Machine Learning Modelling of Critical Care Patients in the Intensive Care Units

Mark Pieroni

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

This research programme was carried out in collaboration with the Liverpool Centre for Cardiovascular Science

Jan 2023

Declaration

I, Mark Pieroni, confirm that the work presented in this thesis is my own. Furthermore, I confirm this has been indicated in the thesis, where information has been derived from other sources.

Mark Pieroni

Word count (excluding acknowledgement, appendices, and references): 39,893 words.

Acknowledgement

I would like first to thank my fantastic supervisory team, Dr Ivan Oiler-Caparroso and Dr Sandra Ortega-Martorell, for their constant support and guidance during my PhD project. Being under your tutelage has given me the tools to progress into the next part of my career and has made it a rewarding journey. You guys have gone above and beyond for me throughout my PhD, and I cannot thank you enough. I could not have asked for better supervisors. Thank you! I would like to thank Prof Ingeborg Welters, Dr Brian Johnston and Prof Gregory Y.H Lip for their time and effort in supporting this project giving expert medical advice and support, which has been invaluable. I want to thank LCCS for the collaborative opportunities and for finding funding to enable me to study for my PhD. I would like to thank everyone in office 608 and the wider team who have been with me during my journey at LJMU. You guys have always kept me going and cheered me along. I could not have asked for a better group of friends and could not have done it without you all.

I want to say special thanks to my mum and dad, Elaine and Joe, who have been nothing but supportive during my PhD years. I cannot thank you enough for supporting and pushing me to pursue what I enjoy. Next, to my wonderful girlfriend, El. Thank you for all the times you have supported me through my PhD. Not all heroes wear capes. Finally, I would like to thank all the maths department staff at LJMU, who have been there through all my academic stages. Finally, I want to say Liverpool John Moores has given me a fantastic learning environment. Thank you to anyone who has been a part of my journey in completing my PhD.

Abstract

The ICU is a fast-paced data-rich environment which treats the most critically ill patients. On average, over 15 % of patients admitted to the ICU amount in mortality. Therefore, machine learning (ML) is paramount to aiding the optimisation and inference of insight in critical care. In addition, the early and accurate evaluation of the severity at the time of admission is significant for physicians. Such evaluations make patient management more effective as they are more likely to predict whose physical conditions may worsen. Moreover, ML techniques could potentially enhance patients' experience in the clinical setting by providing medical alerts and insight into future events occurring during hospitalisation. The need for interpretable models is crucial in the ICU and clinical setting, as it is vital to explain a decision that leads to any course of action related to an individual patient.

This thesis primarily focuses on mortality, length of stay forecasting, and AF classification in critical care. We cover multiple outcomes and modelling methods whilst using multiple cohorts throughout the research. However, the analysis conducted throughout the thesis aims to create interpretable models for each modelling objective. In Chapter 3, we investigate three publicly available critical care databases containing multiple modalities of data and a wide range of parameters. We describe the processes and contemplations which must be considered to create actionable data for analysis in the ICU. Furthermore, we compared the three data sources using traditional statistical and ML methods and compared predictive performance. Based on 24 hours of sequential data, we achieved AUC performances of 79.5% for ICU mortality prediction and a prediction error of approximately 1.3 hours for ICU LOS.

In Chapter 4, we investigate a sepsis cohort and conduct three sub-studies. Firstly, we investigated sepsis subtypes and compared biomarkers using traditional modelling methods. Next, we compare our approach to commonly and routinely used scoring systems in the ICU, such as APACHE IV and SOFA. Our tailored approach achieved superior performance with pulmonary and abdominal sepsis (AUC 0.74 and 0.71respectivly), displaying distinct individualities amongst the different sepsis groups. Next, we further expand our analysis by comparing ML methods and inference approaches to our baseline model and ICU acuity scores. We further investigate extending analysis to other outcomes of interest (In-hospital/ICU mortality, In-hospital/ICU LOS) to gain a more holistic view of the sepsis derivatives. This research shows that nonlinear models such as RF and GBM commonly outperformICU scoring, methods such as APACHE IV and SOFA and linear methods such as logistic/linear regression. Lastly, we extend our analysis in a multi-task learning framework for model optimisation and improved predictive performance. Our results showed superior performance with pulmonary, abdominal and renal/UTI sepsis (AUC 0.76, 0.77 and 0.73, respectively). Lastly, Chapter 5 investigates the classification of atrial fibrillation (AF) in long-lead ECG waveforms in sepsis patients. We developed a deep neural network to classify AF ECGs from Non-AF ECG cases in conjunction with refining a method to gain insight from the neural network model. We achieved a predictive performance of 0.99 and 0.89 regarding the test and external validation data. The inference from the model was achieved through the use of saliency maps, dimensionality reduction methods and clustering, utilising the automatic features learned by the developed model. We developed visualisations to help support the inference behind the classification of each ECG prediction.

Overall, the research displays a wide range of novelties and unique approaches to solving various outcomes of interest in the ICU. In addition, this research demonstrates the implication of ML applicability in the ICU environment by providing insight and inference to diverse tasks regardless of the level of complexity. With further development, the frameworks and approaches outlined in this thesis have the potential to be used in clinical practice as decision-support tools in the ICU, allowing the automated alert and detection of patient classification, amongst others. The results generated in this thesis resulted in journal publications and medical understanding gained from insight available in the developed ML frameworks.

Publication and Dissemination of Results

Journal and Peer-reviewed Journal Papers

- Journal Paper (Frontiers in Medicine | ***First Author***): *In-Hospital Mortality of Sepsis Differs Depending on The Origin of Infection: An Investigation of Predisposing Factors. 2022* *https://doi.org/10.3389/fmed.2022.915224*

- Journal Paper (Frontiers in Cardiovascular Medicine | ***Second Author***): *Development of a Risk Prediction Model for New Episodes of Atrial Fibrillation in Medical-Surgical Critically Ill Patients Using the AmsterdamUMCdb. 2022* *https://doi.org/10.3389/fcvm.2022.897709*

- Journal Paper (Cardiovascular Research | ***Third Author***): *How Machine Learning Is Impacting Research in Atrial Fibrillation: Implications for Risk Prediction and Future Management. 2021* *https://doi.org/10.1093/cvr/cvab169*

Abstracts, Talks, and Conferences

European Society of Intensive Care Medicine (ESICM) Paris 2022 (Abstract submission): *Interpretation and agreement between intensive care unit physicians' ability to diagnose atrial fibrillation on single-lead ECG traces of critical care patients in the ICU using the MIMIC-III database.*

Liverpool Centre of cardiovascular science (LCCS) Research Group meetings. February 2022, Nov 2021, Oral Presentation.

LCCS, AI/ML and Cardiovascular Disease Workshop, January 2021, Oral Presentation.

LJMU, Faculty Research Day, Online, Liverpool, UK. May 2021. Oral Presentation. *Machine Learning Modelling of Critical Care Patients in the Intensive Care Units.*

LJMU, Open Research Week Research Café, Online, Liverpool, UK. February 2021. Oral Presentation. Machine learning with Big Data in Critical Care.

Achievements

Datathon Winner – Galp Oil and Gas Refinery (2019 - $25,000 Grand Prize): To optimise performance and increase the energy efficiency of the Galp refinery. *The pressure difference in the catalyst of the hydrocracker unit increases faster than it should. Although the increase of this parameter is expected, the speed at which this increase occurs is not. The aim was to find a solution to this problem by better understanding why this occurs, allowing for the optimisation of industrial processes.*

Datathon Grand Finalist – ECSIM, Critical Care Datathon (2021): *Multidisciplinary teams and international experts investigated clinical questions by leveraging large datasets of electronic health records to investigate better patient care management. The aim was to develop a risk prediction model for new episodes of atrial fibrillation by exploring associated biomarkers and clinical factors.*

Research Questions

- Can machine learning (ML) be helpful in decision support and management of patients in the ICU?
- Can we successfully model ICU patients for various outcomes with a high degree of accuracy?
- Can we develop ML models for ICU applications which are interpretable?
- Can we apply model optimisation strategies such as multi-task learning to improve model performance?

Research Aims & Objectives

- To exploit needs currently in ML applications to healthcare, which is the ability to effectively model patients in the ICU environment. Therefore, displaying associations between factors and outcomes which is intuitive, hence easily interpreted by clinicians.
- To model ICU patients using various ML methods to discover the most optimal approach to model the outcome of interest.
- To model ICU patients under several ML scenarios, to investigate a deeper understanding of outcomes of interest and how the target may relate to other clinical factors and outcomes.
- To investigate the model optimisation strategies to improve model performance while also leveraging interpretability. Thus, establish the criteria for multi-task learning and compare the results with single-task approaches to compare feature importance, performance, and inference.
- To determine the steps and suitability of publically available ICU data, investigate similarities and differences amongst the data sources, and describe the necessary steps and considerations to conduct ML and statistical analysis.
- To advance current ICU scoring approaches so that we can successfully create interpretable and highly accurate models of clinical multivariate data from a range of diverse data sources.

Table of Contents

List of Figure

List of Tables

List of Formulae

1 CHAPTER 1: Background & Introduction

1.1 Introduction

The Intensive care unit (ICU), also known as critical care unit (CCU), treats acutely ill patients needing radical lifesaving treatments. ICUs are specialised units used for close monitoring and treatment of patients, which in turn could potentially improve outcomes. Typically, ICUs have a more significant number of healthcare professionals compared to other hospital departments; studies have shown to reduce mortality rates, lower hospital length of stay (LOS) and have been associated with fewer illness complications [1]. Every year, more than 5.7 million adults are admitted to the ICU in the United States (US), costing the healthcare system more than 67 billion US dollars each year [2]. Because ICUs accommodate the most critically ill hospital patients, it is intuitive that mortality rates are higher than in a general ward, with approximately 15% of all ICU patients expiring globally [3].

Patients who require life-sustaining treatments or are at high risk would be in immediate need of extensive monitoring and direct attention from healthcare providers. A by-product of this is the wealth of information recorded for each patient in the ICU, including high-resolution physiological signals and various laboratory tests, in addition to detailed medical history in the form of electronic health records (EHR). Nevertheless, vital aspects of patient care are not yet captured in an autonomous manner [4]. With recent advances and developments in artificial intelligence (AI), many researchers are exploring complex autonomous systems in real-world domains. In the ICU environment, clinicians and healthcare professionals are required to make lifesaving decisions while dealing with high levels of uncertainty under strict time constraints to synthesise high volumes of complex physiologic and clinical data. AI technology could assist in administering repetitive patient assessments in real-time, but also in integrating and interpreting these data sources with EHR, thus potentially enabling more timely and targeted interventions [5], [6]. Furthermore, the potential application of AI to healthcare in the ICU could reduce staff workloads, allowing them to prioritise more critical tasks, in addition to aiding in human decision-making.

Over recent years, ICU and healthcare analytics have generated much attention, with much interest from healthcare providers and researchers due to the importance of saving patient lives. However, most clinical-based studies have focused on providing simplistic scores that focus on the severity of disease or illness [7]. The fundamental of these scoring systems is to add a weighting to the degree of abnormality of an organ or a disease based on vital sign measures, historical data and visual inspection of the patient, all to attempt to identify patients at high risk. Currently, available popular acuity scores are APACHE [8], SAP [9] and SOFA [10], amongst a wide range of others. The variables used in these models can be segregated into four main groups: age, comorbidities, physiological abnormalities and acute diagnoses. The purpose of these scoring systems plays a vital status in helping prioritise resources and the best care given to the patient [11]. However, there are some fundamental problems with the current scoring systems in the ICU. Firstly, the scores are population-based and are not patient-specific. Secondly, they are not developed for real-time evaluation of patients [12]. Population-based scores estimate the likelihood of specific outcomes or events occurring in a general population of patients with similar characteristics to those in the data used to develop the model [13], [14]. In opposition, patient-specific models are developed using data from an individual patient and consider the patient's unique characteristics, such as their medical history, vital signs, and laboratory values, among other measurements. These models are designed to provide individualized risk measurements and treatment proposals for that specific patient [15], [16] . Furthermore, these scores are developed using static data from baseline patient characteristics, defined as measures obtained within the first 24 hours of the ICU admission. However, interestingly, there are no universal scoring systems that are currently used in the ICU environment [17].

Due to the above limitations, there is a need for near real-time patient-specific algorithms for patient risk evaluation and modelling. Current modern-day implementations of data mining, ML or predictive analytics approaches differ from current methods used in the ICU environment. No prior hypotheses

are tested. However, instead, the goal is to extract the data's repeated patterns and relationships that are useful in predicting future outcomes. This approach closely resembles how humans acquire expertise and procedural knowledge when interacting with complex medical data [18].

Although methods for quantifying medical status in the ICU are vital in conjunction with patient care, many arguments and issues have been raised. For example, static risk scores, such as those previously discussed, are used to characterise the patients' states. Nonetheless, such scores are limited to the number of variables used in conducting them. Furthermore, these scores do not account for temporal trends, meaning currently, models cannot display adequate behaviour of the patient's state over time [19]. Therefore, real-time forecasting of clinical interventions remains a challenge within the ICU, as this plays an increasingly pivotal role in acute healthcare delivery. Healthcare professionals and clinicians must anticipate the approximate care needed in a fast-paced, data-rich environment. This project proposes a solution for a real-world problem in public health, specifically for improving a range of clinical outcomes of patients with sepsis and cardiovascular (CV) complications in the ICU.

1.1.1 Machine Learning in Healthcare

Basic research in cardiovascular medicine has yielded dramatic insight into physiology, leading to therapeutic advances and a significant decrease in CV mortality over the past 50 years [20], [21]. However, it is increasingly recognised that even highly effective therapies have heterogeneity of effects at the level of the individual. These factors limit the potential impact of scientific advances when implemented in care. Powerful ML and deep learning (DL) methods show promise in supporting more personalised medicine and effective population health management [22]. ML and DL techniques have shown outstanding results recently in versatile tasks, such as the recognition of body organs from medical images [23], classification of interstitial lung disease [24], medical image reconstruction [25] and brain tumour segmentation [26], amongst others. In addition, ML/DL have displayed their value, which is being able to achieve human-level performance in clinical pathology [27], radiology [28], ophthalmology [29] and dermatology [30]. A by-product of this has led to the development of computer-aided diagnosis systems. While ML in healthcare is a very active research topic, most of the health data collected are never used for predictive models, which are successfully integrated into the clinical setting [22], with only 15% of hospitals currently and routinely using ML for limited purposes [31]. Although ML has demonstrated benefits in certain medical domains, the successful utilisation of ML requires considerable effort from human experts, given that no algorithm/methodology can achieve satisfactory performance on all possible problems (i.e., No Free Lunch [32]). One key obstacle relates to the black box nature or opacity, of many ML algorithms, especially in critical use cases, including clinical decision-making. There is some hesitation in deploying such models because the cost of model misclassification is potentially high [33]. There is much opportunity and demand for interpretable ML models in such situations. ML models allow end-users to evaluate the model ideally before an action is taken by the end-user, such as a clinician. By understanding the reasoning behind the predictions, interpretable ML models give users reasons to accept or reject predictions and recommendations [34]. However, historically, there has been a trade-off between interpretable ML models and performance[35]. However, in a real-world application context, interpretability might be judged only according to the specific requirements of the application area in acknowledgement that different applications usually have different interpretability and exploitability needs [36].

1.1.2 Clinical Prediction Models

Clinical prediction models are mathematical tools derived from original research and are primarily intended to assist physicians in their clinical decision-making at the bedside. Typically, they combine multiple predictor variables, including patient demographics, history and physical examination findings, vitals and laboratory test results, amongst others, to calculate the probability or forecast a specific outcome [37]. Developing a multivariate clinical prediction model generally requires the identification of the important predictors out of a set of preselected candidate variables. Traditionally assigning the related weights for each predictor variable in a combined risk score, then estimating the model's predictive performance, calibration, discrimination and reclassification properties, assessing its

potential for optimisation using validation techniques, and if necessary, adjusting the model for overfitting [38].

