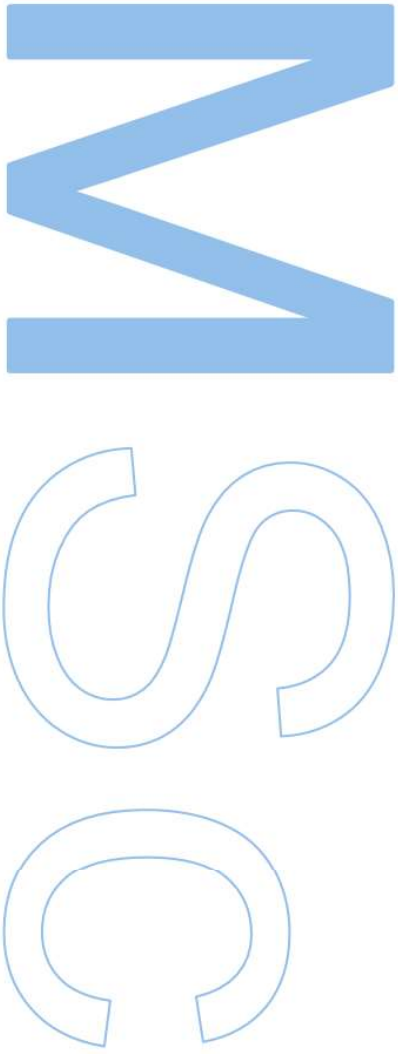
AI-based models to predict the Traumatic Brain Injury outcome

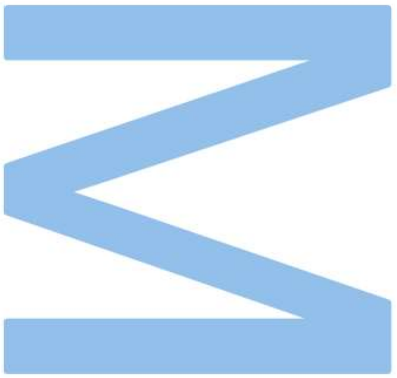


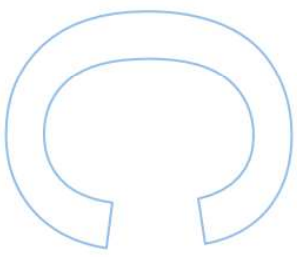
Ricardo Miguel Anjo Noronha Ribeiro
Master's in Computer Science
Department of Computer Sciences
2023

Supervisor
Hélder Oliveira, PhD, FCUP

Co-supervisor
Tânia Pereira, PhD, INESC TEC


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





Sworn Statement

I, Ricardo Miguel Anjo Noronha Ribeiro, enrolled in the Master Degree Computer Science at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Ricardo Miguel Anjo Noronha Ribeiro
28th June, 2023

Abstract

Traumatic Brain Injury (TBI) is a form of brain injury caused by external forces, resulting in temporary or permanent impairment of brain function. Common cases of TBI include falls, violence, sports activities, vehicle accidents, object strikes and intentional self-harm. Despite advancements in healthcare, TBI mortality rates remain around 20% in general cases and can reach 30%-40% in severe cases. By employing Artificial Intelligence (AI) methods and data-driven approaches to predict decompensation, this study aims to assist clinical-decision making and enhance patient care for TBI-related complications. This dissertation uses Machine Learning (ML) and Deep Learning (DL) approaches based on sequential data from the Medical Information Mart for Intensive Care (MIMIC) dataset, using Electronic Health Record (EHR).

In total, 2261 patients were in the TBI cohort from the MIMIC-III dataset, where 1567 are for training, 353 for validation and 341 for testing. Logistic Regressor (LR), Long-short term memory (LSTM) and Transformers architectures were developed and compared using balanced accuracy, precision, recall and Area Under the Receiver Operating Characteristics Curve (AUROC). Two set of features were also explored - original features and added features, with 17 and 47 features, respectively - and missing data strategies by imputing the normal value, data imbalance techniques with class weights and oversampling to try and improve results. Decompensation prediction was performed based on 24 hour in-mortality prediction at each hour of the patient stay in the Intensive Care Unit (ICU). Therefore, a different metric was also applied, by grouping the results for each sample (each hour in ICU) for each patient based on AUROC, meaning that if a patient that decompensated was predicted by the model to decompensate at any hour of the stay, it was considered as positive.

For the original features set, the best model was obtained with LSTM with an AUROC score of 0.9179 with no weights or oversampling strategies. Transformers obtained a score of 0.8788 AUROC with the oversampling technique. Regarding the added features set, the best result was obtained with LSTM with an AUROC score of 0.9294 using class weights. With the decompensation metric, it was possible to also conclude that the best model was also LSTM but with oversampling, with a score of 0.9524 AUROC. The model performance was evaluated across TBI categories (mild, moderate and severe), where the LSTM model with added features and oversampling obtained AUROC scores of 0.9355 and 0.8960 for moderate and severe, respectively. Mild was not possible to calculate AUROC, since there were no patients in both classes to perform this metric.

Overall, the models ability to improve prediction performance vary greatly with the models and methods used for data imbalance. For this study, LSTM proved to be the best model with the added features set, while transformers decreased significantly in performance with the added features.

Resumo

Traumatismos crânianos são uma forma de lesão cerebral causada por forças externas, resultando em danos temporários ou permanentes de funções cerebrais. Causas comuns da complicação incluem quedas, violência, atividades de desporto físico, acidentes rodoviários, golpes com objetos e danos próprios intencionais. Apesar das melhorias nos cuidados de saúde recentemente, a mortalidade desta complicação de saúde continua a obter valores de cerca de 20% no geral e pode chegar a 30%-40% em casos graves. Ao implementar métodos de inteligência artificial (AI) e abordagens que são baseadas em dados para prever a descompensação clinica dos pacientes, este estudo tem como objetivo assistir tomadas de decisão clinicas e melhorar os cuidados dos pacientes com complicações relacionadas a traumatismos cranianos. Esta dissertação aborda métodos de *Machine Learning* (ML) e *Deep Learning* (DL) com dados sequências obtidos no dataset do MIMIC-III, que também contém dados eletrónicos de saude dos pacientes.

No total, foram incluídos 2261 pacientes do MIMIC-III no grupo de estudo, cujos 1567 foram utilizados para treino, 353 para validação e 341 para teste. *Logistic Regressor* (LR), *Long-short term memory* (LSTM) e *Transformers* foram desenvolvidos e comparados usando as métricas de *balanced accuracy*, *precision*, *recall* e AUROC. Dois conjuntos de variáveis foram utilizados: variáveis originais e adicionadas, onde as originais contêm 17 variáveis e as adicionadas um total de 47, respetivamente. Técnicas para lidar com a falta de dados consistem em imputar o valor normal para cada variável. Estratégias para lidar com imbalanceamento de dados foram utilizadas com pesos das classes e *oversampling* para tentar melhorar os resultados. A previsão da descompensação clínica foi baseada no conceito de prever a mortalidade com 24 horas a cada hora da permanência do paciente nos cuidados intensivos. Por isso, uma métrica para comparar os modelos foi também aplicada, ao agrupar os resultados de cada *sample* (cada hora nos cuidados intensivos) para cada paciente, sendo esta métrica também baseada na AUROC. Assim, se um paciente que realmente teve uma descompensação foi também previsto pelos modelos como tal a qualquer hora nos cuidados intensivos, esta variável associa que o modelo fez uma previsão correta.

Para o conjunto de variáveis originais, o melhor modelo foi conseguido com o uso de LSTM, com uma *performance* de 0.9179 AUROC, sem utilizar pesos ou *oversampling*. Os *transformers* tiveram um resultado de 0.8788 AUROC com a técnica de oversampling. Em relação às variáveis adicionadas, o melhor resultado foi também obtido com a arquitetura LSTM, desta vez utilizando pesos nas classes proporcionais ao imbalanceamento dos dados, com 0.9294 AUROC. Ao utilizar

a métrica de descompensação, foi possível verificar que o melhor modelo continuou a ser LSTM, mas desta vez com *oversampling*, com um resultado de 0.9524 AUROC. Foi também feito a análise ao melhor modelo com base em cada categoria de traumatismo crâniano (leve, moderada e grave), o modelo LSTM com as variáveis extra e *oversampling* obtiveram um resultado de 0.9355 e 0.8960 AUROC para os casos moderados e graves, respetivamente. No que toca aos casos leves, não foi possível calcular o valor da métrica AUROC, uma vez que para este caso não existiam pacientes das duas classes.

No geral, a capacidade dos modelos de melhorar a previsão varia bastante com cada modelo e os diferentes métodos utilizados para o tratamento do balanceamento dos dados. Nesta dissertação, o modelo LSTM foi o melhor modelo, ao utilizar as variáveis adicionadas, enquanto que os *transformers* perderam bastante *performance* em comparação com os resultados das variáveis originais. Técnicas para lidar com o imbalanceamento dos dados como *oversampling* e pesos das classes proporcionais obtiveram resultados parecidos em termos de *performance*.

Acknowledgements

I would like to express my deepest gratitude to INESC TEC, and especially my supervisors, Tânia Pereira and Hélder Oliveira, for their guidance and expertise. Their continuous support and constructive feedback helped me in shaping the direction of this dissertation and taught me a lot in the past year.

To my closer family, especially my mother and father, who always supported me throughout all my academic years and personal life that led to this point. I would not have gotten here without them and I am fortunate to have them by my side.

To my close friends, both from childhood and from the past 5 years, who always supported me, laughed with me and experienced the highs and lows of this amazing journey with me.

A special thanks to all my co-workers, especially José, Oscar and Nuno, for their understanding and flexibility, which allowed me to balance my work responsibilities with the demands of my academic journey and this thesis. Their assistance in my professional life have been invaluable and I am grateful for all the opportunities.

Lastly, I would like to express my deepest gratitude to Catarina, for her understanding, patience, belief in me, encouragement and support throughout this entire journey. I am incredibly grateful for her support and having her by my side.

Ricardo Ribeiro

Contents

List of Tables

List of Figures

Acronyms

AI Artificial Intelligence

AUC Area Under the Curve

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristics Curve

ATP Triphosphate Production Adenosine

BBB Blood-Brain Barrier

CBF Cerebral Blood Flow

CT Computed Tomography

CV Cross Validation

CDC Centers for Disease Control

CNN Convolutional Neural Network

CPP Cerebral Perfusion Pressure

CRASH Corticosteroid Randomisation After Significant Head Injury

CSF Cerebrospinal Fluid

DBSCAN Density-Based Spatial Clustering of Applications With Noise

DCC Departamento de Ciência de Computadores

DL Deep Learning

ECG Electrocardiogram

EHR Electronic Health Record

eICU eICU Collaborative Research Database

EV External Validation

FN False Negative

FP False Positive

FCUP Faculty of Sciences of the University of Porto

GCS Glasgow Coma Scale

HPTBI Hackathon Pediatric Traumatic Brain Injury

HR Heart Rate

ICA Independent Component Analysis

ICD International Classification of Diseases

ICP Intracranial Pressure

ICU Intensive Care Unit

IMPACT International Mission for Prognosis and Analysis of Clinical Trials

INESC TEC Institute for Systems and Computer Engineering, Technology and Science

KFS Koehrsen's Feature Selector

KNN K-Nearest Neighbour

LOS Lenght of stay

LR Logistic Regressor

LSTM Long-short term memory

MAP Mean Arterial Pressure

MBP Mean Blood Pressure

mGCS motor Glasgow Coma Score

MLP Multilayer perceptron

MICE Multiple Imputation by Chained Equations

MIMIC Medical Information Mart for Intensive Care

ML Machine Learning

MRI Magnetic Resonance Imaging

OHE One Hot Encoding

PAM Partitioning Around Medoids

PCA Principal Component Analysis

PTS Physiologic time series data

RAIM Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data

RNN Recurrent Neural Network

RR Respiratory Rate

Sa02 Oxygen Saturation

SMOTE Synthetic Minority Oversampling Technique

SVM Support Vector Machine

TBI Traumatic Brain Injury

TN True Negative

TP True Positive

t-SNE T-distributed Stochastic Neighbor Embedding

Chapter 1

Introduction

1.1 Context

Traumatic Brain Injury (TBI) is a form of brain injury, caused by an external force that damages the brain and impairs brain function, temporarily or permanently.

Despite improvements in healthcare in recent years, TBI mortality still represents 20% in general and up to 30-40% in more severe cases. Over 50 million people worldwide have a case of TBI each year. Treatment for TBI patients are often performed by categorizing the severity according to Glasgow Coma Scale (GCS) scores, and usually followed by a head Computed Tomography (CT) scan. Studies suggest this method has reduced fatality in 50% in the last 150 years [37]. However, this method ignores the high heterogeneity of TBI.

1.2 Motivation

As stated before, currently TBI is treated as a 'one treatment fits all'. But, in reality, TBI is highly heterogeneous with different kinds of injuries - laceration, haemorrhage and skull fractures - and different regions of the brain - epidural, subarachnoid, intraparenchymal. As TBI is unique for each case, creating a prognosis is a hard task, and providing a proper treatment and care is difficult, given the current treatment method. In order to deal with all these variances, many data are evaluated for each patient, including medical signals, GCS, imaging such as CT imaging and its features. For physicians, it is hard to aggregate all this data, as it requires many areas of expertise – neurology, laboratory, imaging – and considering that TBI are often urgent cases that need a quick prognosis, the analysis of this data is a priority. Yet, hospital resources and Intensive Care Unit (ICU) are limited, which in turn makes it difficult to get a quick and effective early prognosis.

Therefore, taking advantage of computation power that is capable of analysing huge amounts of data and can also provide a quick prognosis can help drastically patients, as the prognosis is

faster, and can also help physicians and hospital resources.

1.3 Objectives

The main objective of this study is to develop methods based on Artificial Intelligence (AI) to predict decompensation on TBI patients. The datasets chosen for the study are publicly available.

Decompensation prediction was chosen since providing insights on patients with more probability of decompensation is important in identifying patients who are at a risk of experiencing a sudden deterioration in their health, allowing healthcare professionals to detect signs of decompensation and intervene early.

Another important part of this study is being able to take advantage of time series data like data, that have a closer resemblance to how the stays of the patients are in a real situation, which can be crucial when performing decompensation prediction.

1.4 Contributions

In this study, we create AI models based on Machine Learning (ML) and Deep Learning (DL) models based on sequential data that provide decompensation prognosis for TBI patients using the Medical Information Mart for Intensive Care (MIMIC) dataset. This time series data focuses on patients diagnoses, lab results and physiologic data. Another important contribution is the implementation of explainable methods that allow the clinicians to understand and ultimately trust the models decisions.

With the help of explainable methods and data-driven approaches of decompensation prediction, the ultimate contribution is to aid in clinical decision-making and improve patient care associated with TBI related complications.

Chapter 2

Background

This chapter provides a comprehensive overview of Traumatic Brain Injury (TBI) and its multifaceted impact on individuals. The chapter starts with some statistics about the disease, followed by its impact on patients who suffer from it. The chapter also delves into the more complex subjects about the disease, such as the complications of the primary and secondary injuries, as well as the clinical considerations. Medical imaging is also mentioned in the impact it can have on identifying an accurate diagnoses for TBI, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). By providing an in-depth exploration on these topics, the background chapter allows a better understanding of TBI and its research.

2.1 Introduction

TBI is a form of brain injury, caused by an external force that damages the brain and impairs brain function, temporarily or permanently. These injuries can be classified as closed or open, nonpenetrating or penetrating, respectively. Causes include falls, violence, sports activities, vehicle accidents and other transportation related crashes, unintentional struck by an object, and intentional self-harm. The most common cause of injury are falls, followed by unintentional struck by an object and intentional self-harm. According to 2014 Centers for Disease Control (CDC) data in the United States, intentional self-harm is the highest cause of death (32.5%), followed by falls (28.1%) and vehicle or other transportation related crashes (18.7%) [12]. Elders (aged 75 and older) have the highest rate of Intensive Care Unit (ICU) visits, followed by children (0 to 4 years old) and teenagers and young adults (15 to 24 years old). The most common reasons for TBI-related hospitalization according to age groups, falls are the most common cases for children, young adolescents and older adults and elders, while motor vehicle crashes is the most common cases for older teenagers, young adults and adults [12]. In terms of gender, studies from the TBI Model System National Database Statistics from 2021 shows that male cases account for 74% of the total cases, outnumbering female cases [8].

2.2 Problem Impact

Despite the healthcare improvements over the years, TBI mortality represents approximately 20% in general and up to 30%-40% rates in severe cases. Worldwide, more than 50 million people have a case of Traumatic Brain Injury each year, and it is predicted that half the population in the world will have one or more TBI cases over their life cycle. It has also been estimated that costs in healthcare for TBI are approximately 400 billion dollars annually [27]. These costs are mostly related to consequences of the condition, due to the patients and family members decrease in quality of life after severe cases of TBI.

TBI has always been a common health issue and in spite of many efforts in preventing this problem, applying work safety rules and better car safety, no studies have been published that show a decrease of frequency and rate of mortality. In fact, TBI-related ICU visits and deaths have increased from 2006 to 2014, according to the CDC [12]. According to Capizzi et al [4], 80% of Traumatic Brain Injuries in the United States are mild cases, and moderate and severe cases are 10% each, as seen in figure 2.1.

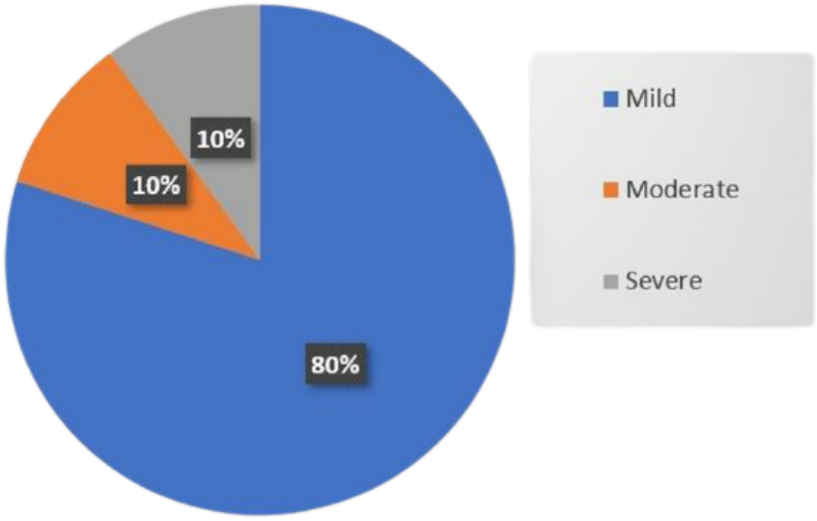


Figure 2.1: Severity distribution in the United States (Taken from Capizzi et al [4]).

