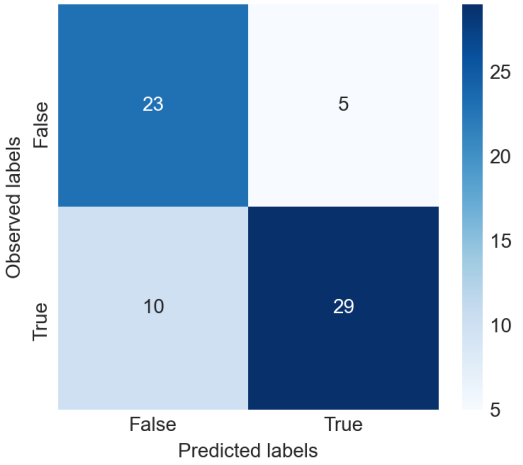


(e) Confusion Matrix: Adaboost.

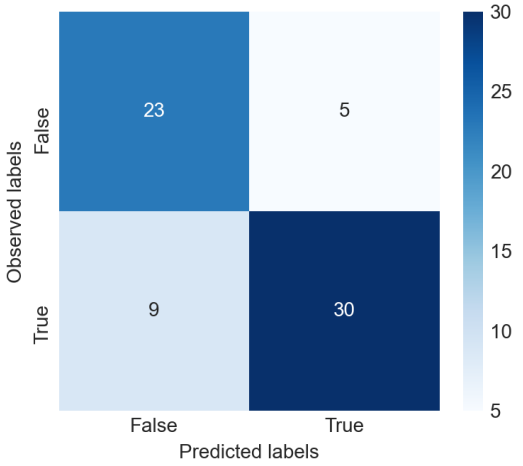


(f) Confusion Matrix: KNN.

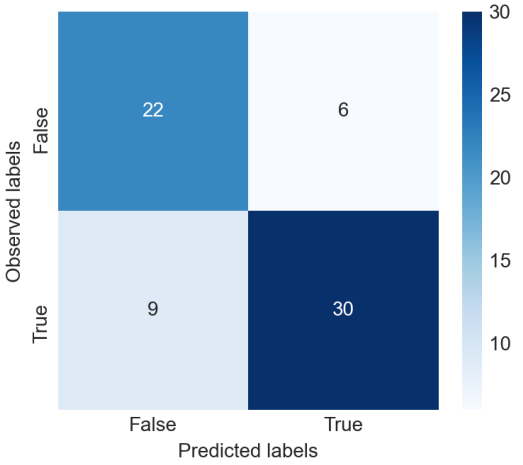
Figure B.5: Confusion matrices for training 2.



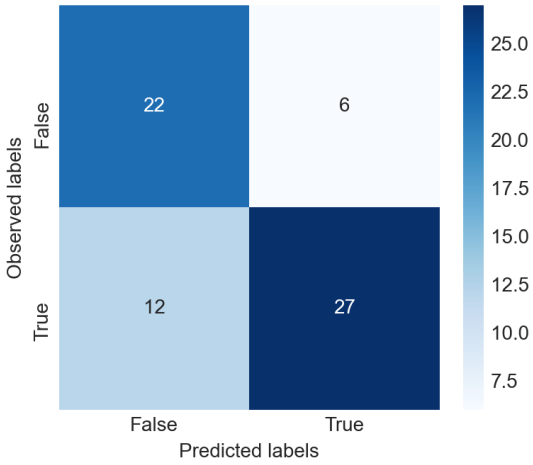
(a) Confusion Matrix: GBDT.



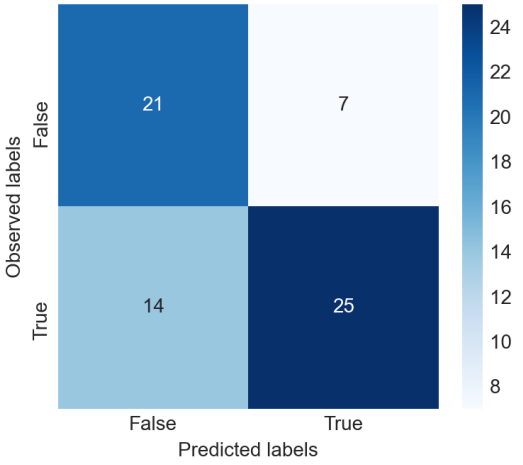
(b) Confusion Matrix: XGBoost.