Research by [35] investigated current approaches to predict cardiovascular risk to identify patients who could benefit from preventive treatments, using only routine collected EHR records. In this study, they deployed a range of ML algorithms (logistic regression (LR), Random Forest (RF), and shallow feedforward neural networks (MLP), amongst others), with the results showing that the MLP achieved the highest performance accuracy. In addition, all other ML models beat the baseline methods and demonstrated that they are superior at predicting the absolute number of CV disease cases correctly. Finally, research presented by [39] highlighted the use of convolutional auto-encoders (CAE) in conjunction with Long short-term memory (LSTM) classifiers to recognise arrhythmias using ECG signals automatically. The implementation of the DL model showed outstanding performance achieving an accuracy of over 99% with the improved computational time and successful compression/reconstruction of the ECGs. This research highlights that ML can achieve outstanding predicting performance. Furthermore, the effectiveness of ML in modelling a complex medical task could potentially optimise current healthcare practices. A review by [36] showed similar results, which systematically reviewed DL model implementations on EHR data. The analysis was conducted with a range of DL approaches (recurrent neural networks (RNN), convolutional neural networks (CNN), autoencoders (AEs), generative adversarial networks (GAN) and model variants) applied to various target applications from papers reviewed between 2010-2018 (98 articles). The results revealed that although DL approaches tend to display favourable performance over traditional ML methods, issues were highlighted regarding the interpretability and transparency of DL models when applied to healthcare applications. Furthermore, this highlights that all models that aim to be developed for clinical applications need to be able to explain the rationale behind each prediction, which is paramount in mitigating risk to the user and the intended subject, especially in critical care.

1.1.3 Interpretable Machine Learning

ML has recently received considerable attention for its ability to accurately predict a wide variety of complex phenomena in various application domains, as discussed previously. Moreover, in recent years, there has been a growing realisation that in addition to predictions, ML models are capable of producing knowledge about the domain relationships contained in data, often referred to as interpretations [40]. As a result, approaches such as DL have become the mainstream method in many important domains. Unfortunately, DL works as a black box model in the sense that, although DL performs exceptionally well in practice, it is difficult to explain its underlying mechanisms and behaviours. Common questions about deep learning models are, first, how has the DL model learnt to make predictions? Secondly, what features are favoured concerning the input data? Thirdly, what changes can be made to the model to improve performance? In recent, only very modest successes have been made in answering these questions. The European Union proposed in 2016 that individuals influenced by algorithms have the right to obtain an explanation. Therefore, the lack of interpretability has become the main barrier to DL and, more broadly, ML in recent research [41]. A black box model could be defined as a function that is too complicated for any human to comprehend. The lack of transparency and accountability of predictive models can have severe consequences. For example, recent work on ‘explainable ML’ for DL commonly display’s a second (post hoc) model, which is created to explain the first black-box model. However, this is problematic in most cases, as explanations are often unreliable and can be misleading [42]. Thus, a fundamental problem facing explanations of such processes is finding ways to reduce the complexity of the elementary operations used in DL architectures. This can be done by either creating a proxy model which behaves similarly to the original black box in a way that is easier to explain (such as the PRN [43]) or by creating salience maps to highlight small portions of the computation to highlight which is relevant [44].

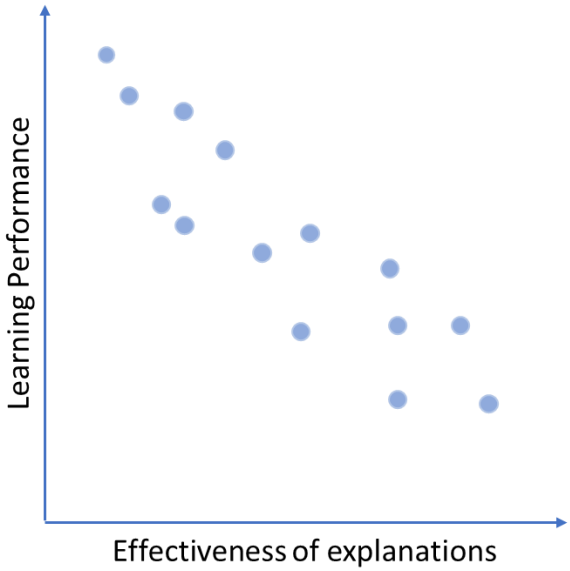


Figure 1: A depiction of the accuracy versus interpretability trade-off

Figure 1 highlights the accuracy interpretability trade-off outlined in the explainable artificial intelligence program broad agency announcement [45]. This research conducted experiments with a static tabular dataset, where several ML algorithms were applied to the same dataset. This research highlights the widespread belief that more complex models are more accurate, meaning that complicated DL is necessary for top predictive performance. However, this is often not the case, particularly when the data is structured with good representations, in terms of naturally meaningful features. When considering problems that have structured data with meaningful features, there are often no significant performance differences in more complex classifiers (deep NN, GBM, random forest) and simpler classifiers (LR, decision trees (DT)) after pre-processing [42]. Similar results were highlighted in research conducted by [46] which reviewed 71 studies between 1/2016 and 8/2017 and compared ML models to LR for binary outcome tasks using static EHR data. Methods implemented were: LR, RF, DT, Neural networks (NN) and support vector machines (SVM). The results showed no evidence of superior performance of ML (nonlinear methods such as RF) over LR when using static linear data.

1.1.4 The Motivation for the Clinical Problems and Tasks

The ICU setting has been depicted by literature as a complex modelling environment due to many challenges, as formally discussed. This project aims to improve healthcare by transforming and optimising outcomes in the ICU. We aim to address these challenges using ML. More specifically, modelling associations between clinical and demographic factors concerning adverse outcomes commonly inherent with sepsis. We aim to take a comprehensive approach to investigate several adverse outcomes associated with sepsis in the ICU. This research consists of multiple modelling frameworks; risk prediction, mortality (in-hospital, in-ICU), and forecasting length of stay(in-hospital/ICU). A critical point in the thesis's focus is to create ML models that cannot only predict the risk of adverse outcomes but are also interpretable. That is, the association map between factors and outcomes is intuitive and hence easily interpreted by clinicians. In statistics and ML, it is assumed there is an existence of a trade-off between model performance and interpretability, as discussed in the previous literature. However, more recent research challenges this assumption. Traditionally, clinical models use simple statistical methods that, although interpretable, tend to display poor performance when the data is complex, noisy or nonlinear. In ICUs, there is a real need for interpretable models without performance sacrifice. By identifying biomarkers associated with different outcomes, we can optimise ICU effectiveness concerning patient treatment or, sub-sequentially, the care provided. In

addition, our developed framework in this thesis can easily be transferable to other related and more general clinical domains.

The work in this thesis would provide the base for models that could be further implemented and utilised in hospitals and ICU environments. In turn, this could optimise patient care and hospital processes, reduce costs and misdiagnosis and potentially inform intervention, thus saving lives. We displayed in this thesis that ML models are predominantly superior to traditional/current methods used to quantify a patient's current status in the ICU. We displayed the steps in modelling complex ICU medical data and the fundamentals of the ML pipeline, which must be considered. We displayed a range of methods to increase interpretability in nonlinear black box models such as RF, GBM and CNNs. Lastly, we demonstrated a framework to increase predictive performance using multi-task learning (MTL). Collectively, the main clinical outcomes modelled are predominantly two-fold. Firstly, we investigate if we can model ICU patients effectively utilising ML and traditional methods and compare these methods to current ICU scoring systems. Secondly, what level of interpretability can we obtain, thus, can we deduce the rationale behind the predictions made.

1.2 Research Novelty

Within this thesis, several areas of novelty build upon existing ideas. The novel aspects are listed briefly below:

- Produced diagnostic models for sepsis subtypes allowing for clinical insight to be obtained. Results from this research resulted in a journal publication with validated clinical insight.

- Compares three large open-source critical care databases, defines the data processing steps for analysis, and compares predictive performance between the data sources.

- Implemented and explored the suitability of MTL in the ICU with sepsis utilising the piling MTL framework.

- Implemented and explored the suitability of a range of ML approaches to model ICU patients and compared this to traditional ICU severity scoring systems.

- We explored a novel framework for AF ECG interpretation, visualisation and classification.

1.3 Thesis Overview

The research in this thesis details the modelling strategies and techniques to model a range of clinical outcomes of interest regarding critical care patients in the ICU. The overall objective is to develop models that can be used in clinical practice to inform intervention or optimise medical or patient processes, therefore, allowing for a level of interpretability to be deduced. The thesis chapters develop from initial traditional modelling techniques to leverage clinical insight to more sophisticated ML approaches to optimise model performance while maintaining a level of interpretability. Chapter 2 outlines the statistical and ML approaches in granular detail used throughout the thesis. The thesis contains three chapters detailing the pipeline and experimental setup of different clinical tasks utilising different ICU data sources and structures. Chapters 3 and 4 investigate different medical outcomes of interest using static tabular data modelling methods. First, we investigate three open-source critical care databases and compare data availability and performance. Next, Chapter 4 utilises data from the eICU database and explores a sepsis cohort in a clinical and ML methodological setting. In exploring various modelling approaches and implementing an MTL framework to increase predictive performance using the other sepsis subtypes as auxiliary tasks in the MTL framework. Research conducted in Chapter 5 investigated the classification of AF in septic ICU patients using long lead ECGs and explored methods of visualising the decision boundary and the rationale behind the model's classification probability. We further extended to explore the potential clusters of the ECGs in the 2-dimensional space created but

the automatic features learned by the CNN model. Lastly, Chapter 6 reflects on the current methods, strategies, and results generated to determine if any improvement or refinement could have been implemented in the experiments. Collectively, the novelties of this research are apparent, in addition to its potential use case application in the clinical setting.

2 CHAPTER 2: Methodological Approaches

2.1 Introduction

ML can be understood as a set of tools and methods that attempt to infer patterns and extract insight from observations made of the physical world. ML is the combination of computer science, mathematics and statistics, with the central element of ML giving computers the ability to learn without being explicitly programmed. ML is one approach that has matured considerably over time and has grown to be the facilitator of the field of 'data science', which has become the facilitator of 'big data' [47]. The term ML refers to the automated detection of meaningful patterns in the data, referring to inductive reasoning and inductive inference. The central theme is to develop tools for expressing domain expertise, translating these into a learning bias, and quantifying the effect of such bias on the success of learning [48]. There are primarily four main classifications of machine learning being supervised, unsupervised, semi-supervised and reinforcement learning. The task and the type of data being utilised deduce the branch of ML used. Our research uses predominantly supervised methods for classifying and forecasting mortality and length of stay. In the later chapters, several unsupervised methods are touched upon, such as PCA and UMAP for dimensionality reduction and Kmenas for clustering. Supervised learning is a learning technique that uses distinct input and output parameters [49]. Whereas unsupervised learning is where the output or labels are learnt, these algorithms discover hidden patterns, structures or grouping within the data without human intervention.

2.2 Methodological Approaches

In this section, we list all the algorithms and models in detail which were used throughout the thesis. We also detail model performance and evaluation methods implemented in conjunction with common statistical tests applied.

2.2.1 Linear Regression

Linear regression is a standard statistical modelling method. Linear regression is the analysis estimate of the 'y' outcome values due to a range of independent variables values 'x' [50]. Linear regression may either be a simple linear regression (i.e., one feature) or multiple linear regression (i.e., features >1). The objective is to model the linear relationships between the independent variable x and the dependent variable y, which will be analysed as shown in Equation 1.

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

2.2.2 Logistic Regression

Logistic regression (LR) is the most commonly used statistical method implemented in medical decision support [51]. LR is a well-established statistical model that, in its basic form, uses a logistic function to model a binary response in terms of a set of feature variables. In general, LR models calculate the class memberships using maximum likelihood estimates to determine the model's parameters. The maximum likelihood method is designed to maximise the likelihood of obtaining the data given its parameter estimates [52]. From these parameters, we can find associations between predictors and outcomes in the form of odds ratios and confidence intervals.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (2)$$

Where X = $X_1, \ldots, X_p$ are p predictors. This can be rewritten as:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+ \ldots + B_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+ \ldots + B_p X_p}} \quad (3)$$

To fit the model, we use a method called maximum likelihood to estimate the parameters β_0 & β_p Since it has better statistical properties than other methods, such as the nonlinear least-squared regression method. The general approach to the method is to calculate β_0 & β_p such that the predicted probability $\hat{p}(X_i)$ for each instance, corresponds as closely as possible to the instance observed default status (true value)[53].

2.2.3 Random Forest

RF is a relatively novel ML algorithm, which is a collection of classification and regression trees (CART) [54] that operate as a collective. Each tree in the mechanics of the RF forecasts a prediction and the class with the most votes become our model's collective prediction [55]. Furthermore, due to RF specific rules for its classification trees, it is robust to overfitting and is considered more stable in the presence of outliers and high-dimensional data compared to other ML algorithms [56].

RF model mechanics steps.

1. *Draw $ntree$ bootstrap samples from the original data.*
2. *Grow a tree for each bootstrap dataset. At each node of the tree, randomly select $mtry$ (m) variables for splitting. Grow the tree so that each terminal node has no fewer than node size cases.*
3. *Aggregate information from the $ntree$ trees for new data prediction, such as majority voting for classification.*
4. *Compute an out-of-bag error rate by using the data, not in the bootstrap sample.*

An example of the architecture of an RF model is shown in Figure 2.

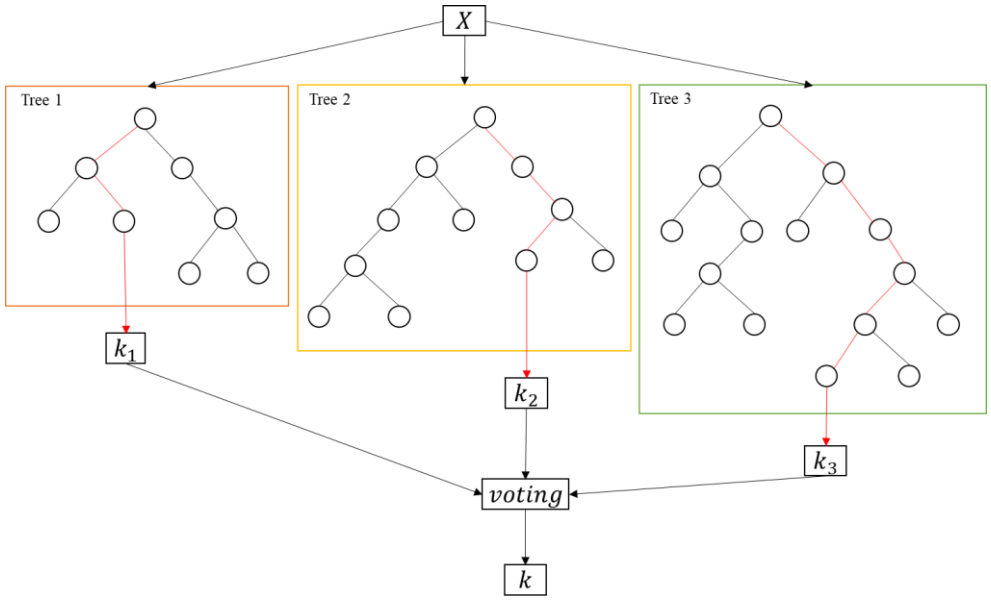


Figure 2: General architecture of the random forest model.

There are two by-products of RF, the first being the test set error estimate, and the second is the variable importance. In our studies, the Gini index was the primary variable importance metric. Given a node t and estimated class probabilities $p(k|t)$ $k = 1, \ldots, Q$, the Gini index is defined as:

(4)

$$G(t) = 1 - \sum_{k=1}^{Q} p^2(k|t)$$

Where Q is the number of classes.

RF tree-based components are grown from a certain amount of randomness. RF provides an improvement over bagged trees by way of a minor tweak that de-correlates the trees. As in bagging, we construct a number of trees based on a bootstrapped training sample. However, when building these trees, each time a split is considered, a random sample of *m* predictors is chosen as a split candidate from the complete set of *p* predictors. This split allows for the use of those *m* input variables only. A fresh sample of predictors is taken at each split, with *m* being selected by Equation 6. Where *m* is the number of random variables chosen from the candidate variables *p*, the square root of the total dimensions of the input data. Therefore, the split allows for only one of these *m* predictors to be used when generating its trees [57].

$$m = \sqrt{p} \tag{5}$$

Therefore, RF, simply at each split in its tree, is not allowed to consider a majority of the available variables. By using only, a subset of predictors, this overcomes the problem of particular input variables dominating the trees produced by the RF, which like bagging alone, can cause fundamental issues. As the algorithm is only considering $\frac{(p-m)}{p}$ of the splits will not consider the stronger predictors, so other predictors will have more of a chance to show their importance in the model.