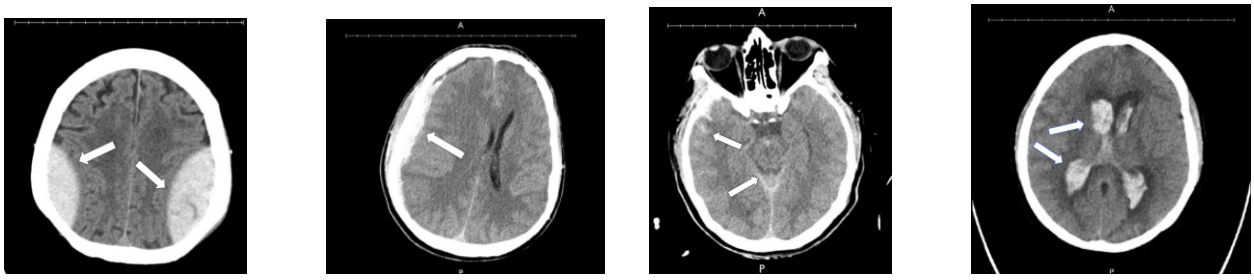
2.3 Primary and Secondary Injuries

These injuries can also be classified according to two different stages - primary and secondary injuries. Primary injuries are related to damages caused by an external force at the time of impact, such as direct hits, penetrating or nonpenetrating. This is considered the period of focal injury, where intracranial pathology resulting from focal injury includes: epidural or subdural hematomas and subarachnoid or intraventricular haemorrhage, as seen in figure 2.2. Evaluating the extension of the injury is important, as it can offer insights about how secondary injuries will

evolve and if further and more invasive treatments are necessary. The secondary injuries relate to damages at the cellular and/or molecular level. This can result in: ischemia, which causes cell death; vasogenic edemas, associated with cerebral contusion; and cytogenic edemas, related to hypoxic and ischemic injuries [4]. The next section is going to demonstrate some of the regions of the brain that were mentioned above.

2.4 Medical Imaging

Some regions of the brain are more prone to damage and proliferation of these secondary injuries, such as frontal and sub-frontal cortex, basal ganglia, diencephalon, temporal lobes and rostral brain stem [4]. Damages in these regions cause complications, being seizures that can show symptoms for up to 2 years after the injury, nauseas and vomiting, headaches, motor and memory deficit. Neuropsychiatric disorders can also be associated with complications after TBI incidents: depressions, anxiety, psychosis, paranoia, and aggression. Next, figure 2.2 shows CT scans of regions of the brain that are affected by TBI.



(a) Epidural Hematoma (b) Subdural Hematoma (c) Subarachnoid Hemorrhage (d) Intraventricular Hemorrhage

Figure 2.2: Focal Injuries (Adapted from Capizzi et al [4]).

In the first image 2.2c, the arrows point to a hyperdense lentiform, where blood appears in white. These types of injury are considered extradural and typically associated with injuries related to the middle meningeal artery. These hematomas can lead to loss of consciousness and lucid intervals, where after it leads to cognitive decline, since the intracranial pressure is increasing and can lead to herniation [4].

In image 2.2b, it is very noticeable a compression of the right lateral ventricle. These injuries are associated with damage to the bridging cortical veins and can suture lines [4].

A subarachnoid hemorrage is seen in image 2.2c, where hyperdensities are seen where the arrows are pointing. In the acute phase, this indicates subarachnoid bleeding and ruptured aneurysm is a common cause of this injury [4].

Finally, the last image 2.2d shows intraventricular bleeding. The lateral ventricles usually

appear black or hypodense on a CT scan and are associated with hydrocephalus [4].

2.5 Clinical Considerations

The primary injury can be not only the result of a direct impact, but also the result of acceleration, deceleration or rotational forces. Since the brain is a viscoelastic organ with little internal structural support, these forces cause a big impact on the brain. These forces lead to inertial forces on the brain tissue and cells, whereas linear acceleration forces lead to superficial brain injuries, since the gray matter closest to the surface of the brain is more susceptible to linear forces, that may cause cortical contusions and hemorrhage [3]; and rotational forces can explain lesions and disruptions on the deeper cerebral white matter axons, called diffuse axonal injury [39]. Apart from the direct forces applied in the initial trauma, the prolonged compressive forces of brain edema and hematomas can significantly decrease brain function by damaging brain tissues, raising Intracranial Pressure (ICP) and lowering Cerebral Blood Flow (CBF) [14].

The biomolecular response to TBI are the secondary injuries. For patients that survive the trauma of the primary injuries, mortality will be determined by the spread of secondary injuries. Excitatory amino acid release, oxygen radical reactions and nitric oxide production are related to the brain cell's response to injury. The associated calcium influx into brain cells, the high calcium and the presence of free-radical molecules create an unstable environment in the cell that lead to increased production and release of nitric oxide and amino acids, such as glutamate, that is the cause of the depolarization of cells in TBI [14]. Disruption of calcium homeostasis by gultamate-mediated ion channels and depolarization are important in the progression of secondary injuries. Calcium affects a huge amount of cellular functions: excess calcium leads to disruption of protein phosphorylation, and other enzyme functions [17]. Mitochondrias with high calcium result in swelling, depolarization and loss of function, which leads to cell death, via apoptosis or through loss of oxidative phosphorylation and failure of Triphosphate Production Adenosine (ATP) [22, 42]. ATP provides energy for the cells and the inability to generate ATP leads to cell death. Cellular injuries caused by TBI leads to changes in cellular protein synthesis. These injuries result in regulation problems, where some genes suffer from up-regulation that code some proteins and others down-regulation. Posttranslational modification of proteins may produce changes in protein levels that can be detected in serum or Cerebrospinal Fluid (CSF) [14].

The brain needs high amounts of oxygen and glucose to be functional. The skull gives protection from the external environment, and the Blood-Brain Barrier (BBB) keeps the integrity of the cerebral milieu [14]. The BBB fails in TBI due to the ischemic changes to vascular endothelial cells, leading to failure of autoregulation of the brain cells. Cerebral Perfusion Pressure (CPP) is the difference between the Mean Arterial Pressure (MAP) and ICP. CPP values below 50 mm Hg leads to brain tissue ischemia and autoregulation failure. CBF is low following TBI and half the normal value by the next day. There are many reasons for the decrease CBF from TBI, since it is highly dependent on blood pressure, oxygenation and blood

carbon dioxide concentrations [14]. The most accurate method of measuring ICP is to insert an ICP monitor. High ICP values lead to intracranial hypertension, and can be reduced by: the shrinkage of the volume of brain tissue, such as mannitol or other hypertonic substances; drainage of CSF; reduction of blood (hyperventilation-induced vasoconstriction); the surgical removal of the pathological process; or opening of the skull to allow expansion of the structures, known as decompressive craniectomy [43].

2.6 Glasgow Coma Scale

One important evaluation metric for TBI is Glasgow Coma Scale (GCS), which provides a practical method for physicians to assess the consciousness level in response to stimuli. GCS tests eye opening, verbal response and motor response of the patient and give points to each component according to specific classifications. Eye opening response is classified from 1 to 4, verbal response from 1 to 5 and motor response from 1 to 6, as shown in table 2.1.

Table 2.1: Summary of GCS categories and scores (Adapted from Capizzi et al [4]).

Score	Eye Response	Verbal Response	Motor Response
NT	Severe trauma to the eyes	Intubation, language disability	Paralysis
1	No response	No response	No response
2	Opens in response to pain	Incomprehensible speech	Extension response in response to pain
3	Opens in response to speech	Inappropriate speech	Abnormal flexion in response to pain
4	Spontaneous	Confused and disoriented speech	Withdrawal in response to pain
5	N/A	Oriented speech	Purposeful movement to localise pain
6	N/A	N/A	Obeys commands for movement

These values are then summed to a total that ranges from 3 to 15, where 3 is the minimum and 15 the maximum score. That total score is used to classify the severity of TBI in three possible categories: mild (13-15 points), moderate (9-12 points) and severe (3-8 points), as shown in table 2.2.

Table 2.2: Classification of TBI (Taken from Capizzi et al [4]).

	GCS (first 24 h)	Loss of Consciousness	Alteration of Consciousness	Imaging	Post Traumatic Amnesia
Mild	13-15	0-30 min	Up to 24 hours	Normal	0-1 days
Moderate	9-12	$>$ 30 min and $<$ 24 hours	$>$ 24 hours	Normal or abnormal	$>$ 1 day and $<$ 7 days
Severe	3-8	$>$ 24 hours	24 hours	Normal or abnormal	$>$ 7 days

There is a correlation between CBF and GCS in the first 24 hours of injury, where low CBF is associated with poor outcome [14].

2.7 Clinical decompensation

Decompensation is characterized by the deterioration or worsening of patient health condition. It is a term widely used in healthcare settings, particularly in ICU where patients require close monitoring. Decompensation can occur when a patient vital signs, physiological parameters or other indicators alter in accordance to normal ranges, indicating a possible decompensation. Idenfitying patient decompensation in ICU can drastically improve outcome by allowing patients with need of better treatment or observation to be identified and prevent morbidity. Physiological degradation or decompensation of patients is common and highly associated with morbidity and mortality [28]. The main challenge for decompensation prediction is the lack of gold standards and standardized clinical criteria that clearly identifies decompensation [28]. Recent studies try to define decompensation based on in-hospital mortality, the need of transfers between ICU and in-hospital cardiac arrest [28], as proxy alternatives to decompensation.

2.8 Conclusions

The current standard of care for TBI cases is a 'one treatment fits all', in which patients are categorized according to the GCS scores shown above. According to these results, further tests or treatments can be decided, typically followed by a head CT scan to conclude the exact location of the injury. Studies suggest that this approach has reduced fatality rates by over 50% in the last 150 years. However, the rate seems to have stagnated over the last 25 years [37]. However, this method ignored TBI heterogeneity. Further alternative approaches are needed for the development of new guidelines, in order to create better prognosis of the disease for each patient and decrease the rate of mortality. Attempts to provide individualised treatments to each patient are challenging due to the high heterogeneity of TBI. Methods that involve blood biomarkers, MRI, genomics and pathophysiological monitoring, combined with computational power able to process and analyse data, can offer new alternatives to improve TBI monitoring, better prognosis and better treatments [27]. Understanding protein regulation and response caused by secondary injuries in TBI could lead to interventions and reduce secondary injuries impact [23]. Also, understanding physiological decompensation of the patients can help in identifying patients in need of better care and in result prevent adverse outcomes.

Chapter 3

Literature Review

As Traumatic Brain Injury (TBI) is a common and impactful complication, there are plenty of case studies and research about the topic. Most of the literature found use a predictive approach for variables as mortality, while others focus on detecting TBI with imaging with Computed Tomography (CT) scans. Some articles are focusing on analysing sequential data and time series, a more complex approach that adds time as a factor, when compared to tabular data. The models used for these tasks also vary, according to the approach and type of data used. The age variable present in the articles seems consistent, mostly using only patients above 18 years old, as paediatric cases show a significant complexity given that the brain is still under development. This chapter of literature review contains the research made about the most common cases of Machine Learning (ML) approaches, current problems that impact the state of the art, the data that is used and also some interesting novel ones that were useful for this thesis research.

3.1 Current Challenges

In comparison to other fields in medicine, TBI and its clinical trials have difficult challenges. On one hand, TBI is not a single disease, it is very heterogeneous and complex with a wide range of pathologies, from the focal injuries (section 2.4) to extracerebral hematomas. On another hand, patients included in trials are also heterogeneous in clinical severity and prognostic risks [26].

The issues in the current models are both related to the nature of TBI and the methods used. Age is an important feature in TBI, even though many researches do not use pediatric age groups. Demographic variables such as age have a big impact on TBI, where it is noticeable that it is one of the most important feature in predicting TBI outcome. However, most researches do not use data from pediatric patients. The International Mission for Prognosis and Analysis of Clinical Trials (IMPACT) [38] and Corticosteroid Randomisation After Significant Head Injury (CRASH) [7] trials are two of the most well known ML models in regards to TBI, but they still disregard pediatric cases, only analyzing patients above 14 and 40 years old, respectively. This is due to the

differences in an adult brain and a brain still in development, in the pediatric cases. Nevertheless, finding effective ways to treat pediatric cases is very important, since children (0 to 4 years old) are one of the most common age group, as seen in section 2.1. Fonseca et al. [10] designed methods to predict TBI mortality in pediatric cases, with 4 different methods, where Area Under the Curve (AUC) values ranged from 0.84 to 0.91.

Other problems current literature has is that many models include very small sample sizes from a single region or center [29, 33]. This could lead to certain problems, since it is known that low and high-income countries have different resources, Intensive Care Unit (ICU) and different causes of TBI [7, 38]. A better explanation is given in section 3.1.1.

Another important factor is that many studies lack external validation. IMPACT [38] and CRASH [7] have a very thorough external validation, where both researchers have collaborated and performed external validation in each one's trials, but the current literature lacks external validation, which is very important before a trial can be approved and advised [20, 35].

3.1.1 Low-middle vs high income countries

Several articles published about TBI outcome prediction state that the model results depended a lot on the type of country that the injury occurs. Typically, low-middle and high income countries have significant differences in predictors' strengths. In the CRASH model [7], it is shown that the main predictors for low-income countries, the strongest prediction was a poor Glasgow Coma Scale (GCS), while for high-income countries it was age, as well as the predictors obliteration of the third ventricle and non-evacuated hematoma had more importance than for low-income. High-income countries also show that findings from CT scans are more relevant than in low-income ones. This could be due to the use of better technology, such as better healthcare, hospitals and medical equipment, a faster response to emergencies, and therefore more accurate diagnosis [7].

Some explanations are given, stating that in high-income countries, access to rehabilitation and healthcare services is generally more available and of a higher quality, which may lead to a better outcome for patients with TBI. Low GCS in low-income countries could be related to the less quality of care and/or use of sedation, or that old age in high-income countries, where younger ages have less risks but, as age increases, both types of countries have similar chances of a poor outcome at an older age [7, 38]. These changes in predictors have a high impact in external validation of the models. IMPACT showed better results in data from high-income countries rather than low-income, since its dataset was from high-income countries. For CRASH to have more discriminatory values, researchers developed specific models for each type of country [7].

3.1.2 Heterogenity

TBI is heterogeneous and has a wide range of pathologies, leading to different endotypes of the disease that lead to different symptoms, treatments and outcomes. In fact, TBI heterogeneity is recognized as a limit to find better prognostics, treatments and outcomes [44]. Further studies need to be performed about TBI heterogeneity and its endotypes. Nevertheless, endotypes could prove incredibly useful in providing insights and enabling better care and treatments for TBI patients.

Kim et al. [21] used unsupervised ML applied to Electronic Health Record (EHR) and Physiologic time series data (PTS) and was able to identify four different and clinically meaningful TBI endotypes. It was concluded in Kim results that the endotypes assigned patients according to outcomes, risks and laboratory signatures.

3.1.3 Missing Data

Most datasets from real scenarios, especially medical datasets, contain a large amount of missing values in the data. Besides that, some values in medical datasets contain defective values. Therefore, require pre-processing methods, in order to have usable data for the models. A summary of the methods found in literature for missing data approaches are mentioned in this section, while the methods for dealing with wrong values is explained in the methodology section.

The eICU Collaborative Research Database (eICU) dataset contained around 40% of missing values in the TBI features that were selected for Kim et al. [21]. To deal with those missing values, a Random Forest imputation method was used. Fonseca et al. [11] and IMPACT [38] used the Multiple Imputation by Chained Equations (MICE), which is a R package that performs imputation on missing values that occur in more than one variable [36]. Fonseca et al. [11] also used several other strategies of imputing missing data: using the value of the previous hour to impute to the missing value, using the value of the next hour, and using the normal value that the literature refers. According to the literature results from Fonseca et al. [11], the best imputation strategy was the normal value imputation, in terms of performance. The previous and next value imputation had close results and the MICE algorithm performed poorly [11]. All of these strategies were based on the benchmark by Harutyunyan et al. [16].

3.2 Data Available

Many studies in literature have access to private clinical datasets, that are not available to use. Those datasets are typically from a specific region or hospital. IMPACT and CRASH have the models available for using them in predicting outcome, but do not give access to the datasets as well. On the other hand, the studies made by Fonseca et al. and Kim et al. both used available datasets, the Medical Information Mart for Intensive Care (MIMIC) and eICU,

respectively. The fact that there are public datasets being used allow us not only to use them for this work, but also to compare the results with studies that use these datasets.

3.3 Previous Studies

In this section, a number of articles related to TBI that were found interesting and relevant to the problem aimed to solve are described and compared.

3.3.1 CRASH Model

The Corticosteroid Randomisation After Significant Head Injury TBI (CRASH) [7], from the London School of Hygiene and Tropical Medicine, is a model with the objective of predicting and validating prognostic models for death at 14 days and death or disability after 6 months of TBI. The variable death or disability (unfavorable outcome) was classified in 5 categories: death, vegetative state, severe disability, moderate disability and good recovery.

The study cohort consisted in 10008 patients that were adults and had a GCS of 14 or lower, excluding thus the mildest cases of injury and patients below 18 years old. The data was collected in studies and trials between 1999 and 2004. The predictors initially selected were age, sex, cause of injury, time from injury to randomization, GCS at randomization, pupil reactivity, CT results, if the patient had sustained a major extracranial injury and type of country (level of income, high or low). The data collected for the models were associated with prognostics in TBI and were included initially in a multivariable logistic regression analysis. Then, each variable was quantified by its predictive contribution by its z score and variables with significant level less than 5% were discarded.

CRASH developed 2 different models - a basic model, that included demographic and clinical variables, such as age, GCS, pupil reactivity and presence of major extracranial injury; and a CT model, which included the same predictors as the basic model, but added CT variables to it, as midline shift, non-evacuated haematoma, subarachnoid bleeding, obliteration of the third ventricle or basal cisterns, and presence of petechial haemorrhages. Apart from these two different settings, relevant interactions between the predictors and the level of income in a given country were found. With that in mind, they also predicted the two different outcomes for the two different types of countries for each model, with the two different models, resulting in 8 different models in total. The performance of the models was based on calibration and discrimination, where calibration was assessed with the Hosmer-Lemeshow test and graphically, and discrimination was assessed with the C statistics, which is equivalent to AUC [15]. The model was internally and externally validated. The first validation was performed using the bootstrap re-sampling method, and the external validation was performed using IMPACT cohort [38]. All the 8 different combinations of the models showed very good discrimination, with C statistics over 0.80 and 6 of the 8 models had good calibration in terms of Hosmer-Lemeshow

test.

The results are described in table 3.1, adapted from the article.

Table 3.1: CRASH Multivariable predictive models, C statistics values (Adapted from CRASH [7]).

Model	Mortality at 14 days		Death or severe disability at 6 months	
	High income countries	Low-middle income countries	High income countries	Low-middle income countries
Basic model	0.86	0.84	0.81	0.84
CT Model	0.88	0.84	0.83	0.84

3.3.2 IMPACT Model

International Mission for Prognosis and Analysis of Clinical Trials in TBI (IMPACT) [38] is a model that uses logistic regression models to predict mortality and unfavorable outcomes at 6 months after injury.

The study cohort contained 8,509 patients with severe or moderate TBI, with GCS of 12 or below. This data was obtained from trials and studies between 1984 and 1997. In contrast with the CRASH model [7], IMPACT had patients with age of 14 and above [38]. In terms of predictors, 26 predictors were initially selected, including demographics, indicators of clinical severity, data from secondary and various CT characteristics.