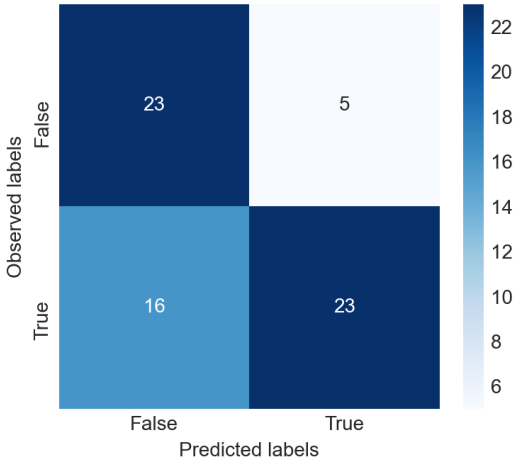
(c) Confusion Matrix: RF.



(d) Confusion Matrix: DT.

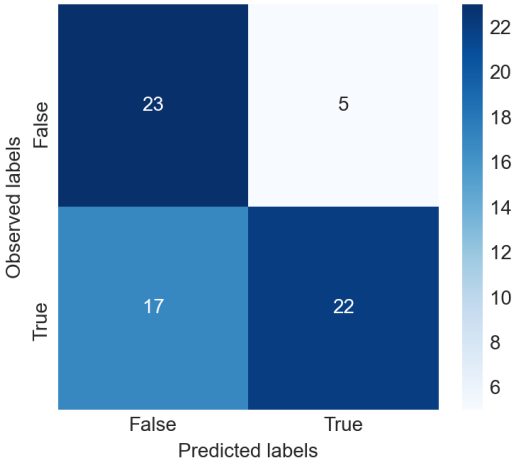


(e) Confusion Matrix: Adaboost.

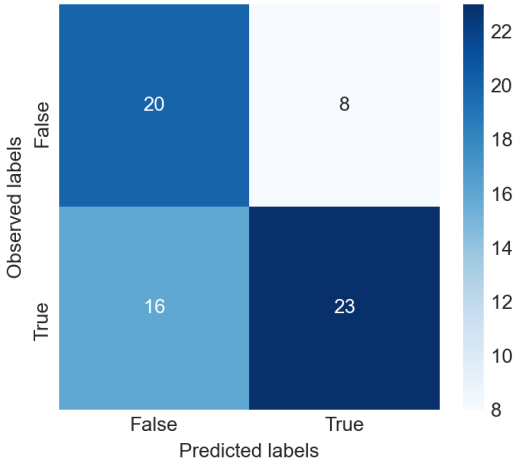


(f) Confusion Matrix: KNN.

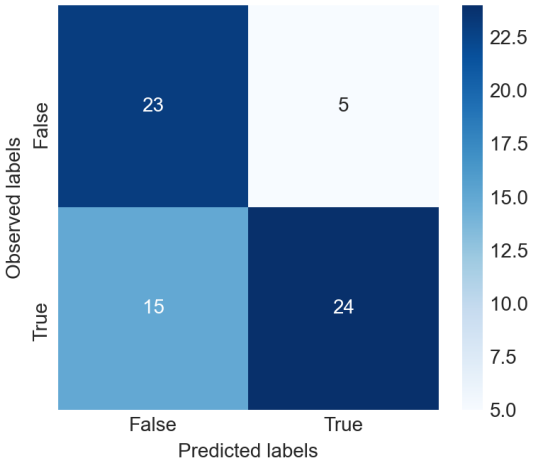
Figure B.6: Confusion matrices for training 3.