Gradient Boost Machines

To design a gradient boost machine (GBM) for a given task, the loss function and hyperparameters need to be specified in order to choose the optimal GBM model for the given application. Like decision trees and random forests, GBM can be applied to a multitude of tasks, both classification and regression. The GBM model framework provides the practitioner with much design flexibility. A particular GBM can be designed with different base learner models. A diverse range of base learners has been displayed in literature. The three main branches are linear models (linear regression, Ridge penalised and random effects), Smooth models (P-splines and Radial basis functions), Decision trees (decision tree stumps and decision trees with arbitrary interaction depths) and other models such as Markov random fields, wavelets and other custom base learner functions [58].

As there are many implementations of GBM, we will restrict our outline regarding the context of decision trees, as this was the implementation used through the analysis conducted.

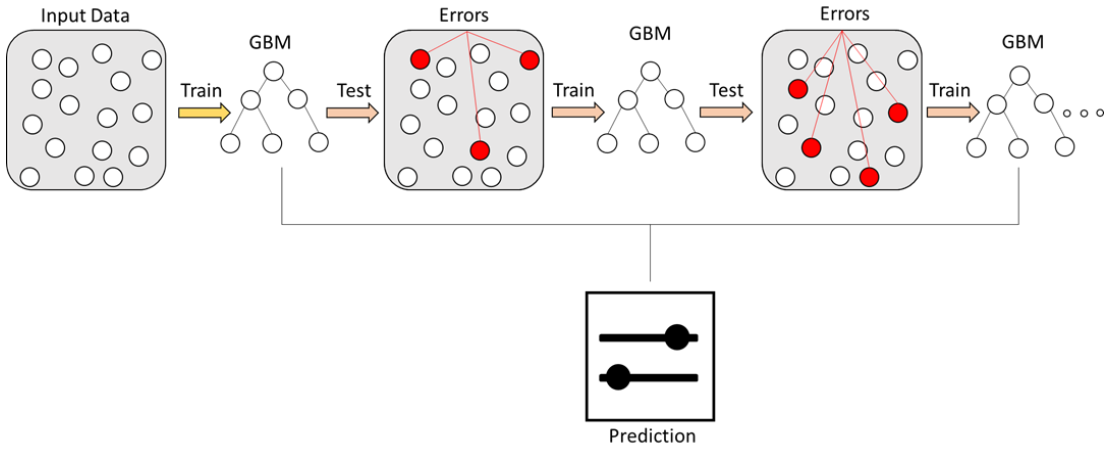


Figure 3: GBM depiction of using information from previously grown trees to reduce prediction error.

GBMs work similarly to bagging, except that the trees are grown sequentially: each tree is grown using information from previously grown trees. Boosting does not involve bootstrap sampling; instead, each tree is fit on a modified version of the original data set.

GBM model mechanics steps.

- *Set* $\hat{f}(x) = 0$ *and* $r_i = y_i$ *for all* i *in the training set*
- *For* $b = 1, 2, \ldots, B$, *repeat:*

a) Fit a tree $\widehat{f^b}$ *with d splits* ($d + 1$ *terminal node*) *to the training*

b) Update $\hat{f}$ *by adding in a shrunken version of the new tree*

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \widehat{f^b}(x)$$

c) Update the residuals

- *output the boosted model (Equation 7),*

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \widehat{f^b}(x) \tag{6}$$

Unlike fitting a single large decision tree to the data, which amounts to fitting the data hard and potentially overfitting, the boosting approach instead learns slowly. We fit a decision tree to the residuals from the model. That is, we fit a tree using the current residuals rather than the outcome Y, as

the response. We then add this new decision tree into the fitted function in order to update the algorithm. Each of these trees can be relatively small, with just a few terminal nodes, determined by the parameter d in the algorithm. By fitting small trees to the residuals, we slowly improve $\hat{f}$ in areas where it does not perform well. The shrinkage parameter λ slows the process down even further, allowing more different-shaped trees to attack the residuals [57].

2.2.4 Neural Networks

Artificial Neural Networks (ANN), commonly referred to as neural networks (NN). There are many types of NN, and the type is dependent on the architecture, for example, Feed-forward neural networks (FNN), recurrent neural networks (RNN), Boltzmann machines (BM) and convolutional neural networks (CNN). ANNs are inspired by biological neural networks. A multilayer perceptron neural network (MLP) is a basic architecture of ANNs and can form a deep NN by stacking multiple hidden layers [59].

Simply put, neural networks are connected graphs with input neurons, output neurons and weighted edges, which are loosely modelled after the human brain. ANN has input and output neurons which are connected by weighted synapses. The weights are affected by how much forward propagation passes through the network. The weights can be changed and adjusted during backpropagation [60]. The way the NN operates is dependent on some critical fundamental hyperparameters. This is briefly summarised in Table 1.

Table 1: Hyperparameter descriptors for neural network

Hyperparameter	Description
input nodes/nodes	No computation is done at this layer, usually a fixed size dimension of nodes which passes the information to the next layer in the NN.
hidden layer /nodes	There can be a series of hidden layers in a NN (deep neural network); this is where the set computation is done for the given application. The performed computation transfers the weights (information) to the following later in the network.
output layer/nodes	This layer is the final layer of the NN and is commonly used with an activation function that maps the desired output format (sigmoid or softmax for classification).
activation function	The activating function takes a single number from the output and performs a specific fixed mathematical operation depending on the activation function selected to give an output in the desired range. Common activation functions are Sigmoid, Tanh, Relu, and softmax.
learning rule	The learning rule is an algorithm that modifies the parameters of the NN for a given input to the network to produce a favoured output, typically modifying the weights and thresholds.

2.2.5 Convolutional Neural Networks

Convolutional neural networks (CNN) generally refer to 2-dimensional CNNs commonly used for image data analysis. However, there are model variants of the CNN implementation, being both 1-dimensional and 3-dimensional CNNs. The CNN type will depend on the data structure inputted into the model. The name CNN comes from the mathematical operation implied in the network called convolutions. Convolution is a specialised kind of linear operation, which CNN deploy. A CNN is simply an NN that uses convolutions in place of general matrix multiplications in at least one of their layers. CNNs have three main advantages over other networks, namely, parameter sharing, sparse interactions, and equivalent representations. First, to fully utilise the 2-dimensional structure of an input data (e.g., grid-like topology, such as image and video), local connections and shared weights in the network are utilised, instead of traditional fully connected networks, a bi-product of this architecture is fewer parameters, therefore, is faster to train [61].

A CNN commonly consists of three layers; a convolutional layer, a subsampling layer(pooling) and a fully connected layer (identical to an MLP) [62]. The CNN layer uses the convolution operation to achieve the weight sharing, while the subsampling is used to reduce the dimensionality. CNN aims to learn abstract features by alternating and stacking convolutional kernels and pooling operations. In CNN, the convolutional layers (convolutional kernels) convolve multiple local filters with raw input data and generate invariant local features, and the subsequent pooling layers extract the most significant features with a fixed length over sliding windows of the raw input data [63].

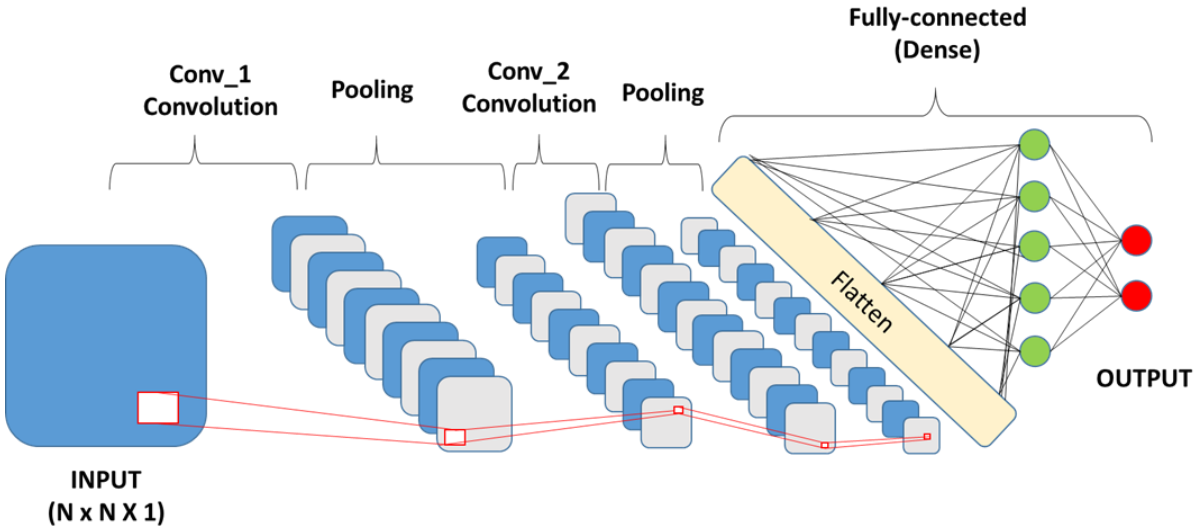


Figure 4: 2D CNN for image processing applications.

The layer in CNNs has inputs x arranged in three dimensions, $m * m * r$ where m refers to the height and width of the input, and r refers to the depth (channels, e.g., RBG: r = 3) for a 2D CNN as displayed in Figure 4. In each convolutional layer, there are several filters (kernels) k of size $n * n * q$. The filters are the base of local connections that are convolved with the input shape and share the same parameters in terms of weights (W^k) and bias (b^k) to calculate the feature map (h^k). Again, similar to an MLP the convolutional layer computes a dot product between the weights and its inputs; however, the inputs are small regions of the original input data. Then the activation function f is applied to the output of the CNN layer.

$$h^k = f(W^k * x + b^k) \tag{7}$$

After that, in the subsampling layer of the CNN, each feature map is down sampled to decrease the parameters in the network, increasing computational efficiency and controlling for overfitting. Then, the pooling operation (e.g., max or average) is calculated over a $p * p$ (where p = filter size) continuous region for all the feature maps. Next, the low/midlevel feature generated from the CNN layer is usually passed to a fully connected layer to generate the high-level abstraction from the data. Lastly, the final layer can be used to generate the desired output. A softmax or sigmoid activation function for classification tasks may be applied to calculate the associated probability [64].

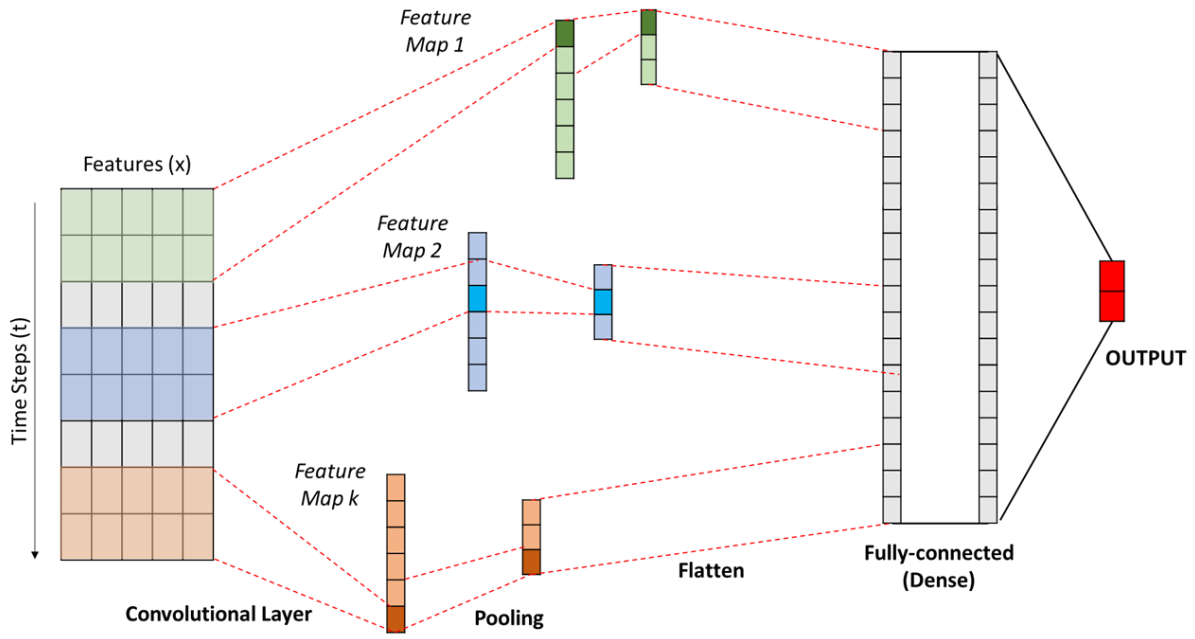


Figure 5: 1D CNN for multivariate time series applications.

2.2.6 Feature Selection Methods

A feature is an individual measurable property of the process being observed. Using a set of features, a battery of ML algorithms can be conducted for classification or regression. In recent years, the number of features has expanded from tens to hundreds of variables for a given application. Several techniques and approaches have been developed to address the problem of reducing irrelevant, redundant variables that burden challenging tasks. Feature selection helps in understanding data, reducing computational costs, the effect of the curse of dimensionality and improving the model's predictive performance[65].

The focus of feature selection is to select a subset of variables from the input data, which can effectively describe the input while reducing the effects from noise or irrelevant variables while leveraging good predictive performance [66]. There are three main classes of feature selection; subset selection (identifying a subset of the p predictors that are significant to the response), shrinkage(also known as regulation, which uses all predictors in the model, however, the estimated coefficients are shrunk towards zero or near zero depending on the penalisation parameter $L1$ or $L2$), and dimensionality reduction(this involves projecting the p predictors into a M- dimensional subspace where $m < p$) [57].

2.2.6.1 Stepwise Logistic Regression

Backwards stepwise regression (BSR) is a subset selection method used to find the optimal number of predictors from a set of feature variables. BSR begins with a full least squared model containing all p predictors, which then iteratively sequentially removes the least important variable.

BSR model mechanics

1. let M_p *denote the full model, which contains all p predictors*
2. For k = p, p -1, … ,1:
   (a) Consider all k models that contain all be one of the predictors in M_k , for a total of k -1 predictors
   (b) Choose the best among the k models and call it M_{k-1} . Here best is defined as having error metric (relating to the target)
3. Select a single best model from among M_0 , … , M_p using cross-validated prediction error: C_p (AIC), BIC, or R^2

In general, the training error will decrease as variables are removed from the model, but the test error may not. The AIC is a statistical method based on in-sample fit to estimate the likelihood of a model to predict future events. The optimal model has the minimum AIC compared to its subsets[67].

Like BSR, forward stepwise regression (FSR) is the same procedure but the opposite. With FBS, rather than starting with a full model, we start with a null model and add a variable to the model with each iteration, increasing the model's performance and minimising the test error until a plateau is reached.

2.2.6.2 Random Forest

RF is often used for variable selection, as the tree-based strategies used by RF rank variables by how well they improve the purity of a node. We used the Gini index as the primary metric of variable importance. The Gini index is a measure of the prediction power of variables in classification domains based on the principle of impurity reduction, which is non-perimetric, therefore, does not rely on the data belonging to a particular type of distribution[68].

2.2.6.3 Gradient Boost Machine

A benefit of using GBM is that after the boosted trees are constructed, retrieving the variable important scores is relatively straightforward. Moreover, features are selected sparsely following and important chance in the impurity function: splitting on new features is penalised by a cost of $\lambda > 0$, whereas re-used of previously selected features incurs no additional penalty. Thus, GBM has several compelling properties. Firstly, as it learns an ensemble of trees, it can naturally discover nonlinear interactions between features. Secondly, unlike RF, it unifies feature selection and classification into a single optimisation loss [69]. Finally, similarly to RF, the Gini index is the primary metric of variable importance.