IMPACT developed 3 different prognostic models: the core model, the extended model and the lab model. The core model is the most basic one and contains the predictors of age, motor score and pupillary reactivity; The extended model contains the predictors from the previous models and adds the information from secondary injuries and CT characteristics; Finally, the lab model includes all the previous characteristics plus the additional information on glucose and hemoglobin. In terms of missing value, the MICE algorithm [36] was used, that will be further explained in the missing values section. The internal validation was performed using the AUC with cross validation from each of the 8 studies and 10 imputed datasets used to train the models.

The results were externally validated on the CRASH trial, although the lab model could not be validated, since CRASH did not include lab results in the trial. Therefore, the only two models to be validated were the basic model and a variant of the extended model. Nevertheless, external validation confirmed both models discriminatory abilities. The core model had AUC values of 0.776 for mortality and 0.780 for unfavorable outcome. For the extended model, the results were even better, with AUC values of 0.801 and 0.796, for mortality and unfavorable outcome, respectively. The discrimination values in AUC are shown in table 3.2.

Table 3.2: AUC values of the IMPACT models at Cross Validation (CV) from studies and External Validation (EV) from the CRASH Trial (Taken from IMPACT [38]).

Dataset	Selection	Study	n	Mortality			Unfavorable Outcome		
				Core	Extended	Lab	Core	Extended	Lab
CV in IMPACT	RCTs	TirInt	1,118	0.70	0.78	0.80	0.75	0.80	0.80
		TirUS	1,041	0.74	0.77	0.79	0.78	0.80	0.82
		Saphir	919	0.66	0.71	0.72	0.70	0.73	0.75
		Pegsod	1,510	0.76	-	-	0.77	-	-
		HIT-II	819	0.72	0.77	-	0.74	0.78	-
	Obs. Studies	UK4	791	0.81	0.81	-	0.81	0.80	-
		TCDB	604	0.81	0.83	-	0.82	0.83	-
		EBIC	822	0.84	0.87	-	0.81	0.84	-
EVin CRASH	All with GCS <= 12	CRASH	6,272	0.78	0.80	-	0.78	0.80	-
	Placebo with GCS <= 12	CRASH	3,075	0.78	0.81	-	0.78	0.79	-
	HC with GCS <= 12	CRASH	1,466	0.80	0.83	-	0.77	0.80	-

As was seen also in the CRASH models, the most important predictors were age, motor score and pupillary reactivity at admission [7].

3.3.3 Identification of Quantitative Data-Driven Endotypes

Kim et al. [21] developed ML models to predict TBI outcome in terms of motor Glasgow Coma Score (mGCS) and mortality, and focused on a novel aspect of TBI research, endotypes.

The cohort used was the eICU and MIMIC was used for external validation. The variables of interest were extracted from EHR and PTS, such as Heart Rate (HR), Oxygen Saturation (Sa02), Respiratory Rate (RR), Mean Blood Pressure (MBP). For the analysis of the time series values, Kim et al. [21] used featurization tools: HCTSA [13] for Matlab and TS fresh [6] for python. These featurization tools were useful, although the number of features was very high and difficult to explain, since the number of features extracted was between 3000 and 9000. Therefore, only four categories from the PTS were selected [21]. The features selected and its pipeline are demonstrated in figure 3.1.

With the pipeline in figure 3.1, the total number of features extracted was 186, which is easier to explain than the 3000 to 9000 features that the extraction tools were able to get.

The ML models used were Generalized linear model, random forests and XGBoost model and aimed to predict mortality outcome. The metrics used for model evaluation was the Area Under the Receiver Operating Characteristics Curve (AUROC). In order for Kim et al. [21] to compare the results and have a baseline, a revised model performance for both IMPACT [38] and CRASH [7] models was performed with the eICU cohort.

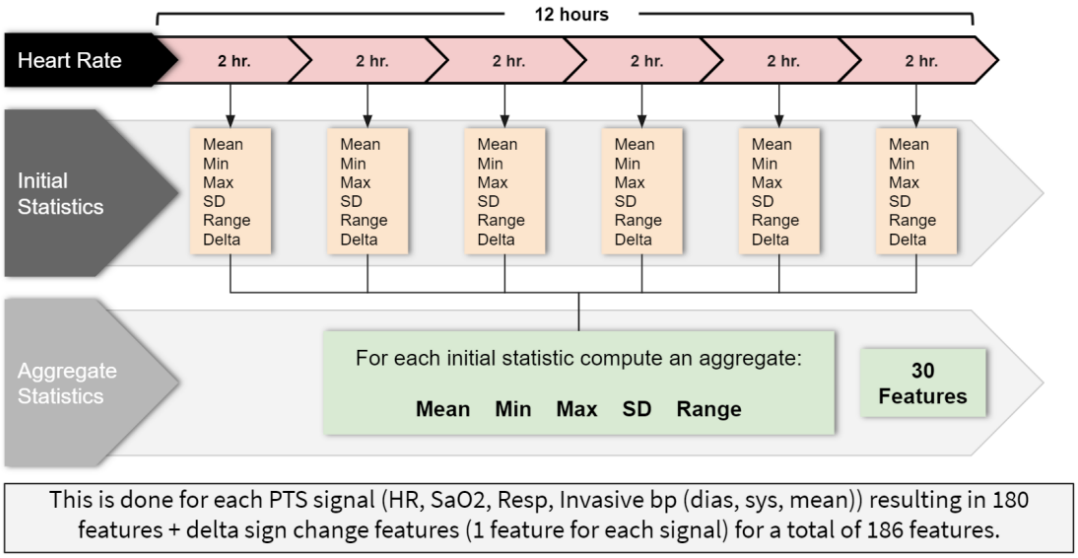


Figure 3.1: PTS feature extraction pipeline (Taken from Kim et al [21]).

Table 3.3: eICU test results (Adapted from Kim et al [21]).

Model	**AUROC**	
	mGCS	mortality
Revised IMPACT	0.790	0.879
Revised CRASH	0.776	0.867
Kim et al. (RF)	0.923	0.895

As seen in the results from table 3.3, the models created in this study outperformed the revised IMPACT [38] and CRASH [7] model. Kim et al. [21] compared two more models (GLM and XGboost), having both also surpassed the results from IMPACT and CRASH.

For the endotypes, an unsupervised ML approach was used with the k-means model with Gower distance, after studying approaches such as Partitioning Around Medoids (PAM) [40], Density-Based Spatial Clustering of Applications With Noise (DBSCAN) [9], spectral clustering [30], hierarchical clustering [32]. In order to visualize the results of the k-means clustering, Principal Component Analysis (PCA) [25] and T-distributed Stochastic Neighbor Embedding (t-SNE) [41] were used [21].

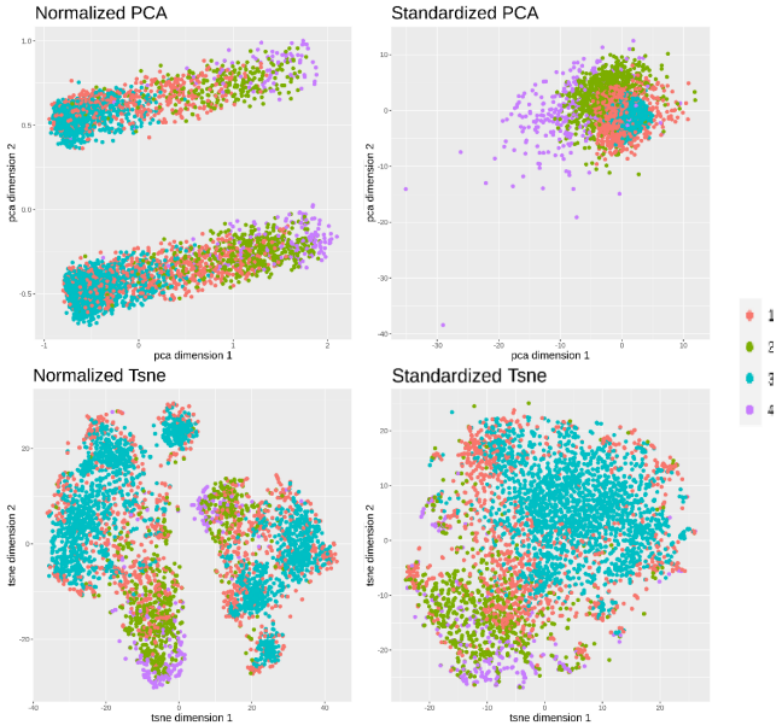


Figure 3.2: PCA and t-SNE endotypes (Taken from Kim et al [21]).

Figure 3.2 shows the visualization of the clustering techniques mentioned. In the normalized PCA plot, we can see two separate chunks in the clusters, that are attributed to the gender feature. Besides that, separate clusters are seen in the Standardized t-SNE plot. Figure 3.3 shows the validation of the clusters in comparison with the actual outcome in the dataset.

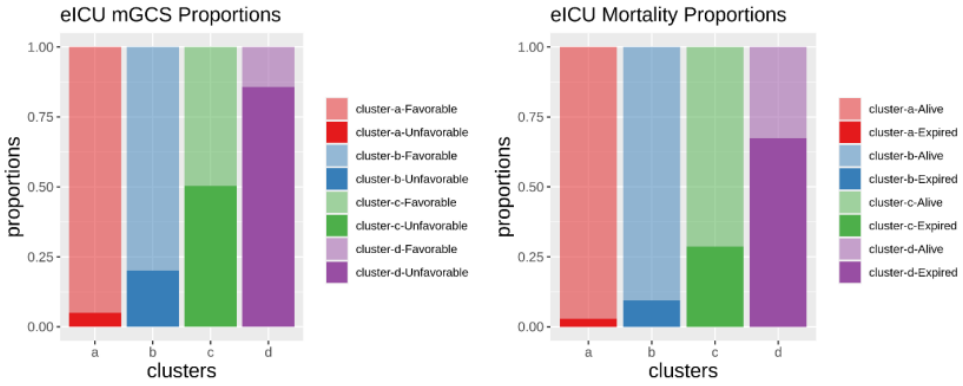


Figure 3.3: mGCS and mortality outcomes according to the four endotypes (Taken from Kim et al [21]).

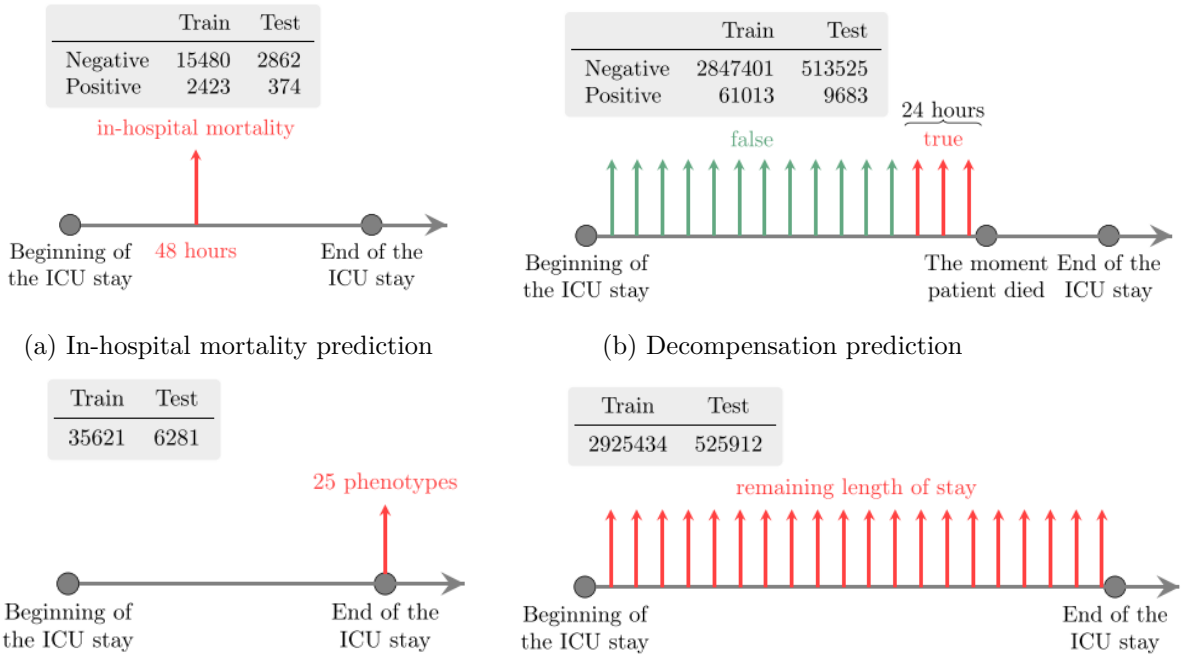
The four endotypes in figure 3.3 clearly show a distiction in terms of unfavorable outcomes. The first cluster, cluster A, seems to be associated with very mild cases of TBI, in both mGCS and mortality, whereas cluster D represents the most severe cases in both predictions.

3.3.4 Multitask learning and benchmarking with clinical time series data

Harutyunyan et al. [16] developed methods for four prediction problems in the MIMIC dataset: mortality, decompensation, forecasting lenght-of-stay and phenotype classification. This study also includes benchmarks to perform all the necessary data preprocessing steps in the MIMIC dataset, although it does not focus on the TBI disease.

In this study, in-hospital prediction is predicted based on the first 48 hours of patient admission in ICU. Decompensation prediction is predicted by identifying if the patient will die in the next 24 hours. This task was defined as mortality prediction for each hour of a patient stay in ICU. It is stated on the study that this method might deviate from the main meaning of decompensation, but it is the closest proxy to decompensation with accurate labels from MIMIC. Lenght of stay (LOS) prediction focuses on forecasting the remaining time on a patient ICU at each hour and finally, phenotype classification tries to classify 25 acute conditions. In terms of metrics, mortality and decompensation use AUROC, phenotyping uses the Cohen's linear weighted kappa score, and phenotype classification uses macro-averaged AUROC [16].

Figure 3.4 shows a visualization of how the prediction tasks mentioned above are performed. In this ilustration it can be seen that while in-hospital mortality (figure 3.4a) and phenotype (figure 3.4c) predictions are performed once for each patient, decompensation (figure 3.4b) and length of stay (figure 3.4d) predictions are performed at each hour.



(a) In-hospital mortality prediction

(b) Decompensation prediction

(c) Phenotype prediction

(d) Length of stay prediction

Figure 3.4: Multitask prediction tasks (Taken from Harutyunyan et al [16]).

In the tables also seen in subfigures 3.4a, 3.4b, 3.4c, and 3.4c the data used for training is much larger for the cases where the predictions are done hourly (subfigures 3.4b and 3.4d). This

is due to the fact that each sample used refers to each hour of each patient in the MIMIC dataset, while for the other cases the samples correspond to each patient stay, where for in-hospital mortality the sample contains only the first 48 hours and phenotypes include the whole stay [16].

This study provides an extensive benchmark to deal with the preprocessing steps for the MIMIC dataset and states some good practices to use in this benchmark. Multiple stays in ICU or transfers between ICU units are excluded for every hospital admission to reduce ambiguity. This study, according to some of the studies previously mentioned, also remove the pediatric ages due to the difference in physiology. Events are also validated that filters events that can not be matched to stays, according to admission ID. After these steps the study extracts episodes for each subject that correspond to the time series of each patients, that can then be used for the multitask predictions mentioned [16].

In terms of results, this study experimented with logistic regressions and Long-short term memory (LSTM), where LSTM were used in different architectures. In-hospital mortality obtained AUROC score of 0.870 using a multitask channel-wise LSTM; for decompensation prediction, the best result was obtained with a channel-wise LSTM with deep supervision, with 0.911 AUROC; phenotyping obtained 0.776 AUROC with a channel-wise LSTM and length of stay obtained a kappa value of 0.451 with channel-wise LSTM with deep supervision [16]. Note that these results were obtained by predicting these multitask methods for all MIMIC dataset and was not specific of TBI.

3.3.5 RAIM Model

Xu Yanbo et al. [45] developed Recurrent Attentive and Intensive Model of Multimodal Patient Monitoring Data (RAIM). This model introduces attention mechanisms for Electrocardiogram (ECG), real-time vital signs and medications [45]. The dataset used was related to MIMIC-III, but it was MIMIC-III Waveform Database Matched Subset, which is a dataset derived from MIMIC-III [1, 2].

This study focuses on two different prediction tasks: decompensation prediction and forecast the length of stay. The models used were various versions of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN), and the model with best results was the RAIM-3, which consisted of a CNN followed by multichannel attention by lab and intervention data [45]. The results obtained with this model were 0.9018 AUROC for decompensation and 0.8291 Kappa for length of stay forecasting, respectively. Figure 3.5 was taken from Xu Yanbo et al. [45] and shows a visualization of a decompensation prediction for a patient that deceased. It can be seen in yellow medium attention, and orange the high attention zones that the model identified.

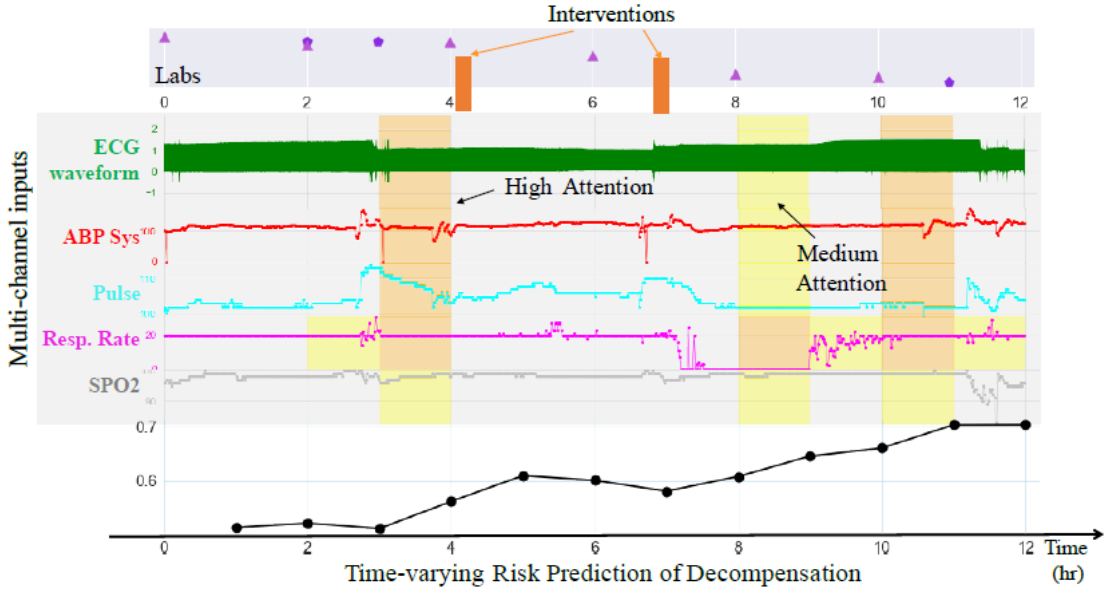


Figure 3.5: Decompensation prediction using the RAIM-3 model (Taken from Xu Yanbo et al. [45])

3.4 Previous work from the group

Fonseca et al. [11] developed two important studies about TBI outcome. The two studies focused on predicting mortality in TBI, although one of them focused on Pediatric TBI, which most studies disregard, despite being one of the most common age groups in TBI patients, as previously mentioned in subsection 3.1.

3.4.1 Mortality prediction using Sequential Time Series Electronic Health Records

The main goals in Fonseca et al. work [11] was to predict TBI mortality and evaluate the features values, so that the models produced in the study have a significant clinical value.