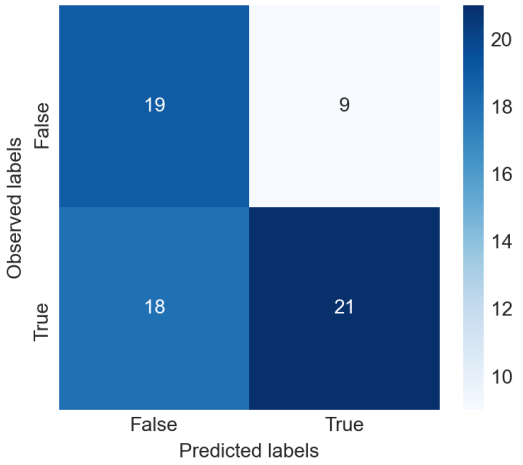
(a) Confusion Matrix: GBDT.



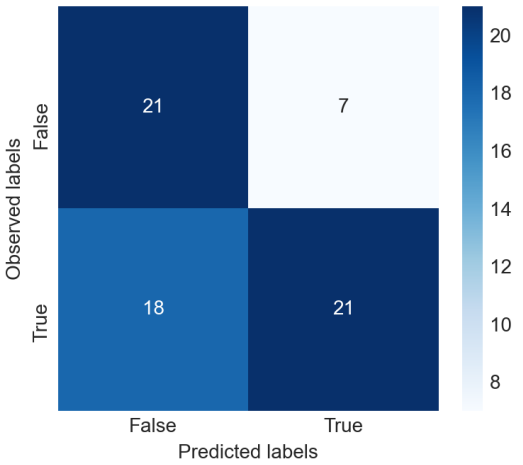
(b) Confusion Matrix: XGBoost.



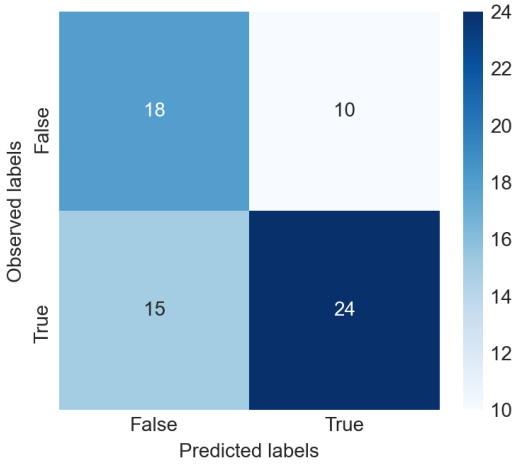
(c) Confusion Matrix: RF.



(d) Confusion Matrix: DT.



(e) Confusion Matrix: Adaboost.



(f) Confusion Matrix: KNN.

Figure B.7: Confusion matrices for training 4.