2.3 Statistical Tests

Throughout the thesis, various statistical tests have been used to compare the summary statistic generated by the data summary tables in the following chapters. The general idea is to use statistical tests to investigate the data at a feature level. The goal of all statistical tests is to determine whether two (or more) variables are associated with one another or independent from each other at the population level. When applying statistical tests, the data in practice and the outcomes we want to measure must be considered. In order to apply the correct statistical test, the data structure is measured (i.e., categorical/continuous). The next factor is the distribution of the data (i.e., gaussian, binomial, Poisson, skew). When applying statistical tests, the last thing to consider is whether the data is matched, indicating that our sampling subjects or data points relate to one another or are independent. If these three factors are considered, we can choose the appropriate statistical test and therefore eliminate any unsuitable statistical approaches [70].

2.3.1 Chi-Squared Test

Pearson's Chi-Square (χ^2) is a nonparametric statistical test to determine if two or more classifications of samples are independent. Chi-squared tests can be applied to only discrete data. However, continuous variables can be used only if they are put into a discreet form by using intervals on a continuous scale (i.e., Age groups). Though the data structure presented must not be continuous [71].

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

Where: O = observed, E= Expected and χ^2 = the chi-square value

$$E_i = \frac{M_r * M_c}{n} \tag{9}$$

Where: M_r = the represents the row marginal for that cell, M_c represents the column marginal for that cell, and n = the total number of samples.

The χ^2 test is essentially the sum over the categories of the squared and standardised differences between the observed and expected frequencies, formulated under the assumption that the null hypothesis is true. In general, under the null hypothesis, this test statistic has an asymptotic χ^2 distribution with degrees of freedom equal to the number of categories minus the number of parameters, if any, that need to be estimated to form the expected frequencies [72].

2.3.2 Kruskal-Wallis Test

The Kruskal-Wallis test (KWT) is a nonparametric statistical test that assesses the differences among three or more independently sampled groups on a single, non-normally distributed continuous variables. Non-normally distributed data is suitable for the KWT. The KWT is an extension of the two-group Mann-Whitney U (Wilcoxon rank) test. Thus, the KWT is a more generalised form of the Mann-Whitney test and is a nonparametric version of the one-way analysis of variance (ANOVA) [73].

$H_0: \theta_1 = \cdots = \theta_k$ Versus $H_1: at\ least\ two\ \theta_i\ are\ unequal$

Reject H_0 for large values of

$$H = \left[\frac{12}{N(N+1)}\right] * \sum_{i}^{k} \frac{R_i^2}{n_i} - 3(N+1) \tag{10}$$

Where $i, i = 1, \dots, k$ number of independent groups. Let R_i denote the sum of the ranks assigned to the observations from groups i. in the rank of the i.i.d variables [74].

2.4 Performance Metrics

A metric of measure is needed to evaluate the model's performance to examine if the model is operating accurately. For different methods, such as regression, the mean square error (MSE) may be calculated or, for classification tasks, a confusion matrix. In addition, ML models for clinical outcome predictions often utilise aggregate discriminative metrics such as the area under the receiver operator characteristic curve (AUC-ROC). However, metrics such as accuracy can deliver an unrealistic measure of performance accuracy when data contains substantial class imbalances.

In this thesis, the primary metric for classification was AUC-ROC. The AUC measures the entire two-dimensional area under the ROC from integral approximations. In addition, several standard metrics can be calculated from a confusion matrix concerning classification [75].

- Sensitivity: Measures the ability to correctly identify those cases with a positive class.
- Specificity: Measures the correctly identified negative case.
- Precision: The proportion of the predicted cases which were positive that were correct.
- Accuracy: Measures the proportion of true positives and negative cases that were correct.
- F1: Measures the harmonic mean of the model's precision and recall.

(11)

$$Sensitivity|Recall: \frac{TP}{TP+FN}$$

$$Specificity: \frac{TN}{TN+FP}$$

$$Precision: \frac{TP}{TP+FP}$$

$$Accuracy: \frac{TP+TN}{TP+FN+FP+TN}$$

$$F1: 2 * \frac{Precison * Recall}{Precison + Recall}$$

In classification problems, we can predict the classed output as a binary classifier or alternately, it can be more flexible to predict the probability for each given class. This provides the capability to choose and calibrate the threshold for a risk prediction model and how interpretable the response probabilities are. In addition, this ability allows the threshold to be adjusted to tune the model's behaviour for a specific problem.

2.4.1 AUC-ROC

The area under the receiver operating characteristic (AUC-ROC) is a performance measure for the classification problem at various thresholds. ROC is a probability curve, and the AUC score measures how well the model is capable of distinguishing between classes. Therefore, the higher the AUC score, the better it will be at distinguishing between classes. The area under the ROC curve is currently considered the typical method to assess the accuracy of predictive classification models. It evaded the supposed subjectivity in the threshold selection process when continuous probability-derived scores are converted to a binary presence-absence variable by summarising overall model performances over all possible thresholds [76].

The AUC-ROC, or simply AUC, measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'). AUC varies between zero and one, with an uninformative classifier yielding 0.5 [77].

$$\begin{aligned} &\textit{True positive rate: } (Sensitivity): \frac{TP}{TP+FN} \\ &\textit{False positive rate: } (1-Specificity): 1-\frac{TN}{TN+FP} \end{aligned} \tag{12}$$

In this thesis, the primary metric for regression was RMSE; however, a range of regression metrics were calculated. The root mean square error (RMSE) has been used as a standard statical metric to measure model performance throughout our studies. The mean absolute error (MAE) is another useful measure widely used in model evaluation. While they have been used to assess model performance for many years, there is no consensus on the most appropriate metric for model errors. When both metrics are calculated, the MAE tends to be much smaller than the RMSE because the RMSE penalises large errors while the MAE give the same weight to all errors[78]. Mean squared error (MSE) is also a popular choice of loss function for regression tasks; however, it is more sensitive to outliers due to its quadratic nature. Thus, MAE is sometimes employed as an alternative to MSE. The formulae for each loss function are displayed below.

(13)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}{e_i}^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}{e_i}^2}$$

2.5 Model Validation

Cross-validation is a model evaluation method. This section will discuss the validation methods for testing the models implemented throughout the thesis.

2.5.1 Holdout-Method

The holdout method is when the data is randomly partitioned into two sections, 'training' and 'testing'; however, most commonly, this is split into three; training, validation, and testing. The holdout method is a basic cross-validation method used to fit the model. The validation set estimates the prediction error and reduces the loss function applied. Finally, the test set is used to assess the generalised error of the final chosen model parameters. However, the disadvantage of this method is that it is usually preferable to the model's loss function[79]. This implies that the test error could be highly variable in this validation approach as only a subset of observations are included in the training set rather than in the validation set that is used to fit the model. This suggests that the validation set error may tend to overestimate the test error of the model fit to the data.

2.5.2 K-fold Cross-Validation

A way to reduce the variance in the model's predictive performance is to use k-fold cross-validation. This holdout method partitions the data into 'k' folds, training and testing on each of the 'k' folds. The average error across the 'k' folds is computed and taken as the true model's performance. This provides an excellent solution to the bias-variance trade-off, allowing for a more accurate representation of the model's performance, however, at a computational time cost [80].

2.5.3 Nested K-Fold Cross-Validation

Nested cross-validation is commonly used to train a model in which hyperparameters must also be optimised. In our case, for models such as RF, GBM and the implementation of a forward sequential search algorithm for feature selection. In each fold of the outer cross-validation, the model's hyperparameters are tuned independently to minimise an inner cross-validation estimate of the performance. This eliminates the bias introduced by the inner cross-validation procedure as the test data in each iteration of the outer cross-validation has not been used to optimise the performance of the model in any way and may, therefore, provide a more reliable criterion for selecting the best model [81].

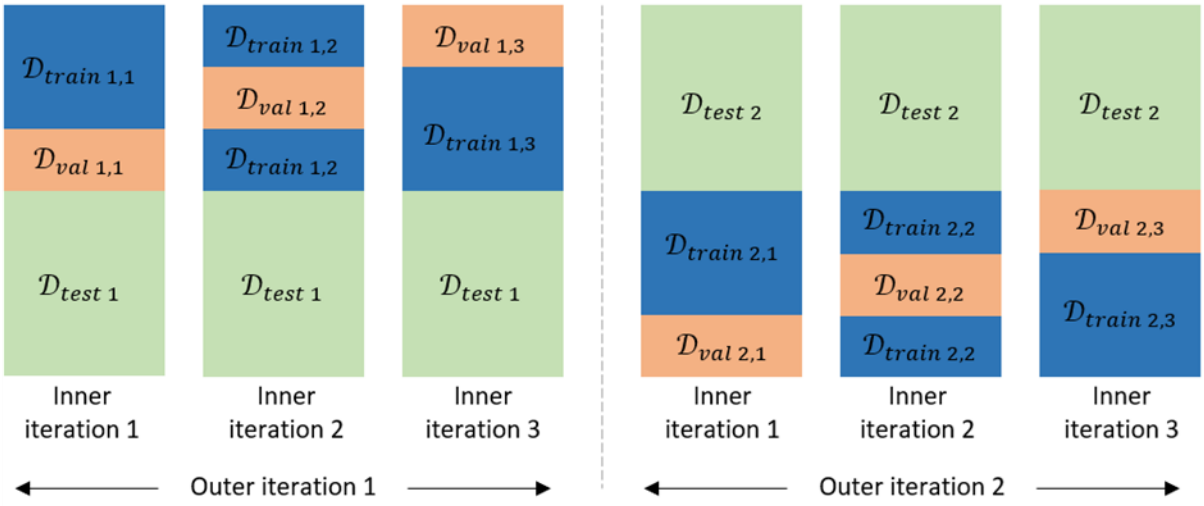


Figure 6: The example displays a 2x3-fold nested cross-validation. $Dtrain$, $Dval$ and $Dtest$ represent a proportion of the training, validation and test data used in each iteration.

2.6 Dimensionality Reduction

Statistical and ML reasoning faces a challenging problem when dealing with high-dimensional data. Typically, the number of input variables is reduced before an ML algorithm can successfully apply. Typically, dimensionality reduction can be performed in two ways, firstly by only keeping the most relevant variables from the original dataset or by exploiting the redundancy of the input data and by finding a smaller set of new variables, each being a combination of the input variables containing basically the same information as the input variables. The critical idea of dimensionality reduction is finding a new coordinate system in which the input data can be expressed with fewer variables without a significant error [82]. DR algorithms aim to reduce the distance between distributions of different data sets in a latent space to allow efficient transfer learning [83]. Furthermore, the finding with DR is much preferable to those without decreased dimensionality [80]. Moreover, the low dimensional data representation of the initial tends to overcome the issues of the dimensionality curse and can be easily analysed, processed and visualised [85], [86].

2.6.1 PCA

The most widely used dimensionality reduction technique is Principal Component Analysis (PCA). PCA [87] is an algorithm for DR based on the maximisation of variance in a lower-dimensional projected space. The original data (formula) are presented by the product of two matrices, namely the scores (T~(n,k)) and loading (P~(p,k))

$$X = TP^T + E \tag{14}$$

Where E~(n,p) is the residual matrix and n,p, and k are the numbers of samples, variables, and components, respectively. The parameters are estimated to capture as much of the variance in the original data in a least squares sense and further to be orthogonal matrices, i.e.,

$$\{T, P\}\ argmax_{T,P}\ (\|X - TP^T\|_2^2) \tag{15}$$

The combination of vectors of T and P is referred to as principal components and used in various ways, e.g., exploratory data analysis to map the multivariate sample distributions as well as interrogating feature2feature correlation structure, furthermore, to represent the data in a few meaningful features used for further analysis. For example, a rewrite of the equation above shows that the score space (T)

is a linear mapping by the orthogonal basis represented by P: T= XP, and hence a coordinate system rotation[88].

2.6.2 UMAP

Uniform manifold approximation and projection (UMAP) has tackled the problem of DR by generalising nonlinear approaches like PCA to be sensitive to possible nonlinear structures in data [89]. They developed the UMAP algorithm by applying a new field of mathematics based on Riemannian geometry and algebraic topology. Using every available data point, UMAP first creates a graph with respect to the distances on the underlying topology and the k-neighbourhood of each element. The Laplacian eigenmaps dimensionality reduction method is then applied to that graph. The resulting graph is further modified by a forced directed graph layout algorithm, which minimises the cross-entropy between this modified graph and the original one. In this manner, the resulting low-dimensional data representation is optimised to preserve the original data's local and global structure. The main advantage of UMAP over PCA is that it can capture a more complex (nonlinear) structure in high-dimensional data, which is a desirable characteristic. UMAP can achieve this by initially constructing a high-dimensional graph representation of the original data, followed by optimising a low-dimensional graph to be as structurally similar to the original as possible. In this manner, the resulting low-dimensional data representation is able to preserve well both the local and global structure of the original data[88].

2.7 Clustering

Clustering is a branch of unsupervised learning and can be categorised as finding structures in a collection of unlabelled data. A cluster is, therefore, a collection of objects that have “similar” characteristics and are “dissimilar” to the objects belonging to other clusters [86]. Furthermore, clustering algorithms can be hierarchical or partitional. Hierarchal algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters simultaneously through an iterative process. Partitioning algorithms are based on specifying an initial number of groups and iteratively reallocating objects among groups to convergence.

2.7.1 K-Means

K-means is one of the simplest unsupervised learning algorithms that solve well-known clustering problems. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed prior. The K-means algorithm assigns each point to the clusters whose centroid (centre) is nearest. The centre is at the average of all the points in the cluster. Therefore, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster [90], [91].

The main idea is to define k centroids, one for each cluster. These centroids should be placed strategically as various location causes different results in the clusters class assigned. Therefore, the optimal choice is to place them far away from each other as much as possible. The next step is to take each point belonging to a given data set and associate it with the nearest centroid. The first step is completed when no point is pending, and an early group is completed. At this point, it is necessary to re-calculate k new centroids as centres of the clusters resulting from the previous step. After these k new centroids, a new binding has to be done between the same data points and the nearest new centroid. A loop has been generated. As a result of this loop, it may notice that the k centroids change their location step by step until no more changes are calculated. In other words, until the centroids do not move. Finally, this algorithm aims at minimizing an objective function, in this case, a squared error function. The objective function

(16)

$$W(S,C) = \sum_{K=1}^{K} \sum_{i \in S_k} \|y_i - c_k\|^2$$

Where S is a K-cluster partition of the entire set represented by vectors y_i ($i \in$I) in the m-dimensional feature space, consisting of non-empty non-overlapping clusters S_k, each with a centroid c_k ($k = 1,2, \dots K$).

K-means model mechanics steps.

1.) Place K points in the space represented by the objects that are being clustered. These points represent the initial group centroids.
2.) Assign each object to the group that has the closest centroids.
3.) When all objects have been assigned, recalculate the positions of the k-centroids.
4.) Repeat steps 2 and 3 until the centroids no longer move.

3 CHAPTER 3: Database Description, Exploration and Analysis

3.1 Introduction

Intensive care units (ICUs) provide complex and resource-intensive treatments for the sickest hospitalised patients. The need for critical care medicine has grown substantially over the past decade [92] and has consumed a vast portion of the income in many countries worldwide [93]. As a result, the healthcare industry has significantly transitioned over the past decade from a paper-based domain to one operating primarily through a digital medium. Beyond the logistical benefits of maintaining and organising patient medical records, the ability to quickly identify and process information from millions of patient records, laboratory reports, imaging procedures, payment claims, and public health databases has brought the industry to the precipice of a significant change. Namely, the opportunity to utilise data science and ML methodologies to address problems across the practice and administration of healthcare. Using such analytical techniques has provided a foundation on which personalised and predictive care models have emerged [94]. These models represent many opportunities, from improved patient stratification to identifying a novel disease, comorbidities and drug interactions, to predicting clinical outcomes [95].

Despite the considerable investments in critical care medicine, medical resources in ICU are usually insufficient to meet the demands of ICU patients. As a result, hospitals are under pressure to improve their efficiency and reduce costs for critical care. Length of stay (LOS) in the ICU is a crucial indicator of medical efficiency [96] and critical care quality in hospitals [97]. Similarly, another popular outcome measured is mortality risk, combined with LOS, which generally allows medical institutions to predict the resources and medical costs of a patient's admission[98], [99]. Therefore, early identification of LOS and mortality can provide an important reference for patients and an essential indicator for optimal clinical intervention.