The dataset used in this study was the MIMIC-III and was focused on EHR, in order to use this data as a sequential time series with deep learning algorithms. In terms of feature selection and analyzing those features, the TSFresh [6] package for python was used, since the multivariate nature of the time series difficult the use of traditional methods [11].

The metric used for the classification and comparison of models was the AUROC. Fonseca et al. [11] compared the LSTM and Transformers model with the original features and the models with the added features. The results are visualized in figure 3.6

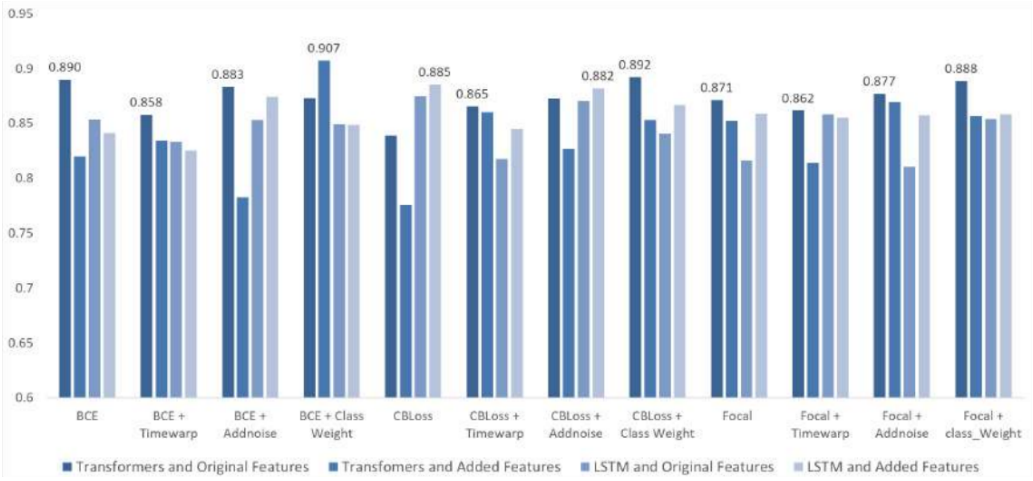


Figure 3.6: Final AUROC results (Taken from Fonseca et al [11]).

The results are combinations of both models with original and added features, the three loss methods and balancing strategies. The results indicate that Transformers perform better than LSTM with the original features, and LSTM perform better with the added features, in general. A statistical T-test was performed to have a better comparison of the AUROC values, and it was concluded that the added features did not improve the results. The best AUROC result was obtained with the Transformers, reaching a value of 0.907 AUROC with the added features.

3.4.2 Mortality prediction using Pediatrics Electronic Health Records

The second study of Fonseca et al [10] focuses on using EHR data for TBI outcome prediction in the pediatric cohort. The dataset used is the Hackathon Pediatric Traumatic Brain Injury (HPTBI) [34], containing 300 patients. Moreover, the study also tries to quantify and qualify the predictive value of the features related to TBI.

For preprocessing, One Hot Encoding (OHE) was used to transform binary variables, and in terms of dealing with imbalanced data, Synthetic Minority Oversampling Technique (SMOTE) [5] was used. In terms of feature selection methods, Koehrsen's Feature Selector (KFS) [24] was the method used. Then, PCA and Independent Component Analysis (ICA) dimensionality reduction was utilized to verify if the method outperform simpler approaches.

Since there is not much knowledge about Pediatric TBI in Literature, baseline methods were implemented [10], such as neural networks, methods based on decision trees and clustering.

The predictive impact of the features in the models is an important factor in explaining the models and giving insights for clinical strategies. Figure 3.7 visualizes the 10 most important features, according to KFS.

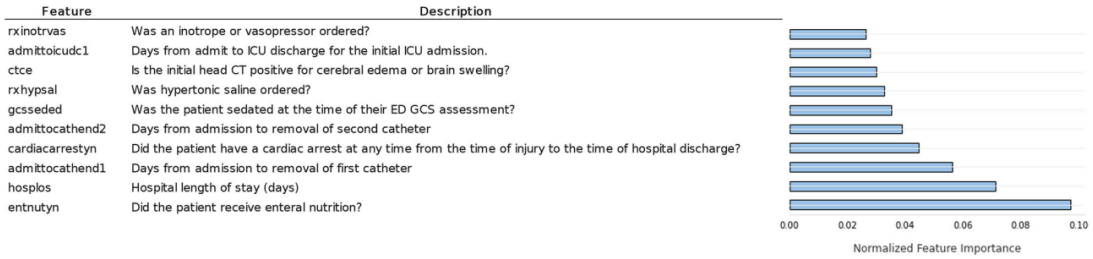


Figure 3.7: Normalized feature importance computed by the KFS method (Taken from Fonseca et al [10]).

Fonseca et al. [10] used five-fold CV and the main evaluation metric used was AUROC. Also, the pipeline used for the models is run for 50 times and the results are averaged, to avoid biased values. The best AUROC results were obtained by XGBoost (0.91 AUROC) and K-Nearest Neighbour (KNN) (0.90 AUROC). XGBoost has better performance when there is no feature selection, while KNN perform better when the KFS method is used.

3.5 Summary

Current challenges in TBI studies in terms of datasets include the fact that they are usually from a single region or center, which in turns makes the models biased towards that region and worse for predicting another regions. The same applies to low income and high income countries, where models from high income countries are not good at predicting cases from low income countries, as the conditions are different. The datasets should also have a decent size in order to be good cohorts. Other problems that arise with the datasets, apart from the size, is the amount of missing data or imbalanced data that occur in medical datasets. A lot of laboratory results or other time dependent variables could be missing, and imputation strategies should be carefully studied to each case. Data imbalance is very common in cases such as TBI, since the amount of patients that survive versus the amount of people that die have very different percentages.

Age is a factor that is commonly disregarded, although pediatric cases are one of the most common cases of TBI. One of the reasons for this is that the pediatric and adult physiology of the brain changes drastically. Therefore, TBI models that focus on Pediatric predictions are very few. Fonseca et al. [10, 11] developed models with time series data and models for Pediatric TBI, which is a very unexplored area of TBI.

The most commonly used models found were all sorts of Artificial Neural Networks, Logistic Regressors, LSTM and Support Vector Machine (SVM). CRASH [7] and IMPACT [38] are the most popular studies and models for TBI prediction. Kim et al. [21] provided a novel approach with unsupervised learning, in order to detect endotypes in patients.

Although decompensation is still not as explored as in-hospital mortality, decompensation methods are being developed in various ways. Harutyunyan et al. [16] developed methods for

decompensation prediction based on 24 hour mortality prediction at each hour of ICU, obtaining a value of 0.911 AUROC. On the other hand, Xu Yanbo et al. [45] developed the RAIM-3 model, based on CNN and RNN models that combined enable a multi-attention mechanism for lab and intervention data, utilizing wavelength and vital signs and obtaining a AUROC score of 0.9018.

Chapter 4

Decompensation prediction using Sequential Time Series Electronic Health Records

In this chapter, the main objective is decompensation prediction utilizing the resources of the Medical Information Mart for Intensive Care (MIMIC) database and the information contained within EHR.

Decompensation, is characterized by the sudden deterioration of a patient physiological state, and provides challenges in clinical practice. With the help of MIMIC-III data, we aim to create predictive models that enable early identification of decompensation events. Decompensation prediction was characterized in this study as a 24 hour mortality prediction at each hour of a patient in Intensive Care Unit (ICU). By monitoring the occurrence of mortality events at regular intervals, we gain insights into the progression and severity of a patient condition and allows for the identification of critical periods.

4.1 Methodology

4.1.1 Dataset

The chosen dataset for this task was the MIMIC-III dataset [1]. Since the dataset is public available and has a huge amount of data, it allows the group to compare the results with various other studies that also use MIMIC-III. The only requirement for the use of the dataset is a data use agreement.

The dataset consists of 57,786 admissions belonging to 46,476 adult patients between 2001 and 2012 [19]. MIMIC contains data from two different sources, CareVue and Metavision. Although this allows for more data to be available in MIMIC, it brings further complications. Some

records are duplicated, others have different units of measurements. Due to this, data engineering techniques were applied to turn the data useful for the models.

In order to use the dataset to the study of Traumatic Brain Injury (TBI), the cohort of the dataset had to be filtered. The first criteria was age, due to the complexity of TBI in pediatric ages, only patients above 18 years were selected. Next, the MIMIC-III cohort had to be narrowed down to only patients suffering from TBI. The best way to do that was to use the International Classification of Diseases (ICD) 9 codes. The TBI code diagnoses are represented in table 4.1, according to Centers for Disease Control (CDC) [12].

Table 4.1: TBI ICD-9 Codes (Taken from Wisconsin Department of Health Services [31]).

ICD-9 Code	Description
800.00 - 801.99	Fracture of the vault or base of the skull
803.00 - 804.99	Other and unqualified or multiple fractures of the skull
850.0 - 850.9	Concussion
851.00 - 854.19	Intracranial injury, including contusion, laceration, and hemorrhage
950.1 - 950.3	Injury to the optic chiasm, optic pathways, or visual cortex
959.01	Head injury, unspecified
995.55	Shaken infant syndrome

A plot of the most frequent diagnosis for the TBI patients was performed in Figure 4.1, in order to check if the cohort seemed accurate with the disease, as well as a graph in Figure 4.2.

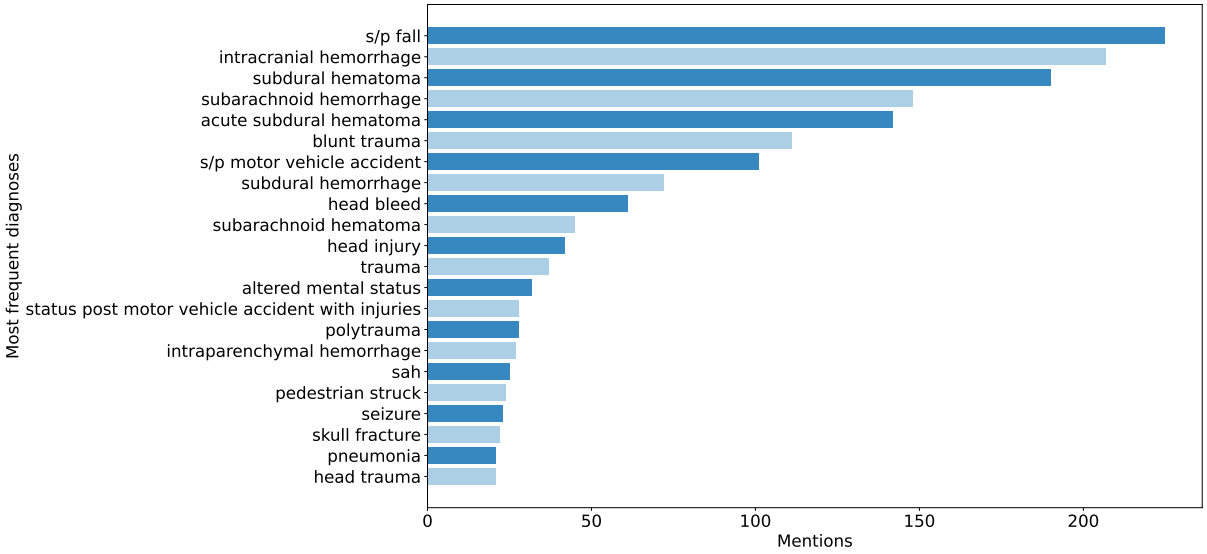


Figure 4.1: Most frequent diagnoses among TBI patients.

Figure 4.2 shows a network that corresponds to the projection of the most common diagnoses among TBI patients in the cohort. This network was produced to show how diagnosis cluster

across TBI patients and was obtained by performing a network projection of the diseases in the cohort shown in figure 4.1. Therefore, it shows, not only the diseases that were diagnosed, but also the relations between those diseases for each patient. Note that a patient can have more than one diagnose. This network shows that the nodes with the highest betweenness centrality are nodes that connect the different communities of diseases, represented in different colors.

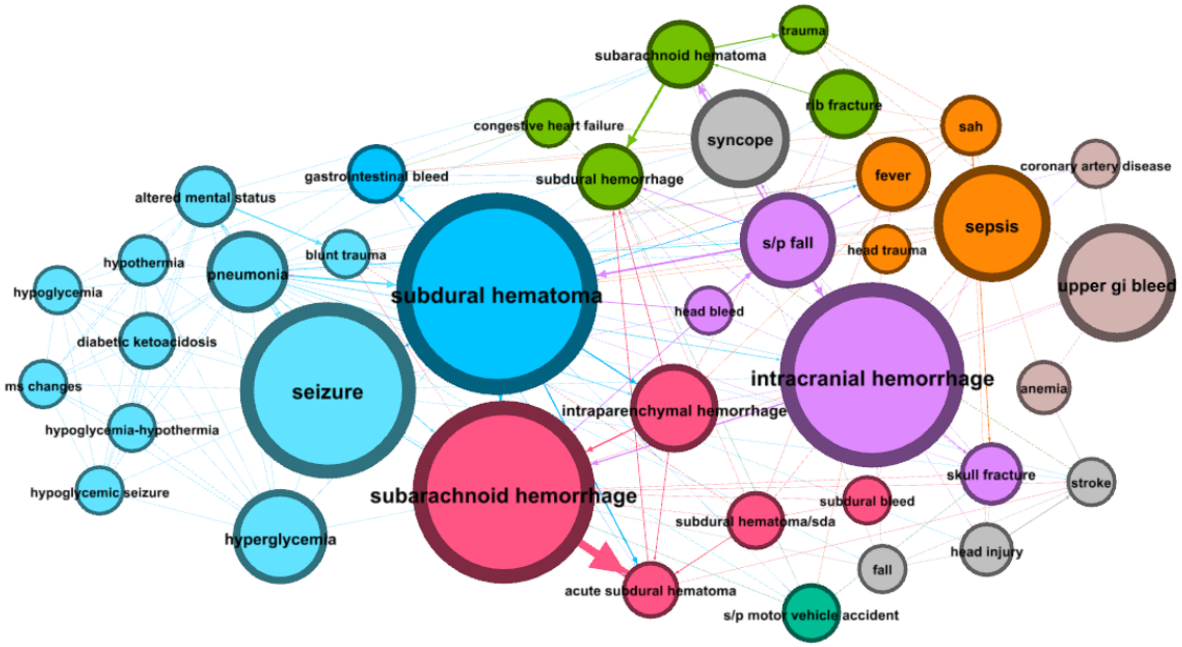


Figure 4.2: Diagnoses Network. This network corresponds to the network projection of all diagnoses for TBI patients. It allows to see the diagnoses that are associated.

After this patient filtering, multiple other steps of data cleansing and preprocessing had to be performed in order to get a coherent dataset for TBI. Following Harutyunyan et al. [16] and previous work from the group [11] guidelines, patients with hospital admissions with multiple stays and/or transfers between ICU units were removed, since it could introduce ambiguity between different admissions. For the clinical events, events that cannot be matched to ICU stays are also discarded. Next, time series of the events for each patient can be extracted [11, 16].

After these preprocessing steps, it is possible to create the decompensation prediction task. Since the logic of physiologic decompensation applied in this approach is to predict 24-hour mortality at each hour of a patient stay, a binary label is created that states if a patient death is detected in the next day of every time point in the time series, and also guaranteeing that the series have over 24 hours, excluding stays with missing Lenght of stay (LOS), LOS less than 24 hours, and stays with no events in those first 24 hours. Figure 4.3 shows a diagram that describes the process described above.

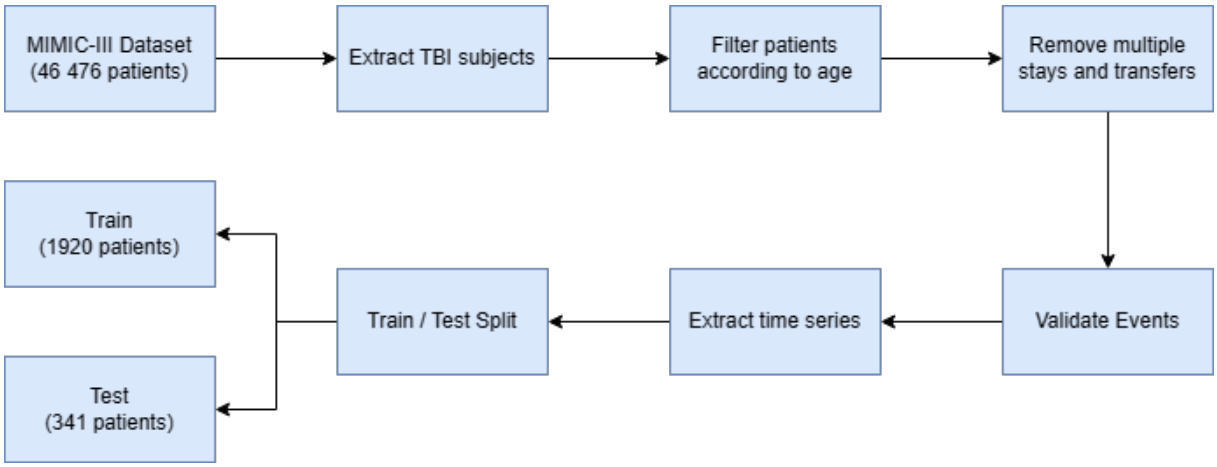


Figure 4.3: Pre-processing diagram following Harutyunyan et al. benchmarks [16]. From the initial 46,476 patients in the dataset, TBI patients were filtered according to age and ICD-9 code. Multiple stays and transfers were removed to prevent ambiguity. The events are also validated, and after that the timeseries for each patient are extracted, removing events that do not contain at least 24 hours minimum.

4.1.1.1 Cohort description

In total, 2261 patients were in the TBI cohort, where 1567 are for training, 353 for validation and 341 for testing. For each patient, with the creation of prediction tasks for every hour of the patient stay, a total of 219,688 samples are created.

Table 4.2: Cohort summary.

Variable	**Value**
Number of patients (N)	2261
Age (mean years ± sd)	58.1 ± 21.2
Gender of patients (M/F)	61.7% / 38.4%
Length of stay (mean days ± sd)	6.4 ± 5.8
Mild TBI (%)	47.616%
Moderate TBI (%)	14.937%
Severe TBI (%)	37.437%
Mortality Rate (%)	10.45%

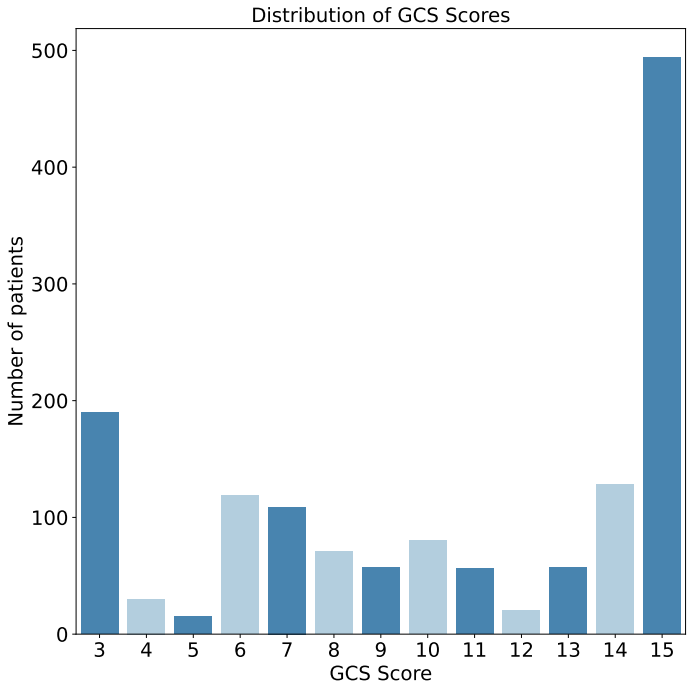


Figure 4.4: GCS total scores for patients at hospital admission.