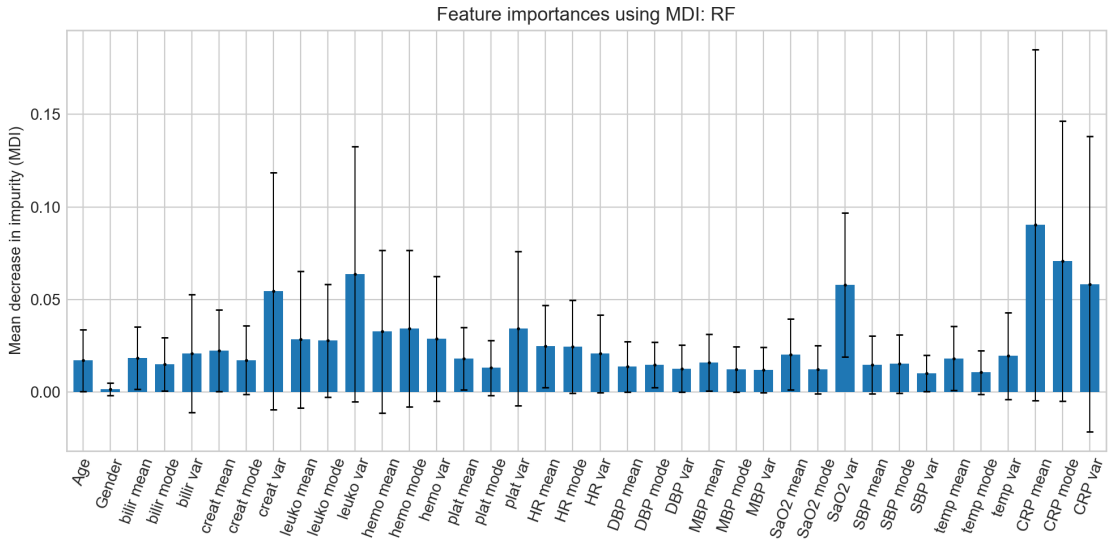


Figure B.8: Feature importance of the RF classifier, the best performing model, for training 3.

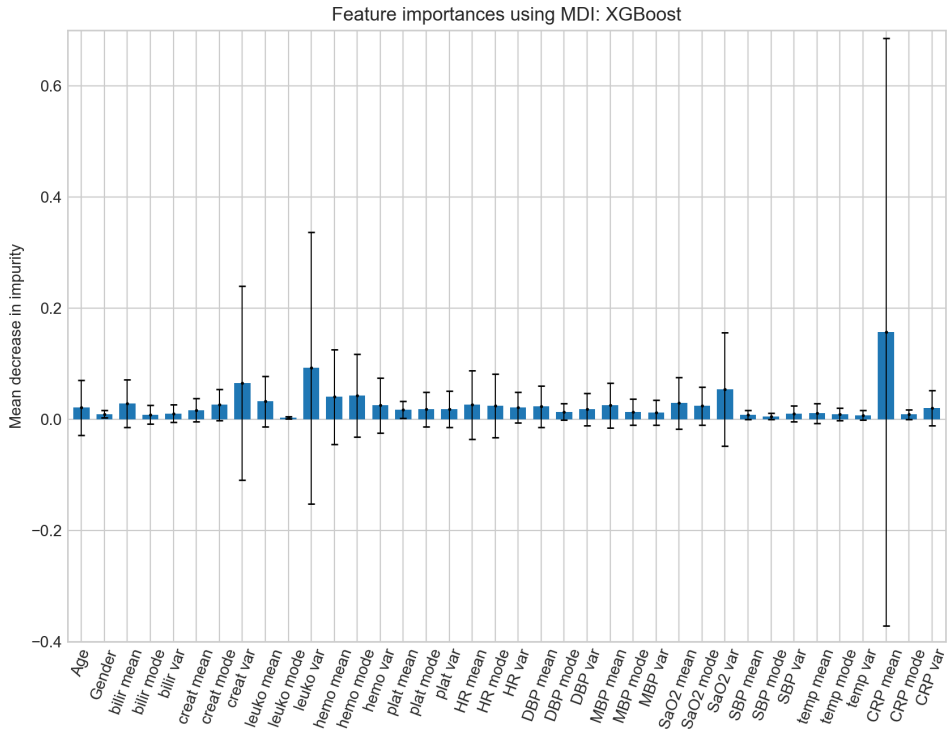


Figure B.9: Feature importance of the XGBoost classifier, the second best performing model, for training 3.