Data preprocessing has a significant impact and effect on the generalisation performance and inference gained from an ML algorithm. It is well-known and recognises the importance of data preprocessing and the steps required, which takes significant time and development to complete successfully. Although ML has shown promising applications to healthcare analytics through personalised and predictive care, there are still obstacles intrinsic to the data being evaluated and the population from which the data is drawn[100]. In many cases in healthcare, the datasets are enormous, complicated or may be prone to a particular issue or limitation. However, generally, when modelling, we aim to either generate some preliminary insight related to a specific domain or outcome in predictive performance or inference gained depending on the study objective.

Clinicians and medical professionals require quick, accurate information to provide care effectively. Therefore, the need to collect and produce high-quality data has become paramount for the application of ML. While the challenges to preprocessing are present in many domains, the dynamics of healthcare necessitate that care is taken to address a number of biological, computational, and representational aspects of the data. In many cases, various medical datasets contain numerical and categorical data from various data sources and structures. This raises questions, such as which set of data gives the most useful information and which data features should be chosen when modelling. Similarly, we often question which sample is appropriate to choose or how large this subset of data needs to be. We must first overcome these obstacles before successfully deploying an ML algorithm to gain insight into the outcome objective.

The data used throughout this thesis is derived from commonly open-source databases used in intensive care, including the MIMIC-III Clinical/Waveform database, eICU Collaborative Research Database [101](eICU) and the AmsterdamUMC (AUMC) database. All three data sources are relational databases, which can be utilised and loaded into any database management system. The MIMIC-III contains 26 comma-separated values (CSV) files with a primary key of 'HADM_ID' interlinking all data sources at the level of hospital admission. The eICU contains 31 CSV files and a primary key of

'PatientUnitStayID' at the level of ICU admission. Lastly, the AUMC contains the least amount of data files, with only 7 CSV files with a primary key of 'admissionid', similarly, at the level of ICU admission. All databases are deidentified following Health Insurance Portability and Accountability Act (HIPAA) [102]. This includes the removal of all protected health information, dates, ages over 89, and person numbers, amongst others, such as removing free text from diagnostic reports and physician notes. Different data sources used in the analysis can render different performances. This can be caused by a multitude of factors: different data collection processes, information granularity, and features available, amongst different geographic factors, such as if the data sources come from multiple centres compared to a single centre with fewer generalisation capabilities. The supplementary materials can view an overview of the relational database structures used in this thesis. A brief comparison of the AUMC, eICU and MIMIC-III is shown in Table 2.

Table 2: Summary characteristics of eICU, MIMIC-III and AUMC.

Items	**AUMC**	**eICU**	**MIMIC-III**
County	Netherlands	United States	United States
Data	Single centre	Multi-centre	Single centre
Year	2003-2016	2014-2015	2001-2012
Number of Units	1	335	1
Number of Hospitals	1	205	1
Number of patients	20,109	139,367	38,597
Number of admissions	23,106	200,859	53,423
Deidentification	All protected health information was deidentified, and no patient's private data can be identified		
Data Content	Vital sign measurements, laboratory tests, care plan documentation, diagnoses information, treatments information, and others		

This chapter explores the three databases in a comparative analysis, comparing data availability, application modelling practicality, and predictive performance. We investigate In-ICU mortality and LOS to examine the heterogeneity in the data among the data sources. To compare model performance metrics, we focused on two commonly reviewed outcomes of interest, ICU mortality and ICU LOS. This was deemed a fair comparison as the commonly used in-hospital mortality indicator and hospital LOS are not available in all data sources, as the AUMC does not contain this information which is a limiting factor. This database strictly provides information solely on ICU admission, consequently directing the comparison in this way. Additionally, this chapter will address the complexities of healthcare data as they impact the ML algorithms which consume them. Broadly speaking, we compartmentalise such work into three major categories, each representing a component of the ML pipeline, as seen in Figure 7.

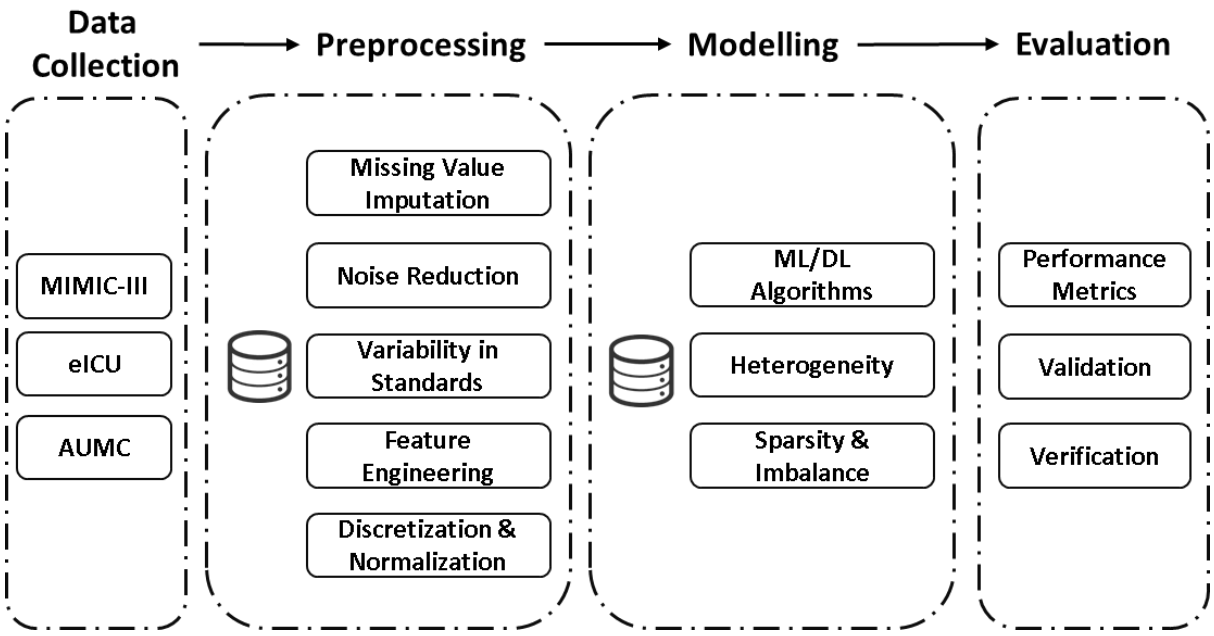


Figure 7: Machine learning pipeline, processes and consideration when modelling critical care data.

3.2 Data Sources Description

3.2.1 Ethics Statement

All legal and user agreements required to use the databases have all been accepted prior to any analysis and are stored on secure servers at Liverpool John Moores University. The analysis is unrestricted once a data user agreement is accepted, enabling clinical research to be undertaken. Researchers and institutes seeking access to the database must request permission, as although the data is anonymised, it still contains detailed information regarding the medical care of the patients, therefore, must be treated with appropriate care. Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. In addition, written informed consent for participation was not required throughout all research in accordance with the national legislation and institutional requirements. Therefore, pose no conflicting interests with the Liverpool John Moores code of practice for research.

3.2.2 eICU Collaborative Research Database

The eICU collaborative research database (eICU) (V2.0), released on 17th May 2018, is populated with data from a combination of many critical care units throughout the continental United States (US). The data from the eICU database covers patients admitted to critical care units in 2014 and 2015. The eICU database is a multicentre intensive care unit database which contains highly granulated data from over 200,000 patient ICU admissions, monitored by eICU programs encountering 139,367 unique patients admitted to one of 335 care units from 208 hospitals across the US. In addition, the database includes vital sign measurements, care plan documentation, severity illness measures, diagnoses information, treatment information and more.

3.2.3 MIMIC-III Clinical Database & Waveform Database

Medical Information Mart for Intensive Care (MIMIC-III) (V1.4) is an extensive single-centre database comprising information relating to 60,000 ICU episodes of 53,423 unique patients admitted to Beth Israel Deaconess Medical Centre in Boston, Massachusetts, between 2001 and 2012. MIMIC-III contains high temporal resolution data, electronic documentation, bedside monitor trends and waveforms, vital signs, medication, laboratory measurements, observation and notes charted by care staff, fluid measurements, procedure/diagnostic codes, imagining reports, and duration of stay, amongst others. In addition, the database supports a diverse range of analytical studies spanning from

epidemiology, clinical decision-rule improvement and electronic tool development. The MIMIC-III Waveform database (V1.0) [103], [104] contains 67,830 record sets for approximately 30,000 ICU patients. Nearly all record sets encompass waveforms recordings containing digitised signals (ECG (electrocardiogram), ABP (Arterial blood pressure), and PPG (photoplethysmogram)) and numeric recordings containing time series of periodic measurements. The ECG signals (I, II, III, AVR, V, MCL) were recorded; however, not all these signals are available simultaneously. A subset of the waveform database contains waveforms and numeric records that have been matched and time-aligned with MIMIC-III Clinical database records. The subset match contains 22,317 waveforms and 22,247 numeric records for 10,282 distinct patients. The waveform records for the subgroup matching the MIMIC-III contain a minimum of three ECG signals (II, V, and AVR) in addition to a respiration signal (RESP) and a PPG signal.

3.2.4 AmsterdamUMC Database (AUMC)

The AmsterdamUMC (V1.0) or AUMC is an extensive single-centre database from Amsterdam University Medical Centre comprising information relating to 23,106 ICU admission of adult patients between 2003-2016. The database was released in November 2019 and is freely available. However, only accessible after completing the mandatory training and guaranteeing the involvement of a practising intensivist in the research team to provide domain expertise. The database contains data from a 32-bed mixed surgical-medical academic ICU and a 12-bed high-dependency unit (medium care unit) [105]. The clinical data contains 23,106 admissions of 20,109 patients admitted from 2003 to 2016, with almost 1.0 billion clinical observations consisting of vitals, clinical scoring systems, device data and lab results data and nearly 5.0 million medication records.

3.2.5 Medical Coding Ontology & Variability

Another challenge in processing healthcare data stems not from a function of its quality but from its representation. In the effort to quantify and standardise the best set of possible conditions, procedures and clinical elements, a wide range of medical coding schemes have been developed, such as the International Classification of Disease (ICD). In many ontologies used, there is an overlap between them. Therefore, the effective processing of such data must consider the possibility that the same attribute may be represented in multiple ways or by multiple coding systems. This situation is exacerbated by the nature of healthcare systems, where in response to documentation or reporting standards, multiple coding standards may be used even within the same institution. Not only does variability arise from using different coding standards, but also from emerging diversity as these standards are revised and updated. For example, the ICD's latest revision (ICD-10) brought approximately 55,000 new diagnostic codes and over 68,000 new procedural codes[106]. Suggestions were made that healthcare organisations use ICD-9 and ICD-10 as starting points to develop their own more precise data crosswalk applications [107]. While variability is clearly a product of the expansive set of coding standards and their revisions, it also results from the medical coding methodology itself. Medical coding is a subjective process, the accuracy of which is dependent on the clinical record of the condition observed and the interpretation of the diagnostic code itself [108]. While it may be straightforward for simple cases where a patient is assigned a single diagnosis, inconsistencies from coders and institutions have been found to increase with the complexity of a patient's condition, specifically when they receive multiple diagnoses [109].

Table 3 displays the patient's top 30 most frequent diagnoses in the MIMIC-III and eICU using the medical coding ontology ICD-9 for the cohorts derived in Figure 10. Firstly, this comparison is unavailable using the AUMC as the primary diagnosis system used in this database is the Apache-IV diagnosis system to record diagnoses given to the patients. Therefore, another limiting factor in the comparisons of the databases. Although this information is partly available in the eICU, a complete Apache diagnoses comparison is still unavailable due to the partly collected nature of this data, as the eICU solely collected Apache diagnoses for ICU admission. Although both databases collect the ICU diagnoses for the patients, the structure and format of each database are considered different, as seen in Table 3, which displays the true representation of the data collected. The eICU used ICD-9 in addition

to ICD-10 to categorise the diagnoses. However, not all diagnoses have ICD codes attached and may not be consistent with diagnoses that were coded and used for professional billing or hospital reimbursement purposes.

All eICU diagnoses have a timestamp of when the diagnoses were given. This can be useful for determining if certain diseases were documented during the ICU stay and at what stages in the patient's ICU stay these diagnoses were documented. The MIMIC-III negates the first issue, as all diagnoses have an associated ICD 9 code concerning each diagnosis. However, issues remain with the MIMIC-III regarding the time the diagnosis was given. The ICD codes are generated for billing purposes at the end of the hospital stay and possess no allocated time stamp. Therefore, depending on the particular condition or diagnosis, it is difficult to determine if the patient developed a condition during that hospitalisation. For example, a patient may have pre-existing atrial fibrillation before ICU admission or may have developed it during ICU. However, the way the data is structured makes it challenging to determine using the ICD codes singularly. Due to this limitation, research has been conducted [110], [111] to gather this information using the 'noteevents' table, which contains handwritten notes from a range of medical professionals regarding the ICU duration and stay of the patient. The text associated with this data is often large and contains many newline characters: it may be easier to read if viewed in a distinct program rather than the one performing the queries. The diagnosis string attached to the eICU allows for a more granular understanding of the diagnosis given, as each '|' in the string allows the user to understand the category and some level of the rationale of the diagnosis. Although, it makes direct comparisons of the data sources difficult. Due to these listed obstacles for comparison of the data source, we will not be focusing on the use of past medical history or comorbidities as a direct comparison is not feasible among the data sources.

Table 3: Comparison of diagnoses ontologies used by the MIMIC-III and eICU for the top 30 diagnoses.

eICU		**MIMIC-III**	
diagnosisstring	icd9code	SHORT_TITLE	ICD9_CODE
pulmonary|respiratory failure|acute respiratory failure	518.81, J96.00	Hypertension NOS	4019
renal|disorder of kidney|acute renal failure	584.9, N17.9	CHF NOS	4280
endocrine|glucose metabolism|diabetes mellitus		Atrial fibrillation	42731
neurologic|altered mental status / pain|change in mental status	780.09, R41.82	Crnry athrscl natve vssl	41401
pulmonary|pulmonary infections|pneumonia	486, J18.9	Acute kidney failure NOS	5849
cardiovascular|vascular disorders|hypertension	401.9, I10	DMII wo cmp nt st uncntr	25000
cardiovascular|ventricular disorders|congestive heart failure	428.0, I50.9	Hyperlipidemia NEC/NOS	2724
cardiovascular|shock / hypotension|hypotension	458.9, I95.9	Acute respiratry failure	51881
pulmonary|respiratory failure|hypoxemia	799.02, J96.91	Urin tract infection NOS	5990
cardiovascular|shock / hypotension|sepsis	038.9, A41.9	Esophageal reflux	53081
hematology|bleeding and red blood cell disorders|anemia		Pure hypercholesterolem	2720
cardiovascular|arrhythmias|atrial fibrillation	427.31, I48.0	Anemia NOS	2859
neurologic|altered mental status / pain|pain		Pneumonia, organism NOS	486
pulmonary|respiratory failure|acute respiratory distress	518.82	Hypothyroidism NOS	2449
endocrine|glucose metabolism|hyperglycemia	790.6, R73.9	Acidosis	2762
pulmonary|disorders of the airways|COPD	491.20, J44.9	Ac posthemorrhag anemia	2851
cardiovascular|chest pain / ASHD|hyperlipidemia	272.4, E78.5	Severe sepsis	99592
cardiovascular|ventricular disorders|hypertension	401.9, I10	Chr airway obstruct NEC	496

cardiovascular|shock / hypotension|septic shock	785.52, R65.21	Septicemia NOS	389
gastrointestinal|malnutrition|protein-calorie malnutrition	263.9, E46	Food/vomit pneumonitis	5070
cardiovascular|arrhythmias|atrial fibrillation|with rapid ventricular response	427.31, I48.0	Long-term use anticoagul	V5861
cardiovascular|shock / hypotension|hypotension / pressor dependent		Chronic kidney dis NOS	5859
endocrine|glucose metabolism|hyperglycemia|stress related	790.6, R73.9	Depressive disorder NEC	311
hematology|white blood cell disorders|leukocytosis	288.8, D72.829	Hy kid NOS w cr kid I-IV	40390
cardiovascular|shock / hypotension|sepsis|severe	995.92, R65.2	Tobacco use disorder	3051
neurologic|seizures|seizures	345.90, R56.9	Thrombocytopenia NOS	2875
endocrine|thyroid|hypothyroidism	244.9, E03.9	Old myocardial infarct	412
renal|disorder of kidney|chronic kidney disease	585.9, N18.9	Septic shock	78552
cardiovascular|chest pain / ASHD|coronary artery disease		Hyposmolality	2761
renal|electrolyte imbalance|hypokalemia	276.8, E87.6	Aortocoronary bypass	V4581

3.3 The ML Pipeline

This far, we have discussed the intrinsic characteristics of data; those properties which influence the statistical foundations guiding ML theory. We now look further, not at the properties of the data, but at the mechanisms through which the data is consumed and represented to build effective ML models. Such attributes range from high-level aspects of integrating heterogeneous data types to low-level considerations when representing and increasing expansive feature space.