Table 4.2 shows the summary of the cohort used in this study. The TBI categories according to Mild, moderate and severe are grouped by their definition: mild TBI ranges from GCS scores between 13 and 15, moderate between 9 and 12 and severe from 3 to 8. The distribution of these results are shown in figure 4.4.

4.1.2 Variable Selection

Although MIMIC-III contains a big amount of data and variables, many of them are very sparse and therefore, are not suitable to be used in this work. When extracting the time series for each patient, a set of variables were selected with this barrier in mind. The first step was to follow guidelines by Harutyunyan et al. and Fonseca et al. [11, 16], described in table 4.3. More variables are explored, where 30 other features are selected, chosen by the number of occurrences in the MIMIC-III dataset and according to previous work [11]. These extra features are also considered relevant in the IMPACT [38] and CRASH [7] studies mentioned before.

In the original features in table 4.3, all but the last feature appear in a good amount. The features that have more occurences are related to Physiologic time series data (PTS) features. Height was expected to have less occurences, since it is not always measured, and when measured it is only registered once in a patient stay.

Table 4.3: Original features in the dataset. The count corresponds to every time point where the value is registered in the TBI cohort.

Original Features (17)	**Count**
Respiratory Rate	278,699
Heart Rate	271,975
Oxygen saturation	270,802
Systolic blood pressure	244,170
Diastolic blood pressure	244,100
Mean blood pressure	242,559
Glasgow coma scale eye opening	97,025
Glasgow coma scale verbal response	96,929
Glasgow coma scale motor response	96,759
Temperature	69,631
Glasgow coma scale total	56,854
Glucose	51,434
Fraction inspired oxygen	36,605
pH	21,104
Weight	6,222
Capillary Refill Rate	2,006
Height	441

Table 4.4: These features correspond to the added features to the 17 original features. The count represents the number of time points where the variable is present.

Added Features (30)	**Count**	**Added Features (30)**	**Count**
Urine output	153,733	Calcium	11,926
Pupillary response left	49,110	Chloride	13,710
Pupillary response right	48,833	Creatinine	13,131
Pupillary size left	49,336	Hemoglobin	11,998
Pupillary size right	49,208	Magnesium	13,564
Urine Color	34,053	Mean corpuscular hemoglobin	11,834
Urine Appearance	33,977	Mean corpuscular hemoglobin concentration	11,955
Potassium	19,177	White blood cell count	11,927
Hematocrit	16,026	Red blood cell count	11,837
CO2 (ETCO2 PCO2 etc.)	15,598	Mean corpuscular volume	11,832
Partial pressure of carbon dioxide	15,487	Partial pressure of oxygen	10,753
Sodium	15,213	Partial thromboplastin time	7,807
Anion gap	13,069	Prothrombin time	7,698
Bicarbonate	13,140	Positive end-expiratory pressure	6,491
Blood urea nitrogen	13,110		
Phosphate	12,398		
Platelets	12,276		

When comparing to the added features in the dataset in table 4.4, there is no relevant feature

with a number of occurrences that could be challenging for the methods.

4.1.3 Pre-processing

After selecting the variables in the original features and in the added features, that data still requires some processing. This section aims to explain the preprocessing steps necessary to use the TBI data extracted from the MIMIC cohort for the decompensation task at hand.

Some variables are in a categorical format, as for example the Glasgow Coma Scale (GCS) features, which require One Hot Encoding (OHE). This process turns those categorical features as binary vectors. These binary vectors are turned into N binary columns, where N is the value of each category in the feature, that will represent 1 or 0, in case of the value being positive or false, respectively. This allows the Machine Learning (ML) algorithms to include these categorical values, now converted into a numerical format. The diagram in figure 4.5 illustrates this process.

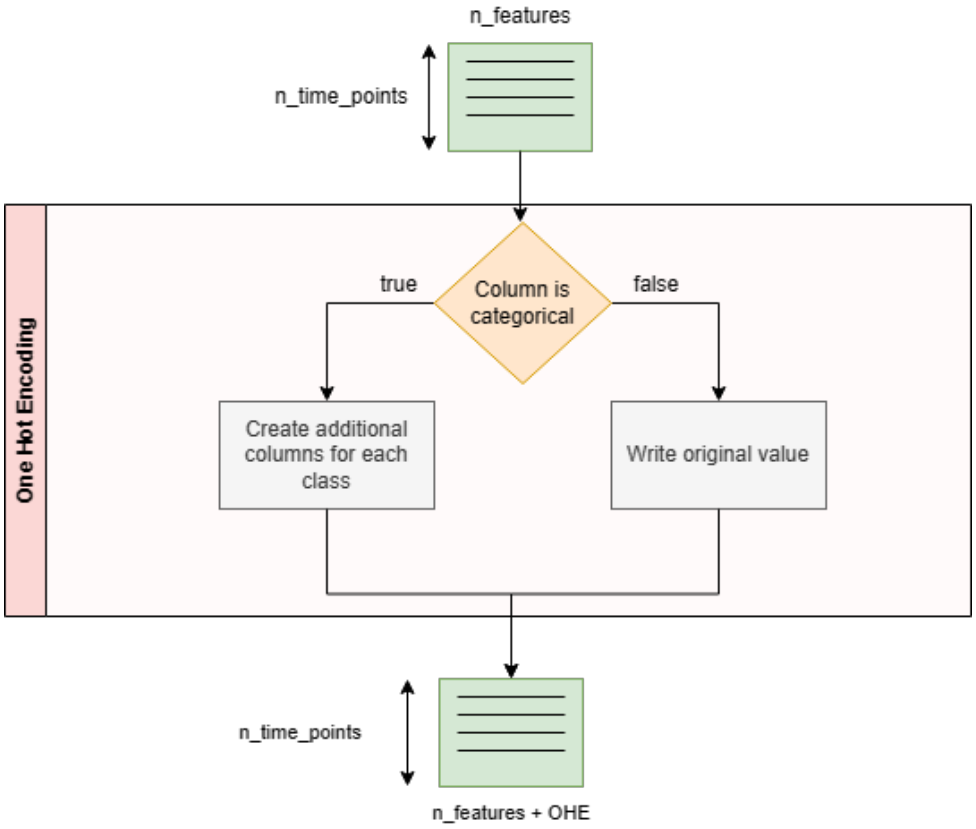


Figure 4.5: OHE diagram process. If a given column is categorical, OHE is used to turn the categorical variable into N other variables according to the number of classes in the categorical feature. If the column is not categorical, the same value is written. In the end, columns that are not categorical stay the same, while categorical columns are turned into N columns, where the values are 1 if the category is true for that value, or false otherwise.

The data also needs to be transformed into time points for each hours in the ICU stay of the

patients. This discretization will convert the data into N bins, each bin representing a time point in the timeseries for each hour of the ICU, until the patient dies or is discharged.

When there is only one value in that hour, that value is imputed into the respective bin. Since there are sometimes more than one value for each hour, only the last value for each hour is considered. When there are missing values in that hour, other imputation strategies are performed, the normal value of the variable is imputed. The diagram in figure 4.6 illustrates this process.

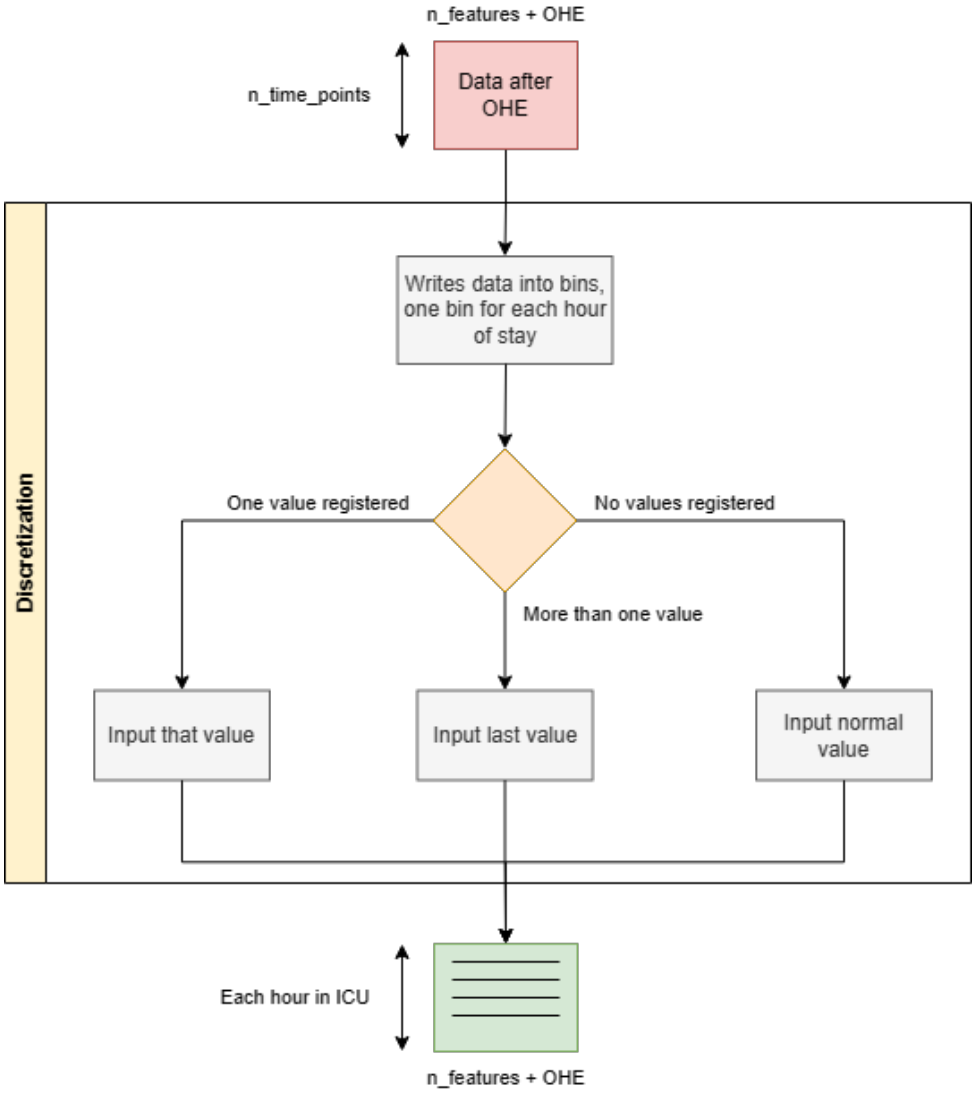


Figure 4.6: Discretization process. The discretization receives as input the data processing from the OHE, where each column is now numerical and not categorical, and groups the data into bins. These bins represent each hour of the stay for each patient. Then, an imputation strategy is performed. For hours where there are more than one value registered, only the last value is imputed. For hours where there is only one value, that same value is used. When there is no value in the hour, normal imputation of that column is used. In the end, the data is discretized into N bins with no missing values.

After OHE, discretization and imputation, the data is normalized. That process was performed using Z-score normalization. The normalization is applied for each column, where for each value the mean is subtracted, and then divided by the standard deviation. The formula for that method is provided in formula 4.1.

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

Where z is the new value, x is the original value, μ is the mean of data and σ is the standard deviation.

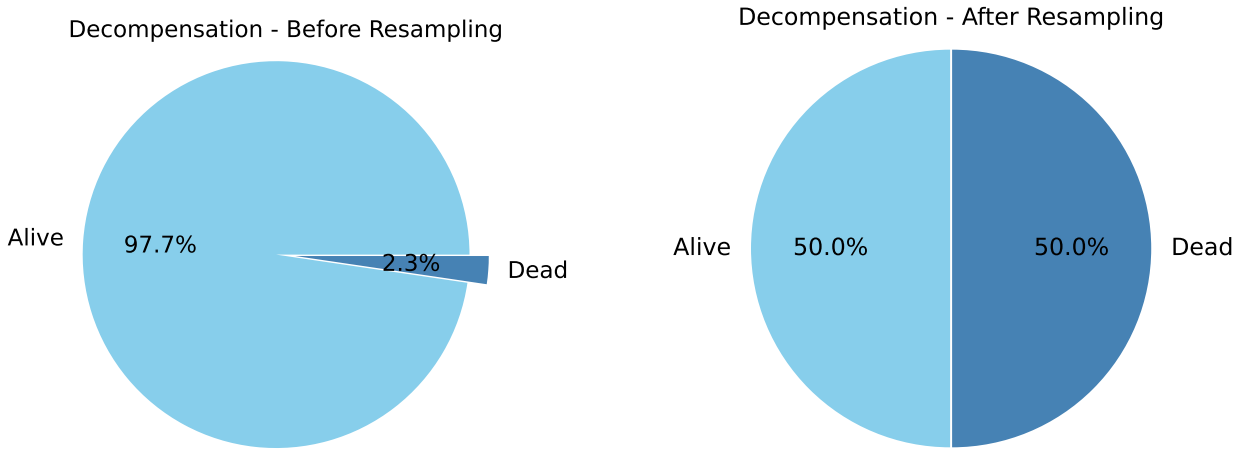
4.1.4 Data Imbalance

One major issue that MIMIC-III and most clinical datasets encounter is the imbalance of data in the predicted label. The TBI cohort used contained 10.45% of mortality detected, in contrast with 89.55% alive patients. When creating samples for the cohort, where a sample for each hour of the patient was created, the total samples are 219,688. These samples suffer from an even bigger difference in classes: in the training data, only 2.3% of labels were categorized as dead and 97.7% as alive. Big differences in the classes proportions usually lead to models overfitting to the majority class. In this case, the majority class is the alive class, which would be overfitted and the models would almost always predict each sample as alive.

To fight this problem, class weights and oversampling were tested, from the knowledge of previous studies from the group by Fonseca et al. [11], where various loss functions were experimented, as well as oversampling techniques and class weights.

Some different class weights were used: 0.023 and 0.977, corresponding to the proportions of the classes; 1 and 2, 1 and 3, were also experimented, giving the minority class a weight double and triple of the majority, respectively.

Oversampling techniques were also used. Due to the nature of decompensation and how we are predicting it, considering decompensation as a 24-hour mortality prediction based on each hour of the ICU, a different technique was tried, because there are various samples belonging to the same patient. Due to this, when oversampling the minority class, we did not only oversample samples for the minority class, but every sample from that patient, when there was at least one sample belonging to the minority class in that patient. The problem with this approach is that a huge amount of oversampling would be necessary, due to the fact that patients that do in fact decompensate, can have multiple other samples of no mortality detected. This means that even when applying oversampling in this way, more samples from the majority class would be also created. With that being said, oversampling was performed with this notion of decompensation, as well as Synthetic Minority Oversampling Technique (SMOTE).



(a) Original mortality distribution in TBI cohort

(b) Mortality distribution after resampling

Figure 4.7: Decompensation distribution in the TBI cohort before and after resampling techniques.

4.2 Classification Methods

4.2.1 Long-short term memory (LSTM)

Long-short term memory (LSTM) models were introduced in 1997 by Hochreiter and Schmidhuber [18] and are designed to avoid long-term dependency problems. LSTM excel at capturing temporal dependencies in sequential data, making them well-suited for predicting complex dynamic processes like decompensation, since LSTM models can capture both short-term and long-term dependencies, allowing for more accurate predictions of decompensation episodes.

Preceding the LSTM layer, a masking layer and bidirectional layer are present. The masking layer handles and is responsible for enabling the model to disregard time points in the timeseries that do not contain any registered value within the sequence. The bidirectional layer allows the model to process the sequence in both forward and backward directions, to primarily enhance the model understanding of temporal dependencies and capture contextual information from both past and future timesteps.

In the LSTM layers, there are three important aspects that allow for the long-term dependencies to be preserved. The cell state carries information from previous timesteps and stores information, with the help of the input and forget gates [46], corresponding to equation 4.2 and equation 4.3, respectively. The forget gate determines what information from the previous cell state should be discarded, while the input gate decides what new information should be added [46]. The output gate determines which information from the cell state should be exposed as the output at each time step, corresponding to equation 4.4. Note that the layer size, as well as the depth of lstm are hyperparameters tuned during the training process.

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \tag{4.2}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.3}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.4}$$

Where i_t represents the input gate, f_t the forget gate, and o_t the output gate. w_x represents the weights of the respective gates x neurons. h_{t-1} represents the output of the previous lstm block, x_t the input at current timestep, and b_x the biases for gates x.

Subsequently, following the LSTM layers, comes the dropout and dense layers. The dropout layer mitigates the risk of overfitting, while the dense layer transforms the output of the LSTM layer into a meaningful prediction. Figure 4.8 illustrates the LSTM architecture used.

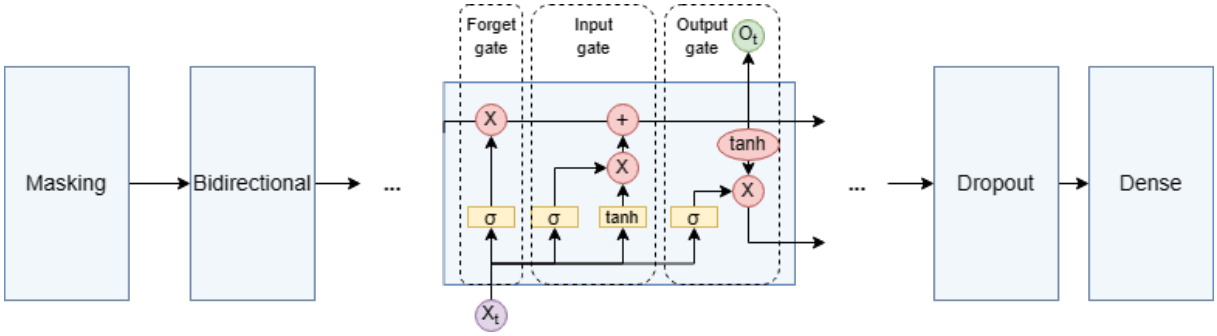


Figure 4.8: LSTM architecture. The masking layer allows the model to ignore timepoints with no values registered. Bidirectional layer allows for forward and backwards to capture more information. Then, the core of the LSTM, with the forget, input and output gates, that enable long-term dependencies. Dropout helps with overfitting, and the Dense layer provides the final prediction (LSTM layer adapted from Yu Yong et al. [46])

4.2.2 Transformer Encoder

Transformer encoding has emerged as a novel architecture that captures complex patterns in sequential data. Unlike Recurrent Neural Network (RNN), transformers use a self-attention mechanism: a component that efficiently captures both short and long-range dependencies within the input sequence space. Self-attention allows the Transformer to extract contextual relationships effectively, contributing to its overall success in modeling sequential data.