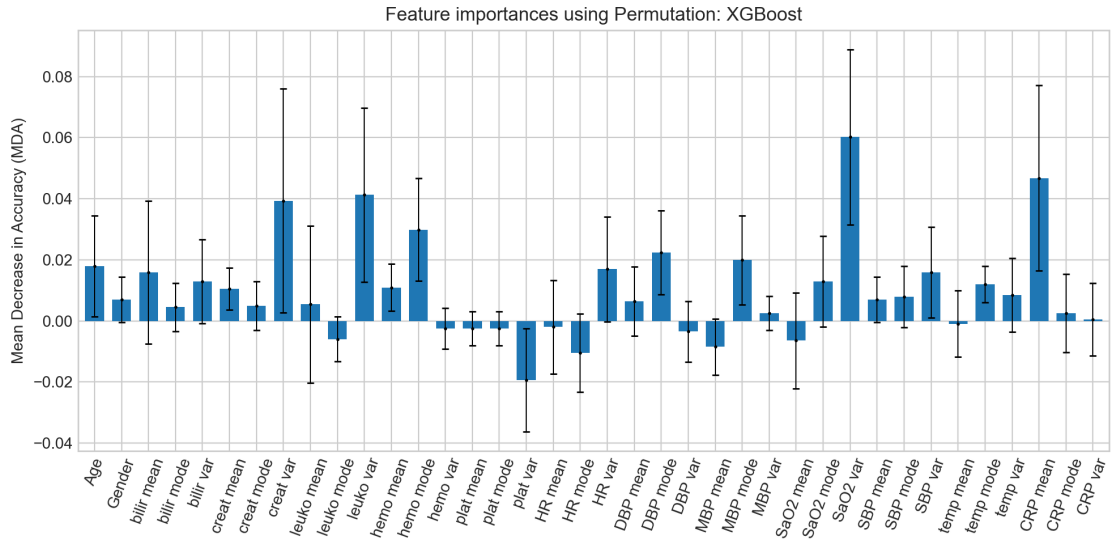


Figure B.10: Permutation feature importance of the XGBoost classifier, the second best performing model, for training 3.

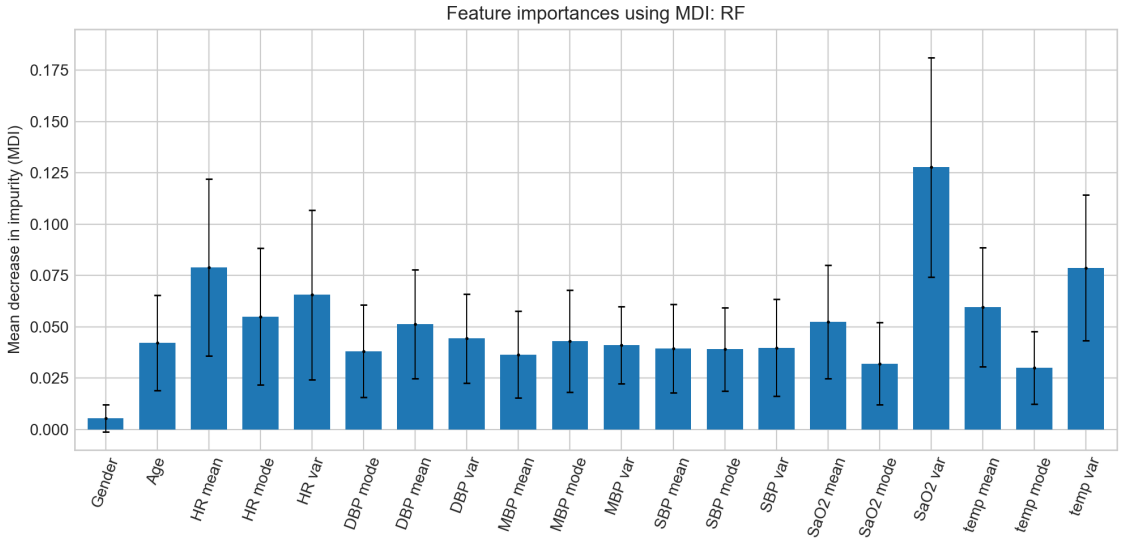


Figure B.11: Feature importance of the RF classifier, the best performing model, for training 4.