3.3.1 Pre-processing

Pre-processing is the first step in the ML pipeline and is characterised by techniques such as cleaning, integration, reduction and transformation [112]. The process intends to address real-world data's noisy, missing and inconsistent properties and ultimately improve data quality prior to modelling. The data preprocessing steps covered are exhibited in Figure 7. In particular, the flow that is followed before applying the learning algorithms to a particular outcome task is shown. Initially, we collect the information or receive datasets from a source. Next, the preprocessing steps are appropriately applied to clean the sample data to make it effective. Then, the data is given as inputs to the learning algorithms. Finally, they are applied to solve a specific problem (e.g., classification, regression) and measure the model's performance and application success [113]. The representations of the dataset play an important role when modelling. For example, a dataset with too many features or features with correlations should not be included in the learning process since these types of data do not offer useful information[66]. Therefore, to use the most informative information, selecting features that maximise predictive performance while minimising unnecessary and unrelated information to the study objective is necessary. In most cases, any preprocessing steps for any algorithm can be used. Regarding the field of application, the problems that can be applied to these steps include, but are not limited to, the classification and regression tasks.

3.3.2 Feature Engineering

Feature engineering is extracting features from data and transforming them into a format suitable for ML algorithms. This is broadly covered in three central field features, extraction, selection and representation. The aim when selecting features is to maintain accuracy and stability, improve the runtime and avoid overfitting the data. The selection of features in healthcare can be a mixture of recommendations of inclusion data by clinicians and algorithmic feature selection methods (e.g., filter or wrapper based [114]), which reduce the number of redundant or irrelevant data when modelling.

Utilising a large number of features increases the chance of the data overfitting the model when the number of observations is less and the need for a significant computational tie when the number of features is more. [115]. Analysis with large numbers of feature variables is computationally expensive. Therefore, we should reduce the dimensionality of these types of variables whenever possible.

3.3.3 Temporal Relations

Among the listed considerations within feature selection, we must reflect on the idea that we are consuming the accurate representation of data concerning the process it was captured. As displayed in the literature, the overwhelming majority of research modelled the outcome task of an individual as a set of discrete observations, representing patient instances in a tabular format. Each row in the dataset is attributed to a single patient admission, and the columns are attributed to a range of clinical features. Although such approaches allow the data to be easily consumed by traditional ML algorithms, such data representation is incomplete. Characterising temporal features as static features can reduce predictive performance, as the ability to utilise long and short-term dependencies in the data is lost. However, although assessing an individual's condition at every minute may be more accurate, such granularity is often not feasible. As a result, there has been an increasing focus on developing innovative ways to utilise data collection as part of existing workflows rather than demonstrating values with models requiring additional data elements. It is important to note that regardless of the analytical approach used, a patient's treatment ultimately remains in the hands of the clinician. As such, there must be a concerted effort to provide appropriate context to the results. Therefore, depending on the modelling task, this may influence the data structures and the algorithm implemented. It is clear that observations, such as a diagnosis or procedure, must occur at a single point in time. However, it would be naïve to believe that such elements of health occur in isolation. Instead, existing temporal relations connect them, representing the variable nature of an individual's health. These relations cannot be described by one feature or a single value but require longitudinal observations with a series of values over time (such as heart rate or blood pressure) [116].

3.3.4 Dynamic Data Representation

To capture the information of temporal data, such as heart rate, statistical properties extracted from the time series can be taken to capture the most relevant information. Common approaches to represent the time series are shown by calculating the mean, standard deviation, variation, minimum, maximum, skewness and kurtosis. Each approach extracts different statistical and informative properties from the time series. Ultimately all these approaches aim to represent the time interval as a single data value. Thus, the data can then be utilised in a traditional ML model such as logistic regression or random forest. Listed in Table 4 are commonly used transformations of temporal data. Commonly only a single transformation is used to represent the data. However, depending on computation limitations, it is also common to use a range of transformations such as the mean, standard deviation and skewness as different input features to represent a single feature variable in a single model (i.e., Heart rate Mean, Heart Rate SD and Heart rate Skewness).

Table 4: Commonly applied data representation for temporal data.

Feature Representation	Description
Mean	The average of the variable values during the interval of time, e.g., the average heart rate of the patient.
Standard Deviation	The standard deviation measures the amount of variation in the variable values during the interval of time. Lower values will indicate little change (values remain around the mean of the variable), and higher values will indicate higher variation in the variable values (less stable).
Minimum Value	The minimum of the variable values during the interval of time, e.g., the min heart rate of the patient.
Maximum Value	The maximum of the variable values during the interval of time, e.g., the max heart rate of the patient.

Skewness	Skewness tells us what is the shape of the time series, indicating whether the mass of the distribution is concentrated on the left (positively skewed), to the right (negatively skewed), or equally/symmetrically distributed (skewness=0). In our context, skewness can be used to identify whether the variable values increase or decrease monotonically over the time window.
Kurtosis	Like skewness, kurtosis is a statistical measure used to describe the distribution. Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures extreme values in either tail. Distributions with extensive kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of the normal distribution.

Our study converted dynamic features such as heart rate into tabular representations. For this, we first calculated the mean or the recorded events per hour (in case there were multiple records in one specific hour but not as many in another hour). Then, we calculated the mean of the hourly averages (the ones that were calculated previously). An example of this process is displayed in Figure 8. Therefore, this exact process can be applied to the listed data representations [117].

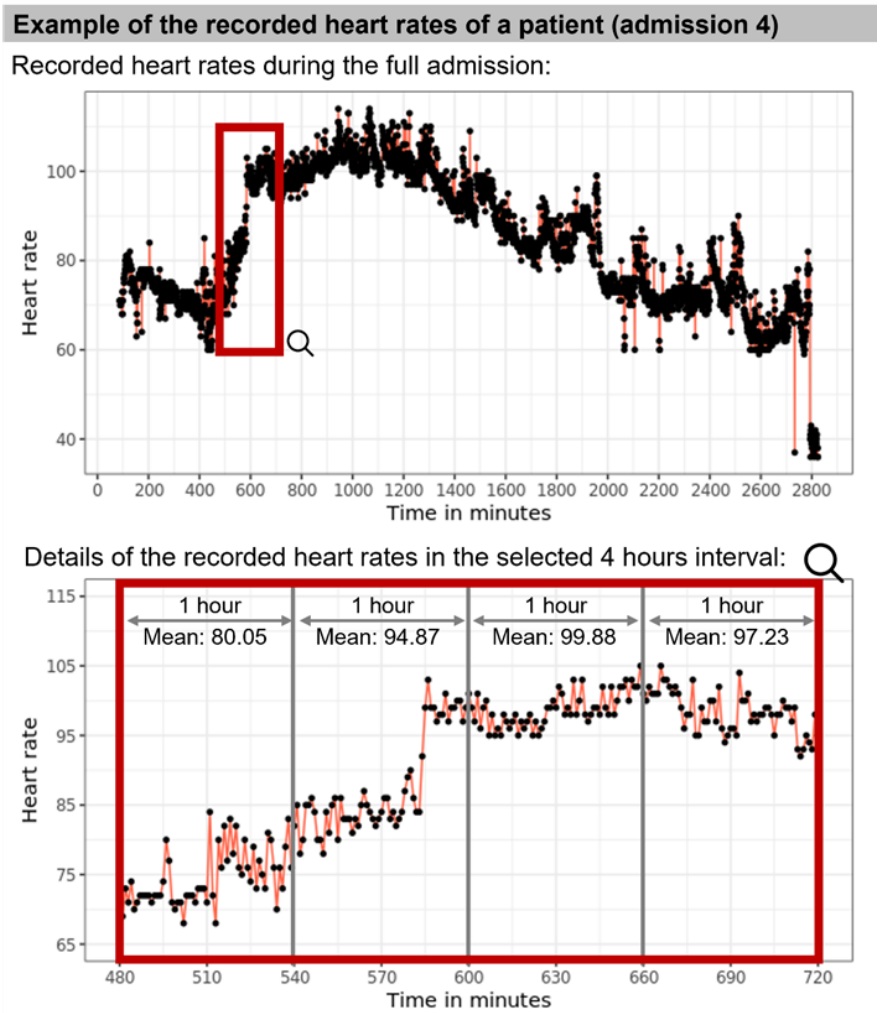


Figure 8: Dynamic features conversion into tabular representations | example of how dynamic features such as heart rate were converted into tabular representations. Firstly, the mean was calculated per hour, and then the mean of the hourly averages was calculated. Top: The recorded heart rates of one of the patients during admission. Bottom: Details of the recorded heart rates in the selected 4 hours interval (zooming in the area marked with a red rectangle on the top plot), showing how the averages per hour were calculated.

3.3.5 Missing Values, Normalization, & Imputation

There are many forms in which missing data manifests across the domain. At an attribute level, data is typically classified as missing in either three forms: completely at random, at random and not at random. However, all forms of missingness present concern, the various forms of missingness can present significantly different considerations while processing a dataset. For example, while data missing completely at random presents a minimal concern to the underlying distribution, allowing for data to be dropped or imputed without the worry of introducing additional biases, such a scenario is often unrealistic. Instead, data is typically missing due to an underlying, sometimes unobserved, pattern known as missing at random or missing not at random, each of which may require techniques such as imputation to help address the inherent bias they present to the data collected [100]. Beyond the type of missingness, the quantity of missing information further influences the preprocessing of the data. Concerning the occurrences of large temporal gaps, we find that although mathematically, we may be able to impute, model, and predict estimations of missing values during processing, there is no guarantee that the values computed accurately reflect the true condition of the individual during that time period. This consideration is particularly relevant in light of the common scenario, where data is collected during a subject's clinical encounter, which may occur months apart. Just as the types of missingness provide a roadmap to the appropriate preprocessing techniques, identifying and assessing which of the possible combinations of these three factors cause missingness to arise presents a critical step in improving the ability to address bias during the processing of such data.

In the various datasets we have to manage, and there are very often considerable differences between the feature's values, such as the maximum and minimum values. In general, this issue is undesirable and requires careful intervention to make a scaling-down transformation so that all attribute values are appropriate and acceptable. This process is known as feature scaling or data normalisation, and it is essential for various classifiers. For example, Min-max normalisation or feature scaling between [0,1] is a common form of normalisation method which scales all continuous values between [0,1], representing the lowest and highest values obtainable. Another common approach is z-score normalisation or standardisation, which essentially constitutes a measurement of how many standard deviations the value is from the mean value [113].

Table 5: List of features used in comparative analysis. This table displays where each feature is located with respect to the data source, the min and max value range, and the imputation method applied.

Features Selected	AUMC-Table	eICU-Table	MIMIC-III-Table	Min Value	Max Value	Imputation
Gender	Admissions	Patients	Patients	Nan	Nan	Mode
Age Group	Admissions	Patients	Admissions	Nan	Nan	Mode
Heart Rate	numericitems	NurseCharting	Chartevents	0	375	Median
BP Systolic	numericitems	NurseCharting	Chartevents	0	375	Median
BP Diastolic	numericitems	NurseCharting	Chartevents	0	375	Median
Respiratory Rate	numericitems	NurseCharting	Chartevents	0	100	Median
Oxygen Saturation	numericitems	Lab	Chartevents | Labevents	0	100	Median
PH	numericitems	Lab	Chartevents | Labevents	3	9.5	Median
Temperature (C)	numericitems	NurseCharting	Chartevents	10	45	Median

Applying noise filters is a well-known preprocessing technique for finding and removing the dataset's noisy instances. A simple filter approach is a variable-by-variable data cleaning. In this approach, the values considered 'suspicious' values are discarded or corrected according to specific criteria. In this approach, the criteria include but are not limited to the following: an expert evaluates the suspicious data as errors or as falsely labelled or, in other cases, a classifier predicts those values as 'unclear' data. Other statistical methods could be used by measuring the distributive properties of the data and applying

some statistical tests, i.e., using the mean and variance or excluding values past a certain percentile. Other methods may include distance or density-based methods to omit outliers/noise.

In all datasets, the missing values of each predictor variable were filled using the median and mode for continuous and categorical variables after excluding unqualified patient records. Min and max values were selected with clinical insight to remove outliers and a level of noise. If a value was outside the min and max range, this value was removed. This process was applied before any transformation of the time series. Multiple imputations by using chained equations (MICE) were considered and would be a more suitable sophisticated method to utilise if the objective was predictive performance[118]. As this was not the case and due to the high computational costs associated with this method, this method was deemed too computationally expensive in conjunction with the nested k-fold cross-validation implemented.

3.3.6 Heterogeneity

Healthcare data represents a remarkably heterogeneous set of data types and sources drawn from multiple sources and encompassing multiple modalities. Perhaps the most prominent examples can be found within the wealth of clinical data now digitalised due to EHR integration across healthcare practices. However, it is essential to note that heterogeneity can exist even within data of the same type, for example, comparing ICD-9 coding diagnosis with nurse-documented notes [119]–[122]. From a technical perspective, one of the primary considerations in applying ML methodologies to the increasingly heterogeneous healthcare data space comes with the acknowledgement that the data captured across each source may span a range of data types (demographic, ECGs, MRIs, amongst others). Additionally, data sources may collect unique data making it more challenging to recreate external validation, shown in Figure 9, where the ICU information type is displayed for each data source. The ICU-type information is collected for all three databases; however, the granularity between them is distinct, especially with the AUMC database.

Figure 9: ICU-level demographic data displays the ICU categories from each of the ICU databases.

Furthermore, the homogeneity of the data sources presents a challenge for learning algorithms. However, the modelling of homogeneous integrated data sources presents several unique concerns, including the ability to reconcile 'dirty' data such as incompatible test results, changes in coded data and the need to ensure trust between the system that share sensitive data [123]. Additionally, the missing data itself can present concern. The impact of this fragmentation compounds during the integration of multiple records between a primary and specialist, or multiple instances of the same record, has severe implications for the ML algorithms used to analyse the data, therefore, must be handled appropriately. Finally, the siloed data source presents a deeper systemic issue. The lack of unique identifiers to track an individual through the healthcare system's various components often presents an incomplete view of an individual health data [124]. This provides duplicated data that can bias the underlying distributions at a basic level. While at a higher level, this removes a valid independence assumption, potentially biasing performance measures by splitting what appears to be a unique instance amongst the train and test sets during evaluation leading to biased results.

3.3.7 Data Creation

The goal is to create the same representative dataset from all databases to reduce bias and ensure a fair comparison. To complete this task, we focused on frequently available biomarkers across the data sources, and to reduce complexity, a small number of clinical features were considered. In this study, all data sources followed the same data extraction. First, all ICU records of patients greater than 18 years old were extracted. Next, all ICU admission with LOS missing or outliers, defined as a LOS-ICU above the 99th percentile of the LOS-ICU in the studied datasets, were omitted. Finally, we randomly selected one record for the corresponding patient for patients with twice or more ICU admissions during one hospitalisation to ensure that all observations were independent in the model development. Compared with the approach of selecting the first admission records for a patient having multiple ICU admissions during a hospitalisation, randomly selecting one ICU record for the patient may help include patients with varying severities [125]. The flowchart of the process for patient inclusion can be seen in Figure 10. We used MySQL V8.0.23, a relational database management system that uses SQL to manage, store and modify the databases for the AUMC, eICU and MIMIC-III databases throughout the analysis.