Figure 4.9 represents the architecture used in this study. The primary component is the transformer encoder block, which is the core of the model and contains the most important aspects of the architecture. It contains the novel multi-head attention, layer normalizations and residual connections, which can useful for stabilizing the weights and facilitate convergence during training. The dropout layers help preventing overfitting. Convolutional layers are used to capture local parameters and spacial relationship in the input data. The global average pooling is used after the transformer encoder block and provides a more contextual representation of

the previous sequences, which can then be passed onto the Multilayer perceptron (MLP) layer, which takes the vector of features from the previous layer, to then pass it to the dense layer. Note that the depth of the encoders, number of heads, dropout size, dense layers, MLP units are all hyperparameters tuned during the training step.

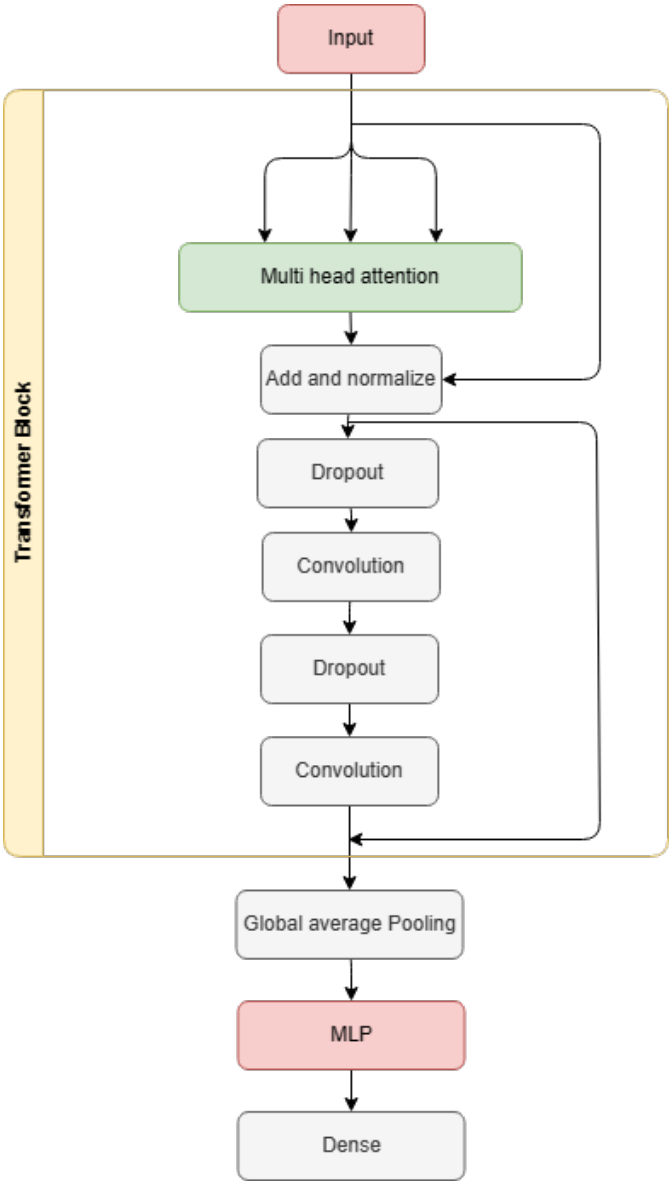


Figure 4.9: Transformer architecture. The transformer encoder block incorporates the multi-attention head, convolutions, normalization and dropout layers. Following this block, global average pooling is performed, followed by MLP and a dense layer for the prediction.

4.2.3 Logistic Regression

Logistic Regressor (LR) is a statistical modeling technique used for binary classification. In contrast with linear regressions, which predicts continuous numerical values, LR predict the probability of an event. The output is a value between 0 and 1, that represent the probability of

the event being in one of the classes.

Figure 4.10 represents the LR used. Net input functions represents the linear combination of the input features and their weights, represented in equation 4.5 The sigmoid activation function maps those linear combinations into a value between 0 and 1 (equation 4.6). The threshold is then a decision based on the value from the sigmoid function that turns into a binary decision. A grid search was used during training to find the best hyperparameters for penalties and coefficients.

$$z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{4.5}$$

Where z is the net input, and $w_0 ... w_n$ are the weights associated with the input features $x_1 ... x_n$

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4.6}$$

Where z is the input of the sigmoid function.

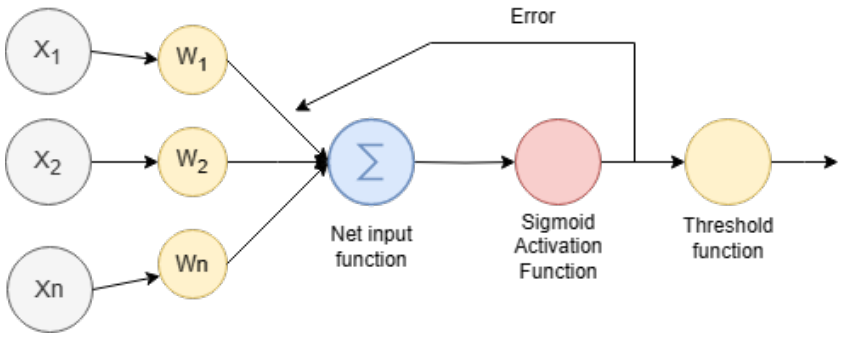


Figure 4.10: LR architecture. Net input function represents the linear combinations of inputs and respective weights. Sigmoid maps them into values between 0 and 1. Threshold turns these values into a binary prediction.

4.3 Summary

In this chapter, the TBI cohort extracted from the MIMIC-III dataset was explored, as well as the variables for the original and added features set. The pipeline to extract the relevant data from MIMIC-III was also explored, with preprocessing steps needed such as OHE to convert categorical values into a numerical format; the discretization process necessary to transform the data into a timeseries, where different techniques were used according to the registered value in each bin.

Data balancing techniques were also explained, with different class weights values explored, such as double, triple and proportional to the dataset imbalance; oversampling, including SMOTE

and oversampling according to decompensation, where every patient classified as decompensated would be included in the method.

Classification methods used were LSTM, Transformers and Logistic Regressors. LSTM architecture consists of a masking layer and a bidirectional layer, LSTM layers, dropout and dense layers. Transformers architecture consist of a transformer encoder block with a multi head attention layer, layer normalizations, dropout and convolution layers. Following the encoder block, a global average pooling, MLP and dense layer. Logistic regression consists of a net input function, sigmoid activation function and a threshold function.

Chapter 5

Results and discussion

This section focuses on presenting the results obtained in this study, as well as a discussion about the values obtained. The models used were Long-short term memory (LSTM), Transformers and Logistic Regressor (LR), with the architectures previously described.

5.1 Overview of model performance metrics

The Area Under the Receiver Operating Characteristics Curve (AUROC), balanced accuracy, precision and recall were the metrics used to measure the models performance. Since the dataset is highly imbalanced, precision and recall are also going to be computed for each class (0 and 1) individually, in order to compare these metrics for both classes between models.

5.2 Hyperparameter tuning

A search for the best parameters was performed using randomized search, before training the models, in order to find the best hyperparameters. Besides that, the models are also trained for different situations, such as the data with original features and the added features; class weights and oversampling. The results seen in table 5.1 were performed on a train with 10 epochs.

Table 5.1: Hyperparameters tuning according to parameters and their respective ranges for LSTM, Transformers, and Logistic regressor.

LSTM			**Transformer**			**Logistic Regressor**		
Parameter	Search	Chosen	Parameter	Search	Chosen	Parameter	Search	Chosen
depth	1, 2, 4	1	num_heads	1, 2, 4	2	Penalty	l1, l2	l2
dropout	0, 0.1, 0.2, 0.4	0	mlp_units	16, 64, 128	64	Coefficient	0.0001, 0.001, 0.01	0.0001
rec_dropout	0, 0.1, 0.2, 0.4	0	mlp_dropout	0, 0.1, 0.2, 0.4	0.2			
learning_rate	0.01, 0.001, 0.0001	0.0001	head_size	64, 128, 256	64			
			ff_dim	2, 4	2			
			dropout	0, 0.1, 0.2, 0.4	0.2			
			depth	1, 2, 4	4			
			learning_rate	0.01, 0.0001, 0.00001	0.0001			

5.3 Classification Results

The final test results for original features and added features are presented in table 5.2 and table 5.3, respectively. The models were trained according to the hyperparameters chosen in table 5.1 and for a maximum of 50 epochs with early stopping criteria: if the AUROC score had obtained 0.85 or higher and decreased below 0.82, the training would stop. The baseline models were performed with the original features and no class weights or oversampling, for LSTM, Transformers and Logistic Regressor. The imputation strategy in these baselines was the normal value imputation, since it was the one chosen for the remaining model trainings, due to better performance. The model performances that are going to be compared with these baselines are the strategies of feature selection and data imbalance, to try and improve these results when compared with the baseline.

5.3.1 Classification using original features

Table 5.2 represents the test values obtained for the decompensation classification using the original features set. Our model performance for the baseline, with no added features and no class imbalance techniques, obtained a AUROC result of 0.8653 with the Transformer architecture. The best value was obtained using the Logistic Regressor, with an AUROC value of 0.9787. This value although very positive, also looks very misleading and can probably indicate that the model could have performed very well in this test set but does not have good generalization. The best LSTM result was obtained using no weights or oversampling, with a AUROC of 0.91879. Although the techniques used to try and improve LSTM results ended up in a lower AUROC score, metrics such as recall and balanced accuracy changed. The high value for the LSTM result with no weights and oversampling is the imbalance of the data, which made the model predict almost only the majority class. The addition of weights and oversampling, although resulting in lower AUROC, also resulted in the recall and balanced accuracy of the model to improve, since the model was capable of predicting for the minority class too. As the rate of true negatives was increased with these changes, so was the rate of false positives. Although the rate of false positives was relatively high, the rate of false negatives was reduced drastically. In a real scenario, this means that there were some patients identified as in risk of decompensation, when in fact they were not, resulting in higher expenses for the hospital, but at the same time the risk of a patient suffering from imminent decompensation being disregarded from the model decreased significantly, allowing patients in need of treatment to be correctly identified.

While with LSTM the model decreased in performance with the changes applied, Transformers suffered from the same problem related to imbalanced data as seen in the precision and recall for class 1 in the baseline transformer, with 0.0 registered for both. But, in this case, the Transformers increased in performance with the use of weights and oversampling, improving from 0.8653 to 0.8788 AUROC. The precision and recall for the minority class also improved a lot with the use of class weights.

Table 5.2: Metric results for original features for LSTM, Transformer and Logistic Regressor.

Model	Methods	Bal Acc	Prec 0	Prec 1	Rec 0	Rec 1	AUROC
LSTM	No weights / No oversamp.	0.6112	0.9873	0.2789	0.9900	0.2325	0.9179
	Class weights	0.7608	0.9933	0.0793	0.8756	0.6447	0.8856
	Oversampling	0.7360	0.9920	0.1001	0.9171	0.5548	0.8871
Transformers	No weights / No oversamp.	0.4999	0.9837	0.0	0.9999	0.0	0.8653
	Class weights	0.7435	0.9926	0.7947	0.8840	0.6031	0.8630
	Oversampling	0.7438	0.9931	0.0614	0.8341	0.6535	0.8788
Logistic Regressor	No weights / No oversamp.	0.8281	0.9966	0.0125	0.8228	0.8333	0.9223
	Class weights	0.8187	0.9996	0.0450	0.6527	0.9846	0.9069
	Oversampling	0.9347	0.9990	0.1757	0.9265	0.9430	0.9787

5.3.2 Classification using added features

The final test results for the added features are present in table 5.3. For LSTM models, the AUROC value for no oversampling and no weights is 0.9035. The same discrepancy in recall from class 0 and 1 can be seen in this model, as there were no methods to prevent this factor. While with the original features, the AUROC values for LSTM models using class weights and oversampling techniques was similar, with the added features there is a noticeable difference, with class weights method yielding 0.9204 and oversampling with 0.9102 AUROC.

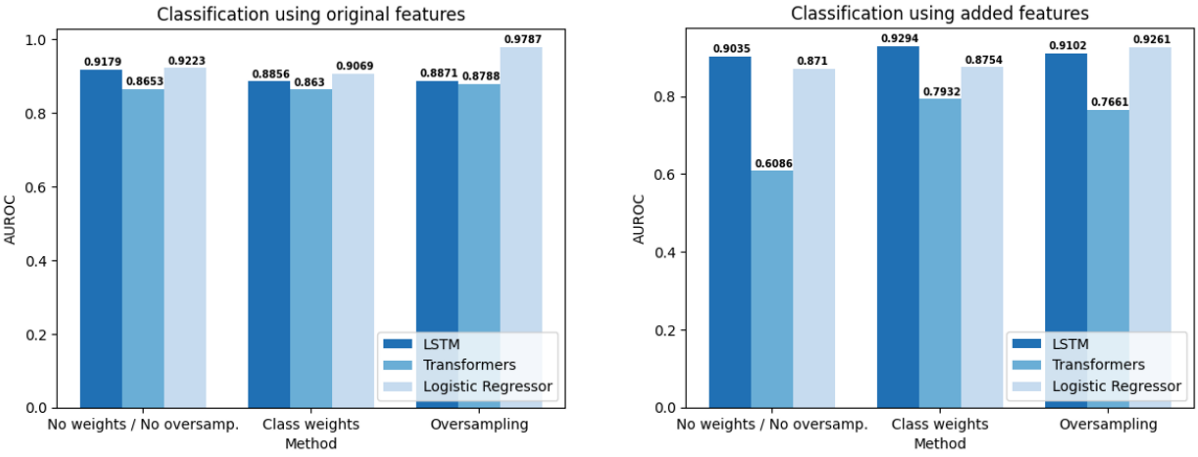
In terms of Transformers, the results were much worse than using only the original features. While LSTM were able to improve results with the added features, Transformers decreased in performance. The Transformer model with no imbalance techniques was not able to discriminate between the two classes and obtained a AUROC of 0.6086. With class weights and oversampling, the results increased by a lot, with AUROC of 0.7932 and 0.7661, respectively. Although this is a big increase in performance compared to the initial value of 0.6086, the values below 0.8 AUROC do not prove very useful when compared with the other values obtained by LSTM. Therefore, it seems that the added features were helpful with LSTM models, but decreased transformers performance by a lot.

Logistic regressors seemed to have slightly worse performance when compared with the original features. The use of no weights and oversampling and the use of class weights have very similar performance and the oversampling technique obtained a better AUROC with value 0.9261. Although this is the best value for logistic regressor, the precision class 0 and recall class 1 obtained values of 1.0, which could be another hint of not very good generalization, since the same seems to happen in the original features results for this model.

Table 5.3: Metric results for added features for LSTM, Transformer and Logistic Regressor.

Model	Methods	Bal Acc	Prec 0	Prec 1	Rec 0	Rec 1	AUROC
LSTM	No weights / No oversamp.	0.5274	0.9845	0.9259	0.9999	0.0548	0.9035
	Class weights	0.8205	0.9951	0.1175	0.9086	0.7325	0.9294
	Oversampling	0.8064	0.9946	0.1116	0.9066	0.7061	0.9102
Transformers	No weights / No oversamp.	0.5	0.9837	nan	1.0	0.0	0.6086
	Class weights	0.7834	0.7386	0.8394	0.8459	0.7189	0.7932
	Oversampling	0.6946	0.7014	0.6915	0.7399	0.6493	0.7661
Logistic Regressor	No weights / No oversamp.	0.7769	0.9863	0.1470	0.8098	0.7441	0.8710
	Class weights	0.7908	0.9895	0.1332	0.7663	0.8152	0.8754
	Oversampling	0.8980	1.0	0.1776	0.7960	1.0	0.9261

The AUROC results obtained in table 5.2 for the original features and table 5.3 for added features are also represented in figure 5.1.



(a) Classification results using original features represented in table 5.2.

(b) Classification results using added features represented in table 5.3.

Figure 5.1: Classification results comparison for original and added features according to the methods explored.

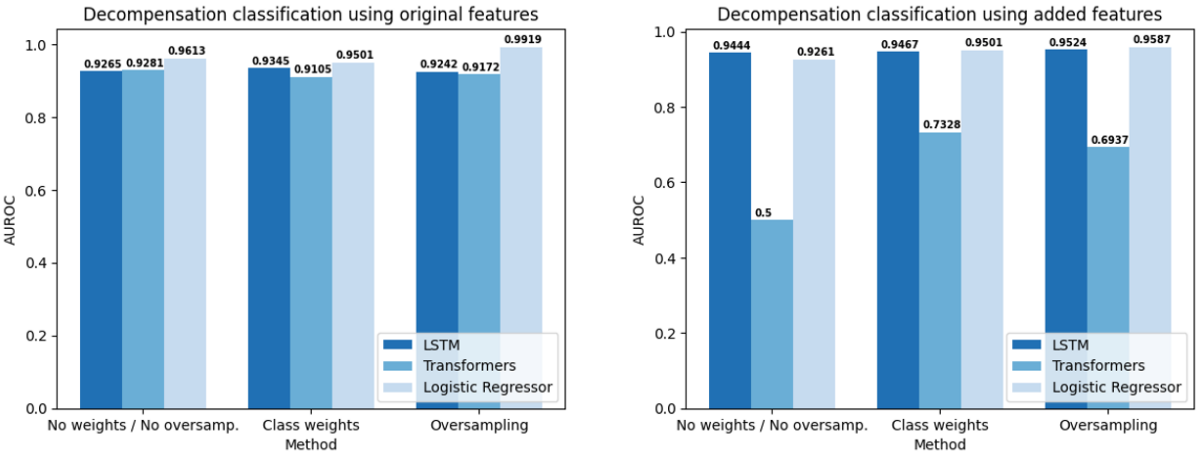
5.3.3 Decompensation classification

A different approach is also going to be used to compare the models. The notion of decompensation is going to be used to calculate different results. Not only the metrics are going to be calculated based on each sample, but also on each patient. This technique was used to check if a patient that actually had a decompensation event, was predicted correctly that at some point during its stay, the model detected decompensation. Therefore, to achieve this, if the model predicts that a sample from a given patient is a decompensation event, then the patient is calculated in this new metric as positive for decompensation, and false otherwise.

Table 5.4 represents the AUROC values adapted to decompensation for each model trained in original and added features from table 5.2 and table 5.3, respectively. Note that these values are results obtained from previous results, being used only to get a better overview of the decompensation task prediction, instead of the previous values considering each sample as individual for all patients. For LSTM, all results showed a very high AUROC value, ranging from 0.9242 to 0.9524. The best result was obtained with the added features and oversampling, followed by added features and class weights. In terms of transformers, the values from original features were relatively high too, while with the added features the values were very poor, also seen in the previous test results. Logistic regressors had very high values, with the original features and oversampling reaching 0.9919 AUROC, demonstrating again the possible misleading values obtained with this model. These values are in accordance with the previous test results, as transformers show the loss in performance with the addition of extra features, the very high values are more emphasized for logistic regressors, and LSTM models have good results for original and added features, showing an improvement. The results from table 5.4 are also represented in figure 5.2.

Table 5.4: AUROC results using the decompensation notion.