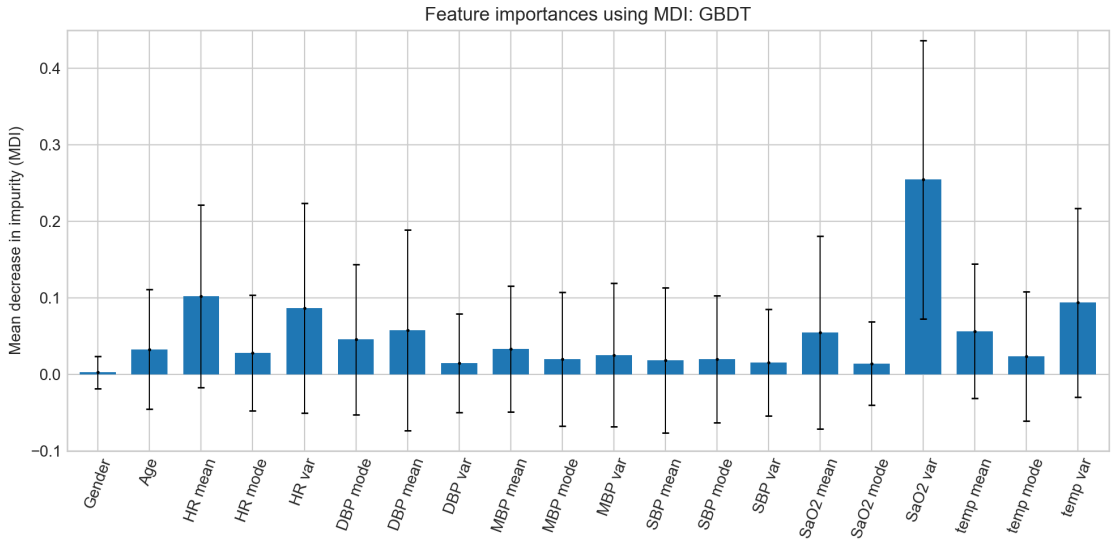


Figure B.12: Feature importance of the GBDT classifier, the second best performing model, for training 4.

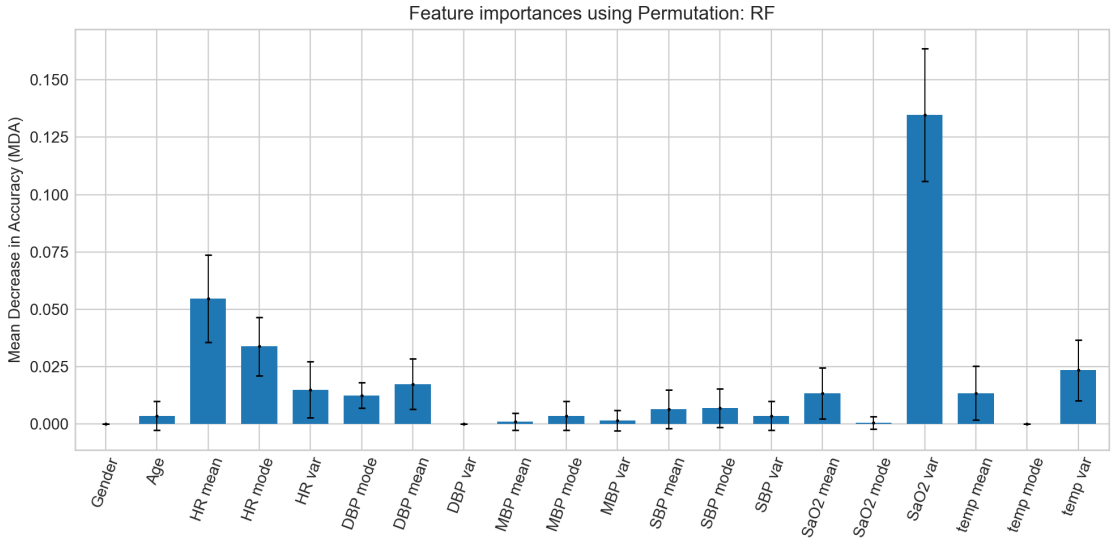


Figure B.13: Permutation feature importance of the RF classifier, the best performing model, for training 4.

Table B.1: Feature importance scores for the RF classifier, with the complete CHULC dataset.