3.4 Modelling

3.4.1 Primary Outcome

There are three primary outcomes for this analysis, first is to compare and explore the three ICU data sources. Secondly, outline the ML steps and methods when constructing a dataset for medical data. Lastly, to model outcomes of ICU mortality and ICU LOS. ICU mortality was coded as a binary variable to indicate whether the patients died in ICU, dead (1) or alive (0). ICU LOS was coded in hours representing the duration of the ICU admission. Each data source is modelled independently using a nested k-fold cross-validation approach. Furthermore, we pooled together all ICU data sources and applied the same modelling approach for supplementary comparison. However, due to the different rates of missingness and value population frequency amongst the different data sources and variables, in addition to varying prevalence rates, we did not directly validate any of the models using another data source (i.e., MIMIC-III models externally validated by eICU data). We acknowledge that death is a correlated outcome of LOS as, in some cases, patients may die during their hospital stay, leading to shorter ICU LOS. Therefore, we considered both outcomes independently.

3.4.2 Univariable Analysis

We choose to perform univariable analyses on the individual variables in the dataset to understand the distribution of the data values appropriately. The most common way to perform a univariate analysis is to describe a variable using summary statistics. We used nonparametric statistical tests for continuous and categorical variables and for the univariate analysis of the three databases. The univariate analysis compares variable distributions for significant differences amongst the database sources. The Kruskal-Wallis test was applied to assess the differences among the database groups for all continuous variables. Similarly, Pearson's Chi-Square was used to assess differences for all categorical variables. P-values < 0.05 were considered statistically significant. Descriptive data are presented, continuous variables are represented as the mean and standard deviation, the median with first and third quartiles, min and max values, and the level of missingness displayed in that given feature in the form of counts. Categorical features were presented as counts and percentages.

3.4.3 Multivariate Analysis

We compare four commonly deployed ML methods, logistic and linear regression, with forward stepwise feature selection (LR) depending on the outcome task (classification/regression), in addition to a random forest (RF) and gradient-boosted machines (GBM). Logistic regression models the outcome probability or risk to be '1' (positive class) as $P(Y = 1) = 1/\left(1 + \exp\left[-\sum_{k=0}^{K} \beta_k X_k\right]\right)$, where $\{\beta_0, \dots, \beta_K\}$ are the model coefficients, which are estimated by maximum likelihood [126]. The logistic regression coefficients are the logarithm odds ratios (LogORs) between the factors and the outcome,

which makes them useful for explanatory analysis. Thus, if a factor increased by one unit, its LogOR measures how much the outcome odd would increase or decrease, depending on whether the coefficient is positive or negative, respectively. Multiple linear regression models the outcome Y (continuous value) to be $\sum_{k=0}^{K} \beta_k X_k$ where $\{\beta_0, \dots, \beta_K\}$ are the model coefficients that estimate each factor's impact similarly to logistic regression. Thus, if a factor increased by one unit, the outcome Y measures how much the outcome would increase or decrease, again depending on whether the coefficient is positive or negative. RF repeatedly fits induction trees to several subsets of random samples with replacements extracted from the training set. RF predicts a new outcome in classification tasks by taking the majority vote[127]. In a regression task, RF predicts a new outcome by averaging across all of the predicted Y values. GBM also uses many decision trees to make predictions, although any ML algorithm could be utilised. Contrasting with RF, GBM implements an iterative learning algorithm, such that a new tree model is fitted on the cases where the previous tree performed inadequately [128]. RF and GBM can also be used for explanatory analysis as they rank the input variables based on their relevance to the model predictions (variable importance).

3.4.4 Model Validation, Performance and Explainability

We implemented a nested K fold cross-validation, which consisted of 5 outer folds and 3 inner folds, for hyperparameter tuning. The input variables were automatically selected using a sequential forward search algorithm over 3 dataset instances for each iteration. For logistic and linear regression (LR), input variables were automatically selected using a sequential forward search algorithm over 5 dataset instances. An inner cycle of 3-fold cross-validation was used for each iteration to select relevant variables. The selection algorithm starts with a baseline model (i.e., all coefficients but the intercept are set to zero, $\beta_{k\neq 0} = 0$), and in each step, the variable which most improves the performance on the validation set is added [129]. Several RF and GBM hyperparameters were tuned using the same validation splits as for LR. For the RF models, we tested a range of variables randomly sampled between 3 and 14, and a range of minimum node size (which controls the depth of the trees) between 3 and 14. For GBM models, a range of shrinkage values (which controls the impact of each additional fitted tree) from 10^{-5} to 10^{1} and a range of minimum number of observations in a node from 3 to 14 was tested. The RF and GBM variable importance of the models with the best hyperparameter set for each cross-validation cycle were also estimated.

The prediction performance of the models was measured using AUC for ICU mortality and a root mean square error (RSME) for ICU LOS. A high AUC value represents a strong discriminant power between the classes. The lower RMSE value represents a stronger forecasting ability for the model to predict the ICU time of a patient. The RSME represents the standard error from the actual value, in our case, in hours.

3.5 Results

3.5.1 Database Cohorts

Overall, for the comparison of databases, a total of 154,070 ICU admissions were selected, 111,797, 32,714 and 9,559 from the eICU, MIMIC-III and AUMC respectfully, as displayed in Figure 10.

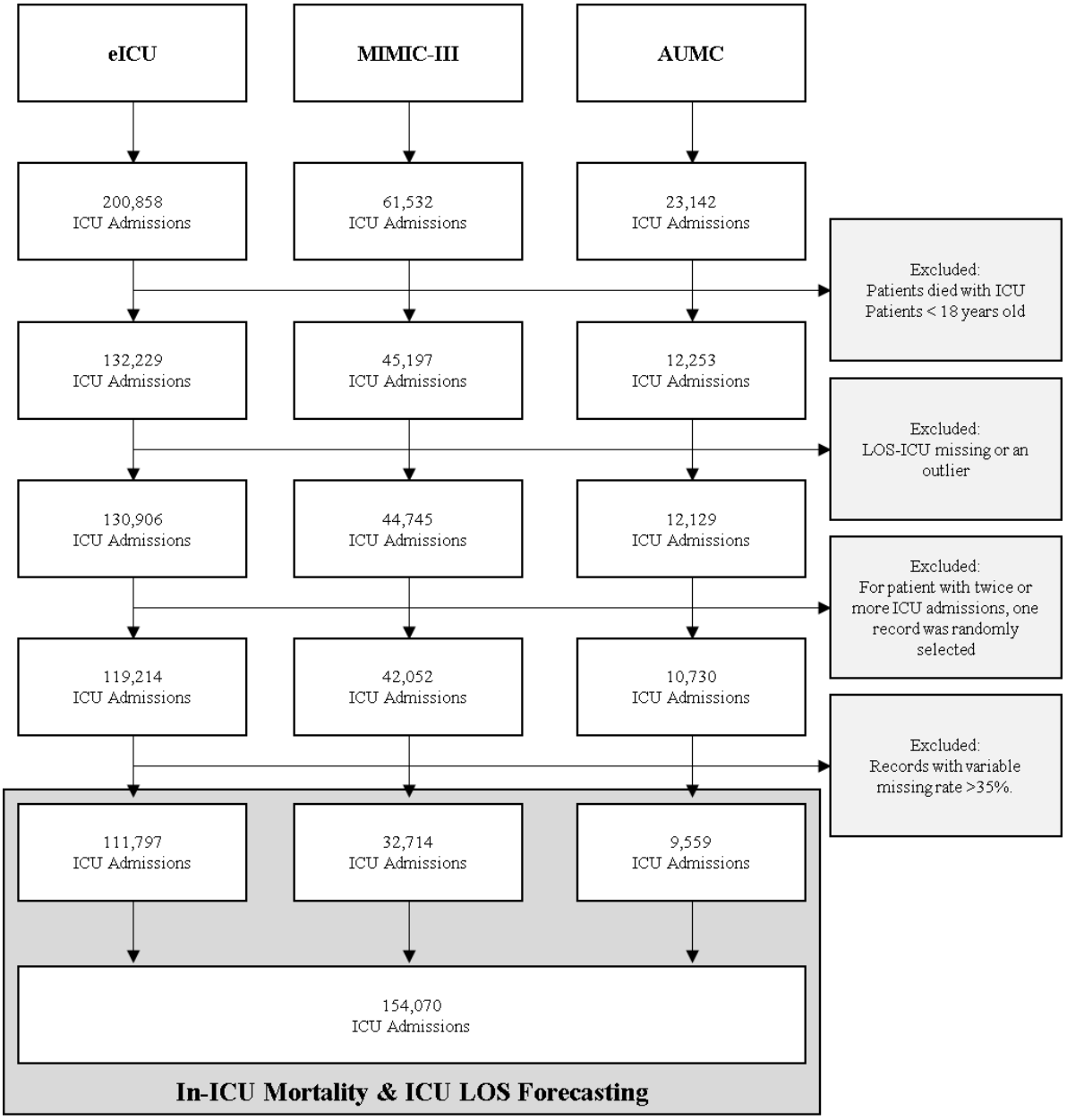


Figure 10: The procedure of the cohort selection for each data source and final cohort count.

The characteristics of the ICU patients were similar in all databases, as displayed in Table 6. The proportion of ICU patients among the data sources, which share the same clinical features, varied, with 72.6% of all ICU cases coming from the eICU. The Proportion of ICU patients who expired in the ICU is significantly different among the databases, with the AUMC displaying high mortality rates, with 14% of ICU admissions resulting in mortality, compared to the MIMIC-III dataset at 7.8% and the eICU at 5.3% which were significantly reduced. A similar result was displayed in terms of time spent in the ICU, where patients from the AUMC dataset had a significantly longer duration of stays compared to the eICU and MIMIC-III, where on average, patients from the AUMC cohort spent twice as long as patients from the eICU dataset, with a similar result shown with the MIMIC-III. Distinct differences

were displayed in both categorical features, gender and age. The AUMC database showed a significant increase in male patients admitted to the ICU compared to the other data sources. A total of 63.6% of AUMC ICU admission were male compared to eICU and MIMIC-III, which displayed 54% and 56%, respectively.

Table 6: Summary characteristics and statistical comparisons of ICU patients in the AUMC, eICU and MIMIC-III.

Features	**AUMC (N=9559)**	**eICU (N=111797)**	**MIMIC (N=32714)**	**Total (N=154070)**	**P-value**
Hospital Discharge Statue	Unknown*	9704 (8.7%)	3523 (10.8%)	13227 (9.2%)	0.001*
ICU Discharge Statue	1369 (14.4%)	5921 (5.3%)	2547 (7.8%)	9837 (6.4%)	0.001*
ICU LOS					0.001*
Mean (SD)	165.0 (218.6)	82.2 (75.5)	101.2 (113.3)	91.4 (101.2)	
Median (Q1, Q3)	72.0 (41.0, 187.0)	54.2 (37.8, 94.5)	59.2 (38.8, 111.0)	56.0 (38.1, 99.1)	
Min - Max	25.0 - 1435.0	24.0 - 532.1	24.0 - 790.1	24.0 - 1435.0	
Missing	0	0	0	0	
Gender	5946 (63.6%)	60352 (54.0%)	18421 (56.3%)	84719 (55.1%)	0.001*
Heart Rate					0.001*
Mean (SD)	82.6 (17.3)	85.2 (16.4)	85.0 (15.8)	85.0 (16.4)	
Median (Q1, Q3)	81.1 (70.5, 93.3)	84.0 (73.3, 95.9)	83.9 (73.7, 95.2)	83.8 (73.2, 95.6)	
Min - Max	35.1 - 168.4	0.0 - 219.7	31.3 - 309.4	0.0 - 309.4	
Missing	1	244	4	249	
BP Systolic					0.001*
Mean (SD)	121.4 (18.5)	121.8 (18.5)	119.3 (17.1)	121.3 (18.2)	
Median (Q1, Q3)	119.0 (108.5, 132.0)	119.6 (108.1, 133.7)	116.9 (107.0, 129.9)	119.0 (107.8, 132.8)	
Min - Max	49.7 - 226.8	39.1 - 254.0	28.0 - 210.0	28.0 - 254.0	
Missing	2	1065	10	1077	
BP Diastolic					0.001*
Mean (SD)	61.2 (9.6)	64.9 (11.9)	61.3 (12.3)	63.9 (12.0)	
Median (Q1, Q3)	60.3 (54.9, 66.7)	63.9 (56.6, 72.2)	60.1 (53.5, 67.6)	62.8 (55.8, 70.9)	
Min - Max	15.4 - 121.0	0.0 - 288.7	13.7 - 361.4	0.0 - 361.4	
Missing	0	1057	20	1077	
Oxygen Saturation					0.001*
Mean (SD)	83.2 (11.5)	96.8 (2.4)	97.3 (2.2)	96.0 (5.1)	
Median (Q1, Q3)	83.7 (75.8, 93.5)	97.0 (95.6, 98.5)	97.6 (96.2, 98.8)	97.0 (95.4, 98.5)	
Min - Max	0.9 - 100.0	0.0 - 100.0	16.0 - 100.0	0.0 - 100.0	
Missing	2	11134	26	11162	
Respiratory Rate					0.001*
Mean (SD)	17.8 (6.7)	19.5 (4.5)	18.8 (4.1)	19.3 (4.6)	
Median (Q1, Q3)	17.0 (14.0, 21.0)	18.8 (16.5, 21.8)	18.2 (16.0, 21.0)	18.6 (16.2, 21.6)	
Min - Max	0.0 - 96.0	0.0 - 100.0	0.0 - 98.4	0.0 - 100.0	
Missing	1097	4027	76	5200	
PH					0.001*
Mean (SD)	7.4 (0.1)	7.4 (0.1)	7.0 (0.7)	7.2 (0.4)	
Median (Q1, Q3)	7.4 (7.3, 7.4)	7.4 (7.3, 7.4)	7.3 (7.0, 7.4)	7.4 (7.3, 7.4)	
Min - Max	5.7 - 9.3	3.4 - 7.8	3.0 - 9.0	3.0 - 9.3	
Missing	199	75294	8999	84492	
Temperature (C)					0.001*

Mean (SD)	36.2 (1.5)	36.8 (0.6)	36.8 (0.9)	36.8 (0.8)
Median (Q1, Q3)	36.6 (36.0, 37.0)	36.8 (36.6, 37.1)	36.8 (36.4, 37.2)	36.8 (36.5, 37.1)
Min - Max	12.6 - 44.4	16.5 - 44.9	10.1 - 45.0	10.1 - 45.0
Missing	81	1118	527	1726
Age Group				0.001*
18-39	1107 (11.6%)	10501 (9.4%)	3280 (10.0%)	14888 (9.7%)
40-49	915 (9.6%)	9994 (8.9%)	3487 (10.7%)	14396 (9.3%)
50-59	1569 (16.4%)	20273 (18.1%)	5724 (17.5%)	27566 (17.9%)
60-69	2318 (24.2%)	25448 (22.8%)	6656 (20.3%)	34422 (22.3%)
70-79	2522 (26.4%)	24586 (22.0%)	6512 (19.9%)	33620 (21.8%)
80+	1128 (11.8%)	20995 (18.8%)	7055 (21.6%)	29178 (18.9%)

3.5.2 Evaluation of Model Performance

Table 7: Prediction performance of the three models on the AUMC, eICU, and MIMIC-III, and the stacked dataset, which comprises all three databases.

Databases	**LR**	**RF**	**GBM**
	ICU Mortality - AUC (SE)		
Stacked Databases	62.48 (0.001)	65.00 (0.002)	64.63 (0.002)
AUMC	76.58 (0.006)	79.50 (0.006)	79.13 (0.005)
eICU	56.39 (0.002)	55.74 (0.004)	56.73 (0.002)
MIMIC-III	70.18 (0.002)	79.15 (0.002)	77.94 (0.002)
	ICU LOS (Hours) - RMSE (SE)		
Stacked Databases	1.709944 (0.005)	1.681901 (0.005)	1.690121(0.005)
AUMC	2.669321 (0.014)	2.459770 (0.012)	2.507542 (0.014)
eICU	1.591408 (0.004)	1.592031 (0.004)	1.589512 (0.004)
MIMIC-III	1.827498 (0.006)	1.710737 (0.005)	1.730920 (0.006)

The prediction of the three models for each of the database partitions is compared in Table 7. The AUMC dataset achieved the best overall performance for In-ICU Mortality (AUC, 79.50%), followed by MIMIC-III (AUC, 79.15%), then the stacked dataset (AUC, 65.00%), with the eICU resulting in the lowest performance accuracy (AUC, 56.73%). The eICU, unlike previously, was ranked highest amongst predictive ICU LOS performance (RMSE, 1.58 Hours), followed by the stacked dataset (RMSE, 1.68 Hours), then MIMC-III (RMSE, 1.71 Hours), and lastly, the AUMC dataset (RMSE, 2.46 Hours). The predictive performance of RF was slightly superior compared to LR or GBM for both modelling outcomes in nearly all cases.