Model	**Methods**	**AUROC**	
		Original features	Added features
LSTM	No weights / No oversamp.	0.9265	0.9444
	Class weights	0.9345	0.9467
	Oversampling	0.9242	0.9524
Transformers	No weights / No oversamp.	0.9281	0.5
	Class weights	0.9105	0.7328
	Oversampling	0.9172	0.6937
Logistic regression	No weights / No oversamp.	0.9613	0.9261
	Class weights	0.9501	0.9501
	Oversampling	0.9919	0.9587



(a) Classification results using original features.

(b) Classification results using added features.

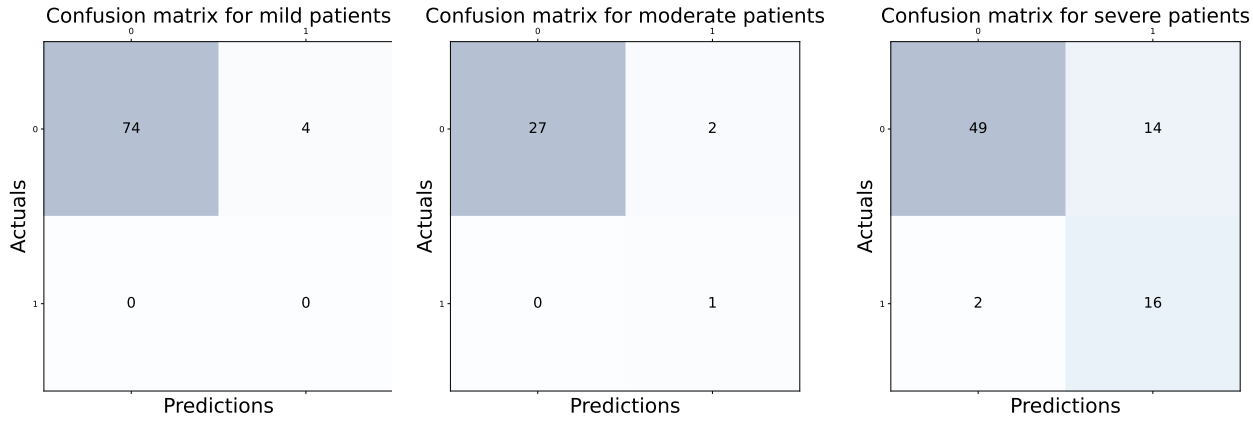
Figure 5.2: Classification results comparison for original and added features according to the methods explored and the decompensation notion represented in table 5.4.

5.3.4 Classification results by TBI category

Other performance factor was evaluating how well the best models perform for each Traumatic Brain Injury (TBI) category: mild, moderate and severe cases. The model used was the LSTM with added features and oversampling technique, presented in table 5.5. For the mild category, the values were not very descriptive of the actual results, since there was not a single result for decompensation present in the mild cohort. Figure 5.3a shows a better representation of the values predicted, where for the 78 patients present, 74 were correctly predicted as no decompensation detected and 4 were false positives. For the moderate category, there were no nan results, with a AUROC of 0.9355. In figure 5.3b, it is seen that there was only one sample in the decompensation class, that was correctly predicted, while 2 were detected as false positives. For the severe class, which had more decompensation samples, the model obtained an AUROC of 0.8960. In figure 5.3c, it can be seen that out of the 18 samples from decompensation, 16 of them were correctly detected as true positives. Out of the 63 patients with no decompensation, 49 were predicted as no decompensation and 14 as false positives. Overall, it is seen that the models are able to predict the minority class of decompensation very well, although some of the majority class are wrongly predicted.

Table 5.5: Test results for mild, moderate and severe TBI categories according to LSTM with added features and oversampling.

TBI Category	Bal Acc	Prec 0	Prec 1	Rec 0	Rec 1	AUROC
Mild	nan	1.0	0.0	0.9950	nan	nan
Moderate	0.5406	0.9882	0.3333	0.9978	0.08333	0.9355
Severe	0.8019	0.9922	0.1297	0.8780	0.7258	0.8960



(a) Confusion matrix for mild patients

(b) Confusion matrix for moderate patients

(c) Confusion matrix for severe patients

Figure 5.3: Confusion matrices for mild, moderate and severe categories of TBI. The values in the matrices correspond to the number of patients in the test cohort for each category and using the decompensation notion.

5.3.5 Feature Selection

A feature selection method was used to calculate which features are more relevant for the models and get a better understanding of their importance in decompensation prediction for TBI. This method focused on identifying which features had more relevance for the model when training, which was performed by taking into account the weights importance in the LSTM architecture. Table 5.6 presents 15 relevant features for the original and added features, ordered by importance.

Table 5.6: Relevant features from original and added features.

Relevant features from original features	Relevant features from added features
Glasgow coma scale eye opening->1 No Response	Pupillary size left->5 mm
Glasgow coma scale motor response->6 Obeys Commands	Glasgow coma scale motor response->3 Abnorm flexion
Glasgow coma scale total->4	Respiratory rate
Glasgow coma scale total->11	Glasgow coma scale total->5
Glasgow coma scale eye opening->4 Spontaneously	Pupillary size right->3 mm
Heart Rate	Urine output
Glasgow coma scale total->6	Glasgow coma scale eye opening->4 Spontaneously
Glasgow coma scale motor response->2 Abnorm extensn	Pupillary size right->6 mm
Glasgow coma scale verbal response->4 Confused	Pupillary size left->3 mm
Glasgow coma scale total->14	Glasgow coma scale total->11
Glasgow coma scale total->10	Urine Color->Orange
Systolic blood pressure	Pupillary size right->5 mm
pH	Glasgow coma scale total->4
Diastolic blood pressure	Glasgow coma scale motor response->6 Obeys Commands
Oxygen saturation	Calcium

In the relevant features from the original features set, it is seen that most of the relevant features are related to Glasgow Coma Scale (GCS) (10 out of 15 features). When comparing with the relevant features from the added features set, 10 GCS are also present, since features related to pupillary size were added for this feature set and are related to GCS eye response feature. Two urine features and calcium are also present in this set, which were also added features. Respiratory rate is the only feature from the added features set that was present in the original features set, although it was not considered relevant in that case. The clinical consideration of these features being considered as important by the models are further explained in section 5.3.8.

5.3.6 Missing data

Missing data was imputed using the normal value imputation. This approach showed good results in performance, except when using the added features for transformers, although it is unlikely that the decrease in performance for this case is due to the missing data imputation technique used.

In general, without the imputation strategies, the models could not have performed well. Missing data imputation played a crucial role for this study, where the discretizer transformed the data into bins according to each hour, and imputation strategies filled the missing data where needed, as well as impute values if more than one value was registered, being an important piece in the discretization process.

Figure 5.4 shows a visualization of the imputation strategy mentioned for Physiologic time series data (PTS) according to each step: the original values with no imputation performed; the discretization and imputation performed; and the final normalization process.

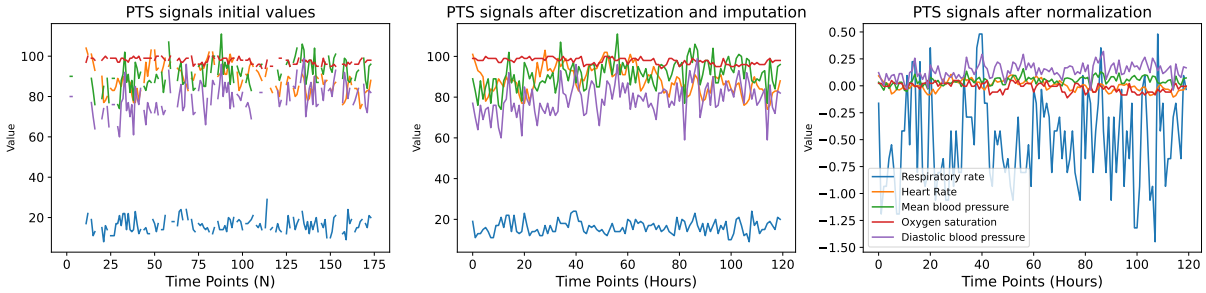


Figure 5.4: Example of the discretization process. The first plot corresponds to the PTS signals without any processing. The second figure plot corresponds to the discretization and imputation, where the missing values were imputed and the data was transformed to fit only in one bin per hour. Lastly, the third plot shows the final normalization process.

5.3.7 Data imbalance

Regarding data imbalance, where the dataset for this study was very disproportional in classes (97.7% and 2.3% for classes 0 and 1, respectively), different techniques of dealing with this issue

were tested: Synthetic Minority Oversampling Technique (SMOTE), oversampling every patient with decompensation, and class weights. During training, SMOTE did not have good results when compared to decompensation oversampling and class weights. In fact, it performed very poorly, and therefore its results were not included. One reason concluded is the fact that SMOTE lost the temporal dependency factors, which lead to very poor results. When comparing the oversampling and the class weights, both showed similar results. Class weights were a good yet simple approach, when using weights proportional to the imbalance in the dataset. Approaches of applying double and triple the weight for the minority class did not prove useful in results.

Although AUROC values without using data imbalance techniques were high, the precision and recall values for both classes was very disproportional. With the use of both class weights or oversampling, those values became more balanced and allowed for a better discrimination between classes, where before data imbalance techniques, the model could not predict the minority class well. When training the models with the original features, class weights and oversampling had very similar AUROC scores using LSTM with scores of 0.8856 and 0.8871, respectively, but for Transformer models the oversampling technique proved better in performance, with an AUROC of 0.8630 for class weights and 0.8788 for oversampling. For logistic regressors the same increase in performance is detected, where an AUROC score of 0.9069 was achieved with class weights and 0.9787 with oversampling, indicating that the oversampling technique for the simpler logistic regressor model caused worse generalization. On the other hand, when using the added features in the models, the class weights seemed to have better performance than oversampling. LSTM obtained an AUROC of 0.9294 with class weights and 0.9102 with oversampling, while transformers obtained 0.7932 and 0.7661, respectively. However, on contrary of the models using only the original features, these data imbalance techniques for both models using the added features proved to have better AUROC scores than without using them, while also increasing the recall for the minority class.

5.3.8 Clinical Considerations

To further understand the results obtained with the models created, a clinical analysis was made, for both results and the feature importance. The feature that were considered more relevant are presented in table 5.6. For both sets of original and added features, most of the most important features are related to GCS features: GCS total, motor response, eye opening, verbal response. These features are interesting to see in the important as they are the first practical method to evaluate the degree of consciousness of patients and it represents the severity. This value is then updated regularly to monitor the patients progress. Another fact about these GCS features is that are often related with either very serious or very mild cases of TBI, such as the high GCS total value are generally high or low, and motor, verbal and eye values, with eye opening of spontaneously or no response; motor response with high values of 'obeys commands' and low of 'abnorm extension'. Pupillary size is also present for both eyes in the added features set, which is one of the features present only in that set and also another form of assessing eye responsiveness.

In terms of the results obtained in this chapter, from original and added features for the best models, along with the decompensation notion used and the results for each TBI category, all of them show a relatively high rate of false positive and a low rate of false negatives cases with the imbalance techniques. In one hand, in a clinical perspective, this means that some patients who are not at a risk of decompensation will be classified as such, perhaps costing more resources than needed. On the other hand, this also means that the rate of patients that actually are in risk of decompensating being disregarded is very low, allowing for patients with actual need of further treatments to receive them.

Figures 5.5, 5.6 and 5.7 contain 3 examples predicted by the LSTM model with oversampling and added features. The first plot contains the patient PTS signals and the yellow boxes correspond to the time points where the model predicted decompensation. The second plot in each figure represents the presence of the lowest values for GCS values registered. The last plot, indicates the predicted decompensation in comparison with the *y_true* variable. These plots help in understanding and visualizing how the model is truly predicting decompensation.

Figure 5.5 shows a visualization of a patient from the test set that deceased. In this case, the *y_true* and the prediction have a perfect match in the last plot of figure 5.5, where the two lines of *y_true* and predicted are overlapped, meaning that the model predicted that the patient was going to decompensate in all of the previous 24 hours before deceasing. It is also noticeable that this patient had some time points registered where the eye opening response was at the lowest value according to GCS, alongside fewer cases for motor and verbal response, which could indicate a problematic case of TBI. For the PTS data, the values seem to have more variance during the decompensation time points comparing with the rest of the timeseries, and also heart rate and systolic blood pressure seem to decrease at a higher rate than in other time points. The plots show only a portion of the features that are constantly changing in values, for PTS, GCS and remaining features. For physicians to keep track of all this data and find trends and perform diagnoses is extremely hard, especially with different patients every day. Therefore, these methods that are being studied and experimented in this study try to help in clinical decisions according to the data.

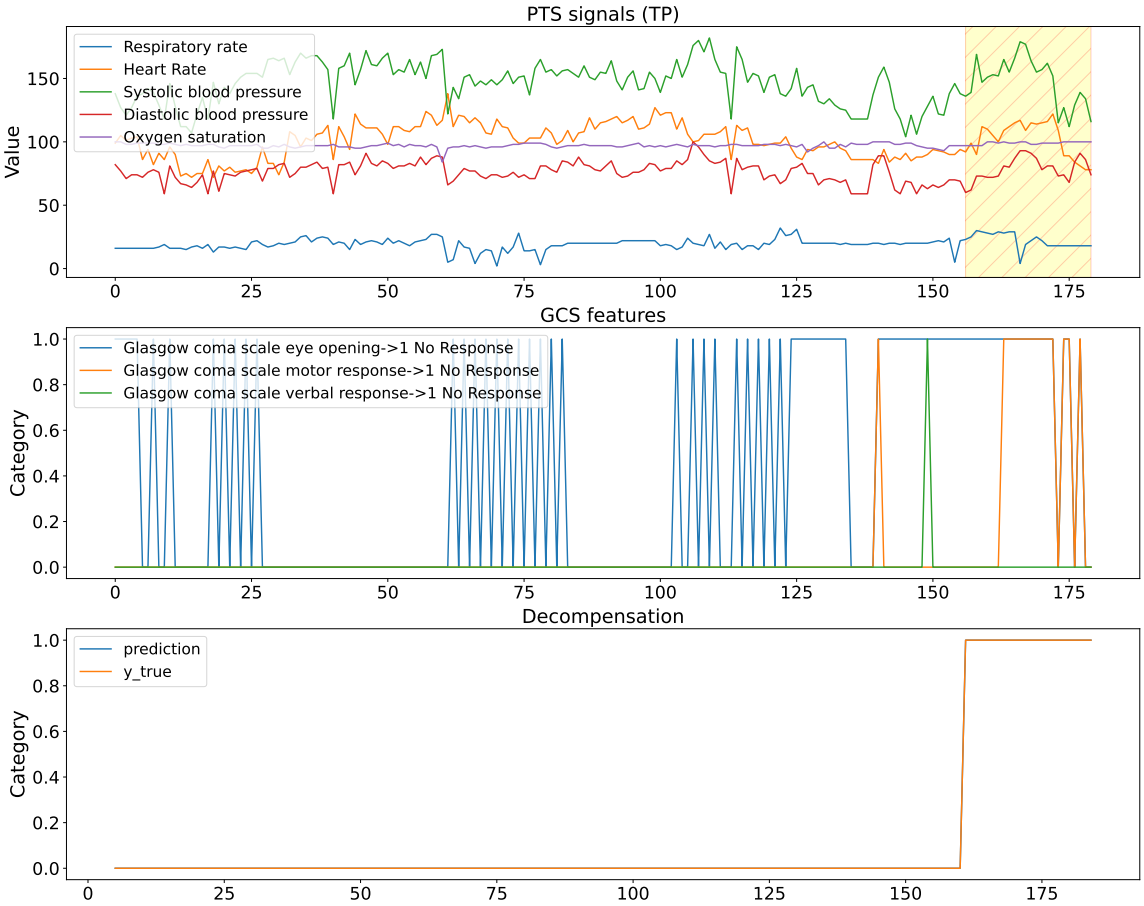


Figure 5.5: Example of a patient decompensation prediction for a TP case. The patient indicates high amounts of GCS variables for eye response, as well as some worrying values for motor and verbal response. The last plot only shows the orange line of the y_true, since the prediction values represent exactly the same values.

Figure 5.6 shows a prediction for a patient which did not decompensate and the model correctly predicted as such. There is a difference that can be seen in this plot comparing with the previous figure 5.5, where the plot for the GCS features does not contain any values. Although values for other values for these categories were not analyzed in this plot, typically the lack of these values indicates a TBI case that is not severe. The prediction also appears as no values registered, since the model did not predict decompensation. In this case, the value for heart rate seem lower for the previous patient predicted as decompensated. The variance also seems to be lower than the previous case, with all PTS data with consistent values during all time points.

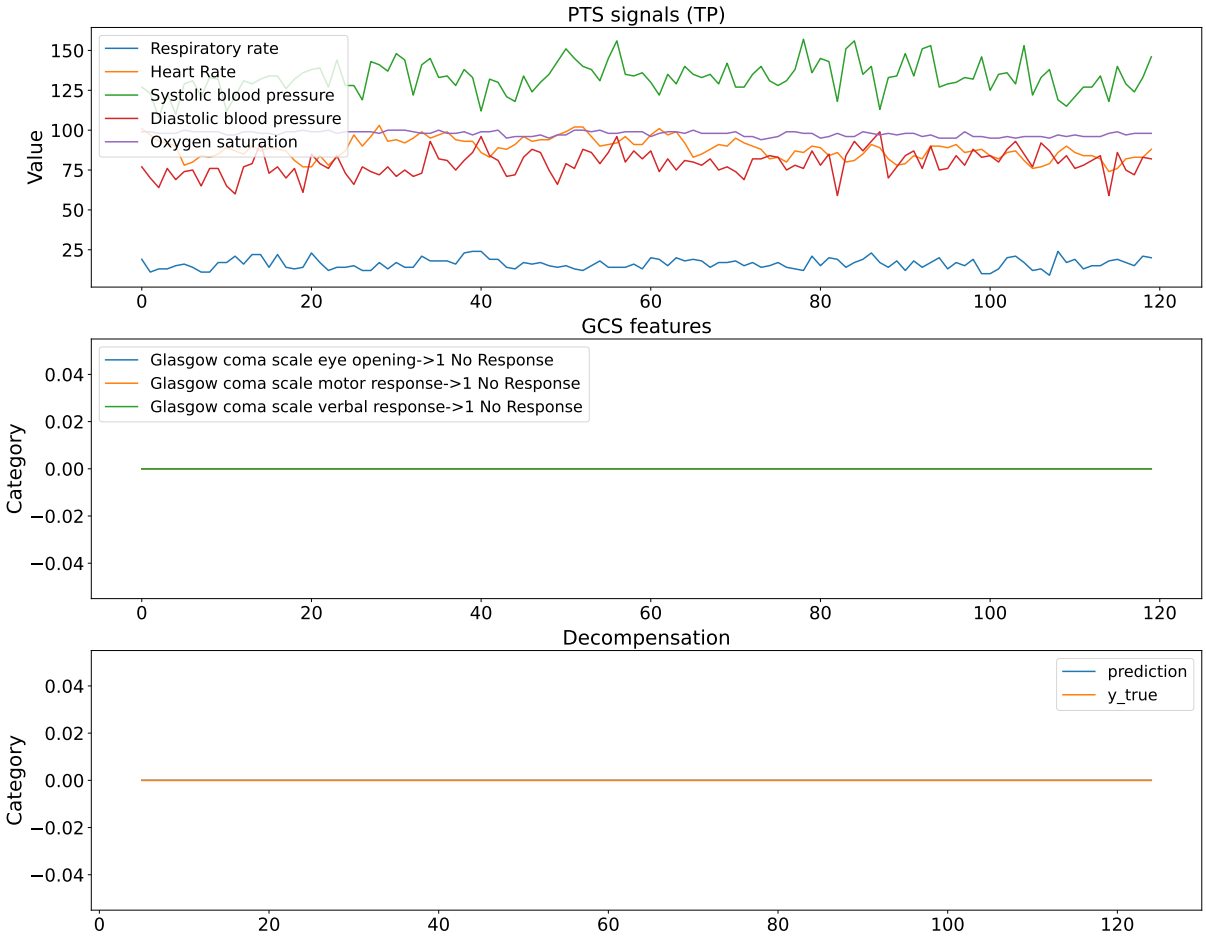


Figure 5.6: Example of a patient decompensation prediction for a TN case. There are no time points where the patient detected mortality, and the plot for the results appear as the model predicted that there was no decompensation. The plot of the GCS features also shows that there were no negative values, which indicates a mild case of TBI.