Feature Importance										
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
SaO_2	mean	18	11	33	19	11	25	23.4	8.5	4.6
	mode	34	18	30	27	14	14	27.4	8.6	
	var	5	6	3	3	3	3	4.6	1.3	
creat	mean	15	25	20	10	12	12	18.8	5.8	6.2
	mode	23	22	14	21	23	17	24.0	3.7	
	var	6	1	13	2	6	3	6.2	4.4	
leuko	mean	11	33	8	14	22	21	21.8	9.1	8.6
	mode	12	27	19	20	27	8	22.6	7.7	
	var	3	2	2	32	2	2	8.6	12.2	
CRP	mean	1	20	1	1	1	29	10.6	12.5	11.9
	mode	2	38	36	33	30	1	28.0	17.1	
	var	4	29	18	30	7	5	18.6	11.9	
plat	mean	21	23	29	35	32	17	31.4	6.9	14
	mode	31	34	16	8	25	37	30.2	11.2	
	var	7	17	6	4	32	4	14.0	11.1	
DBP	mean	30	31	28	29	13	18	29.8	7.5	16.8
	mode	28	5	12	12	21	6	16.8	8.9	
	var	32	13	24	22	10	10	22.2	8.9	
hemo	mean	9	4	32	25	26	26	24.4	11.1	18.8
	mode	8	28	38	28	19	12	26.6	11.3	
	var	10	12	27	6	8	31	18.8	10.6	
temp	mean	22	15	37	5	13	5	19.4	12.1	19.4
	mode	36	14	15	18	20	17	24.0	8.1	
	var	19	32	7	11	30	7	21.2	11.2	
bilir	mean	20	24	4	15	28	9	20.0	9.1	20
	mode	27	35	22	13	5	8	22.0	11.7	
	var	16	16	21	16	27	13	21.8	5.0	
SBP	mean	29	8	11	31	4	18	20.2	11.2	20.2
	mode	26	19	34	36	32	7	30.8	11.0	
	var	37	30	26	38	9	14	30.8	11.9	
HR	mean	13	37	5	37	31	22	29.0	13.2	20.4
	mode	14	36	25	24	33	14	29.2	9.2	
	var	17	7	35	9	17	17	20.4	9.9	
MBP	mean	25	21	31	7	34	33	30.2	10.2	21.6
	mode	33	9	23	34	27	16	28.4	9.8	
	var	35	3	17	26	16	11	21.6	11.2	
age	-	24	26	10	17	28	6	22.2	9.0	22.2
gender	-	38	10	9	23	24	30	26.8	11.3	26.8

Table B.2: Feature importance scores for the XGBoost classifier, with the complete CHULC dataset.

Feature Importance										
Parameter	**Feature**	**MDI**	**MDA**	**SFS**	**SBS**	**SFFS**	**SBFS**	**Average rank**	**SD**	**Best rank**
SaO_2	mean	8	35	25	11	19	3	20.2	11.9	3.4
	mode	15	13	24	6	21	21	20.0	6.7	
	var	4	1	3	2	3	4	3.4	1.2	
leuko	mean	7	22	18	25	12	11	19.0	7.0	6
	mode	38	34	22	34	27	9	32.8	10.7	
	var	2	3	2	19	2	2	6.0	6.9	
CRP	mean	1	2	28	1	33	1	13.2	15.2	13.2
	mode	33	26	1	29	1	17	21.4	14.1	
	var	19	27	6	7	6	8	14.6	8.8	
bilir	mean	9	11	14	10	21	5	14.0	5.4	14.0
	mode	34	24	4	20	4	8	18.8	12.3	
	var	29	12	5	26	5	24	20.2	10.8	
creat	mean	24	16	9	31	9	15	20.8	8.7	14.6
	mode	10	23	11	22	11	12	17.8	6.0	
	var	3	4	26	3	32	5	14.6	13.2	
hemo	mean	6	15	10	24	10	25	18.0	7.9	17.4
	mode	5	5	16	21	23	17	17.4	7.8	
	var	12	32	20	17	21	19	24.2	6.6	
HR	mean	11	29	12	23	11	24	22.0	7.9	17.8
	mode	14	37	8	16	8	6	17.8	11.5	
	var	17	9	13	38	13	38	25.6	13.2	
MBP	mean	13	36	7	14	7	23	20.0	11.1	20
	mode	26	7	37	18	30	17	27.0	10.7	
	var	27	25	15	5	18	21	22.2	7.9	
DBP	mean	16	21	32	30	33	31	32.6	7.0	24.0
	mode	25	6	33	13	34	9	24.0	12.3	
	var	21	33	31	27	25	27	32.8	4.3	
temp	mean	28	28	27	4	21	13	24.2	9.8	24.2
	mode	32	14	30	37	25	17	31.0	8.9	
	var	36	17	23	8	37	20	28.2	11.3	
SBP	mean	35	19	38	9	17	4	24.4	13.7	24.4
	mode	37	18	17	15	35	18	28.0	9.9	
	var	30	10	19	35	30	35	31.8	10.0	
plat	mean	23	30	35	12	29	10	27.8	10.2	27.8
	mode	22	31	36	28	14	16	29.4	8.7	
	var	20	38	29	36	20	5	29.6	12.3	
gender	-	31	20	21	33	15	22	28.4	6.9	28.4
age	-	18	8	34	32	28	32	30.4	10.3	30.4

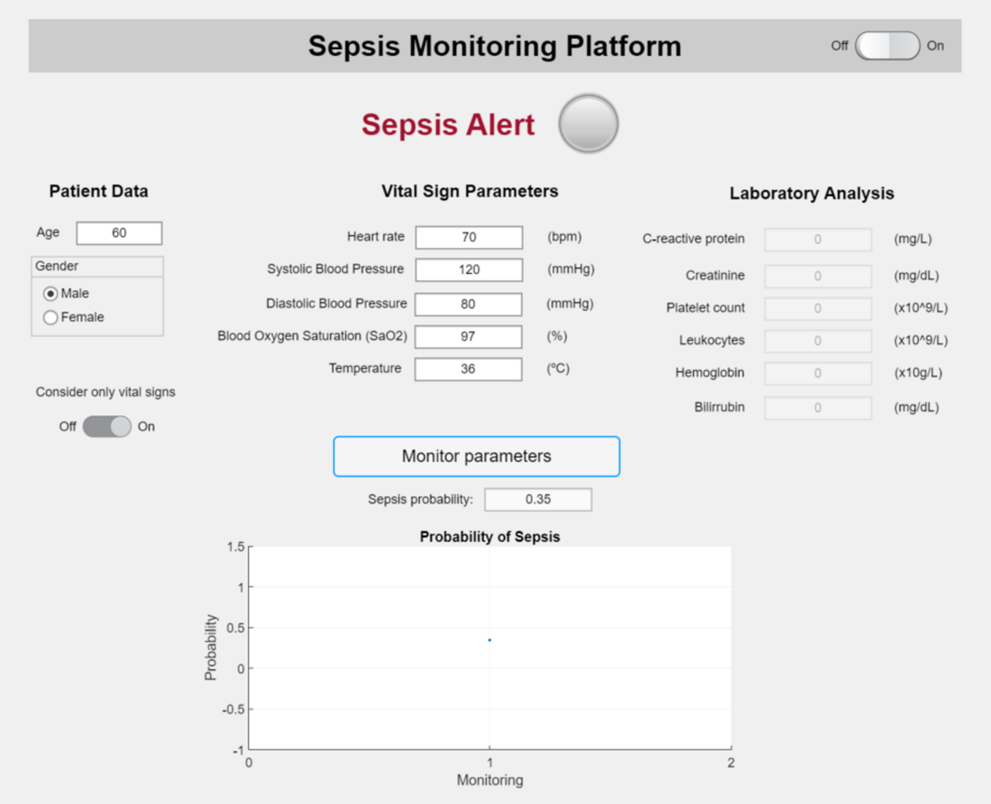


Figure B.14: Sepsis monitoring platform interface, with the option of considering only vital sign parameters for the detection, i.e., with the fields corresponding to the laboratory analysis work disabled.