Finally, for figure 5.7, there is another case of a True Positive (TP) case where the model predicted a patient that decompensated as such. The difference in this case in comparison with figure 5.5 is that the model predicted the decompensation happening some time points before the actual decompensation characterized as 24 hours prior to mortality detected. In the PTS values in the first plot, there can be seen that the variance is much higher than the previous shown cases for all PTS values shown, and also a trend during and after the time points predicted as decompensation, where the PTS values seem to decrease. The GCS values also show higher amounts of eye and motor response with the worse possible values, indicating a very serious risk of severe TBI.

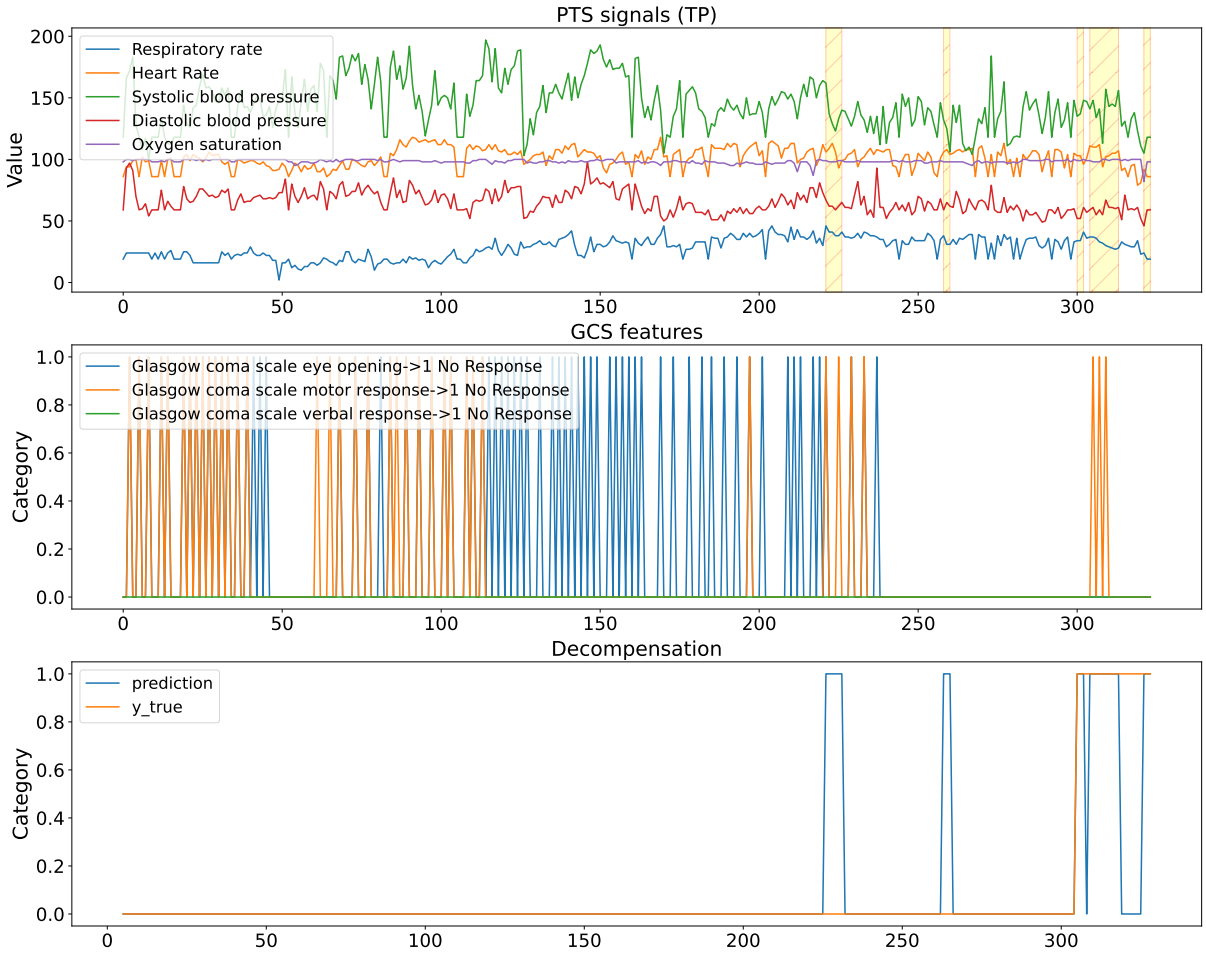


Figure 5.7: Example of a patient decompensation prediction for another TP case. In this case, although the model predicted decompensation for this patient, it predicted some time points before the 24 hours before mortality was registered for this patient. There were also a lot time points registered for GCS values for eye and motor response with serious values.

5.3.9 Limitations

The research for the decompensation prediction task in this work, like any exploratory study, has limitations. The size of the cohort used was the first problem encountered. 1567 patients for the training prediction are not enough for deep learning models to obtain the best possible results, hence the data imbalance techniques used such as class weights and oversampling. Although test and validation sets for this studied that were separated from the training set, external validation could also be useful to have a better perspective of the real life prediction capability of the models, while also allowing the validation and test set to be used in the training set and hence increase the number of patients for training. This study also did not focus on pediatric ages, since a cohort for this age group would have a lesser amount of patients and more time to fine tune the architectures and other possible methods to get the best possible results.

Although imputation strategies showed good results, the high missing data rate is also a

limitation. Imputation data when there are a lot of 'gaps' in the real data introduces limitations, as the new imputated data could be thought of as invented or not representative of the patient case. Unfortunately, missing data in health based datasets is a common challenge.

The use of three different architectures, although provided more results to compare the methods used, also constrained some chances and time to optimize one of the architecture and improve the results obtained, such as analyzing the very high score values obtained in the logistic regression models, the worse results from transformers when using the added samples, or the rate of false negatives obtained.

The decompensation prediction method could have been further explored. Instead of using each hour of the stay and predicting 24 hour mortality, other methods could also have been explored and compared, such as focusing more on physiologic data only.

5.3.10 Summary

Three different models were presented in this chapter - a ML model (logistic regressor) and two DL models (LSTM and Transformer) - with the task of predicting decompensation in TBI patients. These models used sequential time series EHR. Hyperparameter tuning was performed to each model and then the prediction task consisted on using the three models mentioned in different methods: using the original and added features that consisted in 17 and 47 features, respectively; class imbalance techniques such as class weights and oversampling.

The best model obtained from these methods was the LSTM model with the added features and class weights with a score of 0.9294 AUROC. Comparing the three architectures, LSTM obtained better results for the original features and much higher results for the added features, since transformers did not perform well in the later case. Logistic regressors had very high results, with the best being 0.9787 AUROC with the original features and oversampling, which suggests that the logistic models could have poor generalization. Another metric to measure the decompensation prediction task, where if a sample was predicted as decompensating in any given hour for the next 24 hours, it would be classified as decompensation, and false otherwise. In this case, the best model overall was the LSTM, where the best result was obtained with added features and oversampling with a score of 0.9524 AUROC. This metric allowed to check if the model performance was consistent, as it was possible to see that transformers had similar performance to LSTM for the original features and was also more visible the worse results with the added features. The lack of generalization from the logistic regressor was also able to be detected with this metric, where almost all scores for the logistic regressor for each method was above 0.95 AUROC. When comparing each TBI category (mild, moderate and severe), it was possible to have a better understanding of how well the models performed for different scenarios. The models obtained good testing results overall, with the mild category predicting 74 out of the 78 patients correctly, moderate obtained an AUROC of 0.9355, and 0.8960 AUROC for the severe category. With the confusion matrices from figure 5.3 it was possible to notice some false positives being predicted, especially in the severe category. Although this false positive rate was

noticeable, the model performed extremely well for the minority class, allowing patients with real need of better care to be correctly detected.

Chapter 6

Conclusions and Future Work

6.1 Main Conclusions

With the Deep Learning (DL) model results obtained and discussed in previous chapters, this study demonstrated the potential to forecast decompensation events in advance using Electronic Health Record (EHR), allowing for timely interventions and ultimately improved patient outcomes. The technique used to predict decompensation with this data was based on a 24 hour mortality prediction at each hour of the patient stay. Although this notion diverges a little from the original definition of decompensation, other metrics that allowed to group the model results by patient instead of each hour/sample allowed to prove the efficiency of the models. Long-short term memory (LSTM) models outperformed the transformers architecture used, in both original and added features, as well as class weights and oversampling techniques with a score of 0.9102 Area Under the Receiver Operating Characteristics Curve (AUROC). Nevertheless, transformers showed good performance for the original features and with some more time, the architecture could possibly achieve similar results to the LSTM. In terms of the logistic regressors, the model needed some fine tuning to deal with the possible lack of generalization found in the high scores obtained, as the 0.9787 AUROC with original features and oversampling. Throughout the study simpler strategies proved to sometimes be the best option in terms of results. Synthetic Minority Oversampling Technique (SMOTE) was explored but did not capture the temporal dimension well, while simple class weights were able to deal with the huge imbalance, alongside more basic oversampling techniques. Missing data was also imputed with the simple strategy of imputing the normal value for each feature according to previous literature and obtained good results, as it is was an important step in the discretization process. Feature importance was explored and verified the importance of Glasgow Coma Scale (GCS) values in the models created in both sets of features. Also, with the added features set, other features related to eye response were also categorized as having high importance, which is a feature related to GCS and Traumatic Brain Injury (TBI) assessment. In terms of TBI prediction for the three categories of the disease (mild, moderate and severe) the best model obtained good performance in all categories, achieving an AUROC of 0.8960 and 0.9355 for severe and moderate

categories, respectively. AUROC for the mild category was not possible to calculate since there were no patients in the minority class in that case.

6.2 Future Work

As any research goes, further improvements can also be performed. In terms of this study, the first ideal future work would be to externally validate the methods obtained with other datasets. Alongside externally validate the models, improving the size of the TBI cohort is also a very important step, using datasets such as the eICU Collaborative Research Database (eICU) mentioned in previous chapters. The problem with merging different health datasets is that each one is unique and in order to implement methods to obtain a bigger TBI cohort from different sources can be challenging.

An important case that unfortunately fell out of scope in this study is the pediatric section of TBI. This age range represents a relatively big percentage of total cases, and although it is usually harder to predict TBI outcomes in pediatric ages, it is highly important. For future work, models such as the implemented in this study should be trained with this age range to predict the decompensation outcome using pediatric datasets.

The methods demonstrated in this study would benefit with more time spent in fine tuning the models. One of the examples of this problem was the transformer architecture that failed to achieve good results with the added features set. As different methods were approached in this study, time to optimize each task to their maximum capability was not possible. In the future, investing time in each of the techniques could certainly improve the results presented in this study, namely the transformer methods, which are novel models with high potential of achieving great results, as seen with the original features set. Fine tuning this model and architecture could possibly surpass the results obtained with LSTM methods.

Bibliography

[1] Roger Mark Alistair Johnson, Tom Pollard. Medical information mart for intensive care. 2016.

[2] Mauricio Villarroel Gari D. Clifford Ikaro Silva Benjamin Moody, George Moody. Mimic-iii waveform database matched subset. 2020.

[3] PC Blumbergs, G Scott, J Manavis, H Wainwright, DA Simpson, and AJ McLean. Stalning af amyloid percursor protein to study axonal damage in mild head injury. *The Lancet*, 344 (8929):1055–1056, 1994.

[4] Allison Capizzi, Jean Woo, and Monica Verduzco-Gutierrez. Traumatic brain injury: an overview of epidemiology, pathophysiology, and medical management. *Medical Clinics*, 104 (2):213–238, 2020.

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[6] Maximilian Christ, Andreas W Kempa-Liehr, and Michael Feindt. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016.

[7] MRC Crash Trial Collaborators et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *Bmj*, 336(7641): 425–429, 2008.

[8] National Data and Statistical Center (NDSC). National database: 2021 profile of people within the traumatic brain injury model systems. Accessed December 9, 2022.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[10] João Fonseca, Xiuyun Liu, Hélder P Oliveira, and Tania Pereira. Learning models for traumatic brain injury mortality prediction on pediatric electronic health records. *Frontiers in neurology*, 13, 2022.

[11] João Pedro Barbosa Fonseca. Ai-based models to predict the traumatic brain injury outcome. 2022.

[12] Centers for Disease Control, U.S. Department of Health Prevention, and Human Services (CDC). Surveillance report of traumatic brain injury-related emergency department visits, hospitalizations, and deaths. *CDC Injury Center*, 2014.

[13] Ben D Fulcher, Max A Little, and Nick S Jones. Highly comparative time-series analysis: the empirical structure of time series and their methods. *Journal of the Royal Society Interface*, 10(83):20130048, 2013.

[14] Mark W Greve and Brian J Zink. Pathophysiology of traumatic brain injury. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, 76(2):97–104, 2009.

[15] Frank E Harrell Jr, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

[16] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

[17] Ronald L Hayes and C Edward Dixon. Neurochemical changes in mild head injury. In *Seminars in Neurology*, volume 14, pages 25–31. © 1994 by Thieme Medical Publishers, Inc., 1994.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

[19] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[20] Amy C Justice, Kenneth E Covinsky, and Jesse A Berlin. Assessing the generalizability of prognostic information. *Annals of internal medicine*, 130(6):515–524, 1999.

[21] Han Biehn Kim et al. *Development and Validation of Traumatic Brain Injury Outcome Prognosis Model and Identification of Novel Quantitative Data-Driven Endotypes*. PhD thesis, Johns Hopkins University, 2020.

[22] Jae-Sung Kim, Lihua He, and John J Lemasters. Mitochondrial permeability transition: a common pathway to necrosis and apoptosis. *Biochemical and biophysical research communications*, 304(3):463–470, 2003.

[23] Patrick M Kochanek, Rachel P Berger, Hülya Bayr, Amy K Wagner, Larry W Jenkins, and Robert SB Clark. Biomarkers of primary and evolving damage in traumatic and ischemic

brain injury: diagnosis, prognosis, probing mechanisms, and therapeutic decision making. *Current opinion in critical care*, 14(2):135–141, 2008.

[24] W Koehrsen. A feature selection tool for machine learning in python. *Towards Data Science*, 22, 2018.

[25] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis. *Nature methods*, 14(7):641–643, 2017.

[26] Andrew IR Maas, Anthony Marmarou, Gordon D Murray, Sir Graham M Teasdale, and Ewout W Steyerberg. Prognosis and clinical trial design in traumatic brain injury: the impact study. *Journal of neurotrauma*, 24(2):232–238, 2007.

[27] Andrew IR Maas, David K Menon, P David Adelson, Nada Andelic, Michael J Bell, Antonio Belli, Peter Bragge, Alexandra Brazinova, András Büki, Randall M Chesnut, et al. Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. *The Lancet Neurology*, 16(12), 2017.

[28] Oscar JL Mitchell, Maya Dewan, Heather A Wolfe, Karsten J Roberts, Stacie Neefe, Geoffrey Lighthall, Nathaniel A Sands, Gary Weissman, Jennifer Ginestra, Michael GS Shashaty, et al. Defining physiological decompensation: An expert consensus and retrospective outcome validation. *Critical Care Explorations*, 4(4), 2022.

[29] Nino A Mushkudiani, Chantal WPM Hukkelhoven, Adrián V Hernández, Gordon D Murray, Sung C Choi, Andrew IR Maas, and Ewout W Steyerberg. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of clinical epidemiology*, 61(4):331–343, 2008.

[30] Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[31] Winsconsin Department of Health Services. Wish injury related traumatic brain injury icd-9-cm codes. 2016.

[32] Sakshi Patel, Shivani Sihmar, and Aman Jatain. A study of hierarchical clustering algorithms. In *2015 2nd international conference on computing for sustainable global development (INDIACom)*, pages 537–541. IEEE, 2015.

[33] Pablo Perel, Phil Edwards, Reinhard Wentz, and Ian Roberts. Systematic review of prognostic models in traumatic brain injury. *BMC medical informatics and decision making*, 6(1):1–10, 2006.

[34] DeWitt TDB. Peter E. Harmonized pediatric traumatic brain injury hackathon. 2021.

[35] Brendan M Reilly and Arthur T Evans. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine*, 144(3): 201–209, 2006.

[36] Patrick Royston and Ian R White. Multiple imputation by chained equations (mice): implementation in stata. *Journal of statistical software*, 45:1–20, 2011.

[37] Sherman C Stein, Patrick Georgoff, Sudha Meghan, Kasim Mizra, and Seema S Sonnad. 150 years of treating severe traumatic brain injury: a systematic review of progress in mortality. *Journal of neurotrauma*, 27(7), 2010.

[38] Ewout W Steyerberg, Nino Mushkudiani, Pablo Perel, Isabella Butcher, Juan Lu, Gillian S McHugh, Gordon D Murray, Anthony Marmarou, Ian Roberts, J Dik F Habbema, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS medicine*, 5(8):e165, 2008.

[39] LE Thibault and TA Gennarelli. Brain injury: an analysis of neural and neurovascular trauma in the nonhuman primate. In *Association for the Advancement of Automotive Medicine (AAAM), Conference, 34th, 1990, Scottsdale, Arizona, USA*, 1990.

[40] Mark Van der Laan, Katherine Pollard, and Jennifer Bryan. A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, 73(8):575–584, 2003.

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[42] Bon H Verweij, J Paul Muizelaar, Federico C Vinas, Patti L Peterson, Ye Xiong, and Chuan P Lee. Impaired cerebral mitochondrial function after traumatic brain injury in humans. *Journal of neurosurgery*, 93(5):815–820, 2000.

[43] Jean-Louis Vincent and Jacques Berré. Primer on medical management of severe brain injury. *Critical care medicine*, 33(6):1392–1399, 2005.

[44] Victor Volovici, Ari Ercole, Giuseppe Citerio, Nino Stocchetti, Iain K Haitsma, Jilske A Huijben, Clemens MF Dirven, Mathieu van der Jagt, Ewout W Steyerberg, David Nelson, et al. Intensive care admission criteria for traumatic brain injury patients across europe. *Journal of critical care*, 49:158–161, 2019.

[45] Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2565–2573, 2018.

[46] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.