3.5.3 Variable Importance

Table 8: Ranked variable importance identified by RF for ICU LOS.

Rank	**AUMC**	**eICU**	**MIMIC-III**	**Stacked**
1	PH	Heart Rate	PH	Oxygen Saturation
2	Oxygen Saturation	BP Diastolic	Temperature C	Temperature C
3	Heart Rate	BP Systolic	Heart Rate	Heart Rate
4	Temperature C	Respiratory Rate	Respiratory Rate	Respiratory Rate
5	Respiratory Rate	Temperature C	Oxygen Saturation	BP Systolic
6	BP Systolic	Oxygen Saturation	BP Systolic	BP Diastolic

7	BP Diastolic	PH	BP Diastolic	PH
8	Gender	Gender	Gender	Gender
9	Age Group 70 79	Age Group 50 59	Age Group 80+	Age Group 50 59
10	Age Group 60 69	Age Group 60 69	Age Group 70 79	Age Group 60 69
11	Age Group 18 39	Age Group 18 39	Age Group 60 69	Age Group 70 79
12	Age Group 50 59	Age Group 70 79	Age Group 50 59	Age Group 80+
13	Age Group 80+	Age Group 40 49	Age Group 40 49	Age Group 40 49
14	Age Group 40 49	Age Group 80+	Age Group 18_39	Age Group 18 39

The predictive variable importance identified by the RF models is listed in Table 8. All models ranked five variables, namely heart rate, PH, oxygen saturation, respiratory rate and temperature, among the top important variables. All RF models implemented ranked the categorical variables as least important compared to continuous variables, with gender displaying more importance than age in all cases. Uniquely, all ranks of the variable important are different for all databases. The top-ranked most important feature for each data source is likewise independent. AUMC ranked PH as the most important, similarly to MIMIC-III, however, eICU and the stacked datasets differ, which ranked heart rate and oxygen saturation as the most impactful features when modelling.

3.6 Discussion

This chapter has illustrated several processes from which data challenges arise within the healthcare domain. In addition, we have touched upon various preprocessing issues and obstacles that must be considered. To this point, we have discussed the intrinsic characteristics of data, those properties which influence the ML algorithmic performance. It is imperative to understand the process from which preprocessing challenges arise within the healthcare domain but also the implication of the preprocessing phase on the ML pipeline. While a significant effort has been undertaken to develop models able to process the volume of data obtained during the analysis of millions of digitalised patient records, it is essential to remember that volume represents only one aspect of the data. In reality, generating data from an increasingly diverse set of sources presents an incredibly complex set of attributes that must be accounted for throughout the ML pipeline. This chapter highlights such challenges and is broken down into distinct components, each representing a distinct pipeline phase from data collection to model evaluation. Unfortunately, the ability to derive accurate and informative insight requires more than the ability to execute ML models. Instead, a deeper understanding of the data on which the models are run is imperative for their success.

We implemented four ML methods, namely logistic and linear regression (LR), RF and GBM, to construct LOS and mortality prediction models based on each database. Furthermore, we combined all the data sources into a stacked, cohesive dataset to test generability and predictive performance across all data sources. One of the contributions of this study was to explore and compare each critical care database for the discussed outcomes. The four ML-based models were used to gather a baseline performance and assess the heterogeneity in the different data sources concerning predictive performance. The overall prediction performance of all four models for the eICU was less superior in predicting ICU mortality compared to MIMIC-III and AUMC. In contrast, the eICU datasets showed the most outstanding overall predictive performance for forecasting ICU LOS compared to the other data sources. These results suggest that a possible degradation of prediction performance occurs depending on the data source. This could be due to the granularity of which the data is collected or related to the single and multi-centric nature of the data sources.

ML applications in healthcare have increased emphasis on ML explainability[130]. For most ML systems, improved predictive accuracy may often be achieved through increased model complexity [131]. Suppose a clinical application is taken into consideration. In that case, a pragmatic prediction model could help physicians identify patients at high risk and this may provide timely individualised

interventions, and finally, patients' prognoses may be improved. Therefore, from an application perspective, the prediction models developed in this study are an innervational tool though it has limited contribution from the perspective of ML methods. Although many prediction models have been developed in the literature using these data sources, none has compared the predictive performance among these three ICU databases using a controlled environment, as each model and dataset followed the same preprocessing and used the same biomarkers/features available. May studies focus on one source or compare the performance between two data sources, most commonly the eICU and MIMIC-III, both US-based data[132]–[134]. This study showed that different open-source databases used in critical care could deliver different performances concerning outcomes of interest with the same set of feature variables. This is further displayed in the feature importance calculated, as each data source delivered a unique feature importance ranking. The top important variables for ICU LOS ranked by RF were similar to those retained by GBM. However, the ranking between the data sources differs slightly, which we would not expect if all data sources represented the same patient population. This finding may provide a clue for future research, meaning that studies from a single ICU data source could potentially be insufficient for true medical insight.

This study has several strengths. First, the databases used to derive the prediction models differ geographically, as two data sources are single-centre databases, in addition to an extensive multicentre database with a relatively wide representative population. Secondly, the AUMC is a European database. We compare the predictive performance of the same models with two US data sources, allowing us insight into the feasibility of international scoring systems based on a single population. Thirdly, all the predictor variables used to construct prediction models are routinely collected during the first 24 hours in ICU, thereby ensuring the feasibility of applying the prediction models in clinical practice to assist physicians in decision-making. Lastly, we explored and compared the databases by looking at data availability and comparing the data available for modelling, comparing basic statistical and demographic features in both data sources.

However, this study has some limitations. First, the databases used to derive the prediction models come from different time periods. The eICU database only contains data on ICU patients admitted between 2014 and 2015 across 290 + medical practices from the US, the MIMC-III database only contains data on patients who were admitted between 2001 to 2012 to a single centre in Boston, Massachusetts, and the AUMC contains patients who were admitted between 2003 to 2016. All data sources are collected at different regions in time and over varying time durations. Therefore, a truly fair comparison may not have been achieved. We acknowledge that propensity score testing may have been used to combat heterogeneity in the data and deliver a stricter comparison. Therefore, the clinical utility of the prediction models needs further assessment before application in other regions. Secondly, selection bias may exist since we excluded patients who died in the ICU before the 24-hour period, as it is a competing outcome regarding LOS. Accordingly, the models developed in this study may not apply to patients who die in the ICU within the first 24 hours of admission. Thirdly, to compare the prediction performance of four ML-based models in an objective manner, we only included a small range of prediction variables for training the models. Other potential predictor variables may have been neglected in our study, as the aim is not to build the most optimal ML model but to compare the performance of the developed models using the same ML pipeline. Although, adequate predictive performance was achieved with several models using few features, achieving AUC performances of 79.5% for ICU mortality prediction and a prediction error of approximately 1.3 hours for ICU LOS.

3.7 Conclusion

This work has served as a foundation highlighting a broad set of considerations that must be addressed as ML establishes its place in the healthcare industry. In summary, this study demonstrates that the different ICU databases provide different clinical attributes regarding data availability, predictive performance, and modelling capabilities. This study lays the foundation for future applications of ICU analysis based on clinical prediction models in addition to various other outcomes and sample populations. We present a framework to model different data sources using routinely available clinical

data from large publicly available databases. We demonstrated that factors of importance show homogeneity depending on the data source. Similarly, we demonstrated homogeneity regarding predictive performance depending on the data source.

4 CHAPTER 4: Medical Records Analysis: A Sepsis Study

In this chapter, we cover three different experiments with the same cohort. Firstly, from a clinical perspective, we analysed sepsis and its subtypes using traditional statistical methods, resulting in a journal publication[135]. Next, we extended the first analysis by implementing a range of ML models to four outcomes: In-hospital mortality, In-ICU mortality, hospital-LOS, and ICU LOS. The aim was to test various ML models for predictive performance and explainability. The last experiments focus on model optimisation and predictive performance by implementing a multi-task learning (MTL) framework. The layout of this chapter follows an introductory setting, defining the exploratory setting and how the dataset was derived.

For each of the studies undertaken, the primary outcome for classification tasks was coded as a binary variable to indicate whether the patient was dead ('1') or alive ('0') for mortality classification. Regression tasks such as LOS prediction were coded as a continuous variable representing the duration in hours. Model performance was measured using the area under the receiver operator characteristic (AUC) curve for classification and mean squared error (MSE) for regression. AUC and MSE means, and confidence intervals (CI) were calculated for each sepsis type using the outer folds of the nested k-fold cross-validation method implemented. Bootstrapping was not considered as the method to provide parameter estimates due to the computational costs.

4.1 Introduction

Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection [136]. It is not a uniform disease, but a complex syndrome of physiologic and biochemical abnormalities. Clinical experience supports the concept that prognosis, treatment, severity and time course vary depending on the source of infection [137], [138]. Consequently, attempts have been made to characterise different types of sepsis based on clinical data, routine blood results and biomarkers [139]. Mortality of sepsis ranges from 15% in patients with sepsis without shock to 56% in patients with sepsis with shock [140]. However, mortality prediction for sepsis remains satisfactory at best [141].

Although numerous trials have been designed to explore treatment options for sepsis, so far, none of these has resulted in new therapies [142]. A major shortcoming of many of these multi-centres randomised clinical trials is the patient cohort investigated. Patients with sepsis manifest striking heterogeneity, not only with respect to the site or microbiology of the inciting infection but also with respect to the comorbid conditions present in the patient at the time of onset [143]. Comorbidities, site of infection and pathogen factors impact the mortality attributed to sepsis. However, in most clinical trials differentiation between groups of sepsis is lacking and may have contributed to the negative outcome of these studies. Recently, attempts have been made to discriminate sub-phenotypes of sepsis based on panels of immunological markers. Although promising, these clinical phenotypes for sepsis [141] are complex, rely on the measurement of biomarker profiles, and are thus not easy to implement into routine clinical applications.

Electronic health records are now commonly used to record all routine clinical data. This allows the construction of large databases, which not only structure and aggregate clinical data but also record outcome measures such as mortality, length of stay and duration of ventilation. Alongside routinely applied scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE), the Sequential Organ Failure Assessment (SOFA) or the Simplified Acute Physiology Score (SAPS), novel outcome prediction models are being developed based on these large patient populations.

In this research, we investigate in-hospital mortality and predictors thereof in different cohorts of sepsis based on the origin of infection using data from the eICU Collaborative Research Database, a freely available multi-centre database for critical care research [144]. We hypothesise that mortality and factors influencing mortality risk differ between pulmonary, urinary, and abdominal sepsis as the three

most relevant clinical presentations. We aim to identify unifying and distinct features in these groups. Comparisons will be made with established outcome prediction scores such as APACHE IV and SOFA to determine if more sophisticated models show superior performance in predicting hospital mortality in these different groups of septic patients. All Acuity scores that measure patient's status used throughout the research are defined in the supplementary materials.

4.2 Data Source & Extraction

In this study, we used the eICU database (eICU)[145]. We extracted data from the medical ICUs (MICU), surgical ICUs (SICU), and medical-surgical ICUs (Med-Surg ICU). Specialist critical care units such as cardiothoracic and cardio-surgical ICUs were excluded because of their specific patient cohorts with distinct presentations of sepsis. Patients after elective surgery and those with an underlying haematology diagnosis were also excluded, as their clinical presentation and course are distinct from patients with sepsis as the primary diagnosis. We then used the admission diagnosis codes, which are coded using the APACHE IV diagnosis system, to extract the admissions related to sepsis, and excluded patients <18 years of age and with an ICU stay < 72h. Lastly, all records with more than 35% missing data were excluded. These inclusion and exclusion criteria are represented in Figure 11.

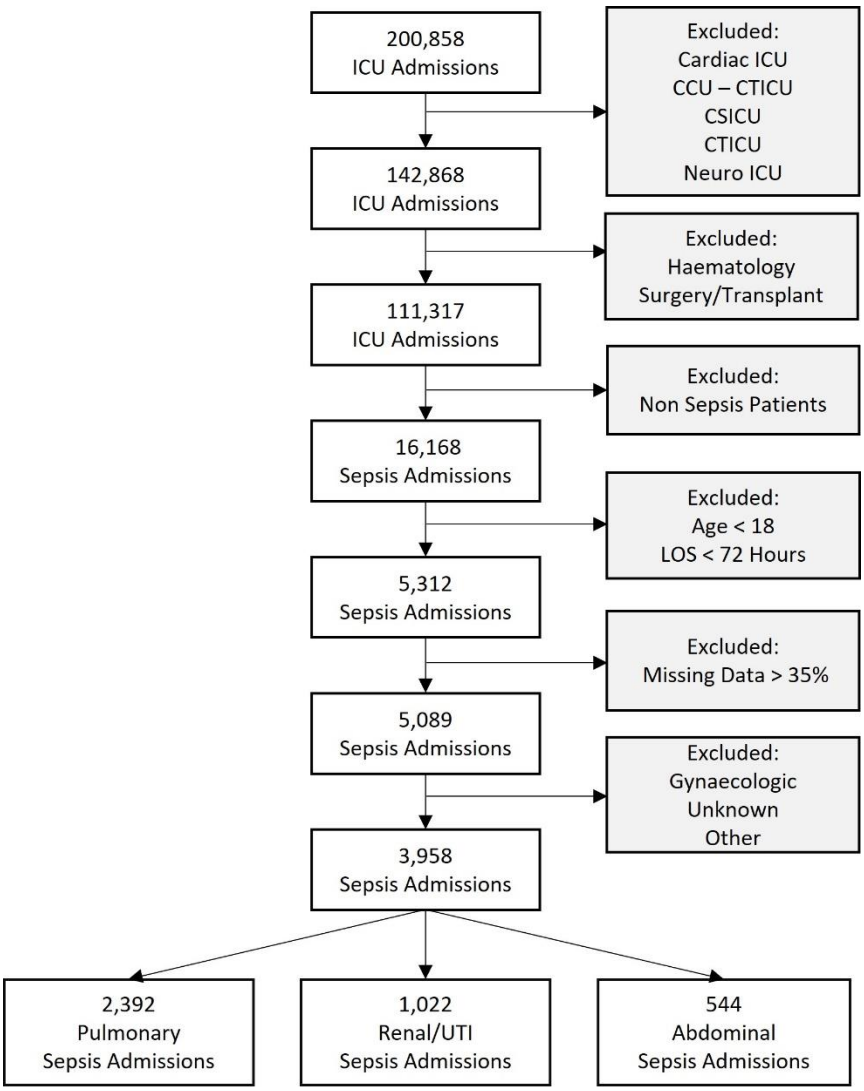


Figure 11: Flowchart of sepsis cohorts analysed showing the inclusion and exclusion criteria. ICU: intensive care unit, CCU-CTICU: critical care unit-cardiothoracic intensive care unit, CSICU: cardio-surgical intensive care unit, LOS: length of stay, UTI: urinary tract infection.

We collected all electronic health record data from the acute phase of the ICU admission, defined as the first 72 hours after admission. From this dataset, we excluded the first 6 hours (resuscitation phase), where the priority is to stabilise the patient. Previous studies have used data from different time windows for outcome prediction, e.g., the first 24h of the ICU admission [146]. All dynamic features were organised into 1-hour non-overlapping time series bins when extracting the data from the eICU database. This was to accommodate for different sampling frequencies of available data and the balance between missing data points and bin size. All time-varying variables were converted into tabular representations by extracting their means and standard deviations. The mean value of these time-varying variables, which represents the average of each time series, was named "Average" (Avg), e.g., the mean of the heart rate signal was coded as "Avg Heart Rate". Similarly, the standard deviation, which is representing the variation in the time series, was coded as "Variations" (Var), e.g., Heart Rate Var.

4.3 Definition of Sepsis Types

A cohort of patients with sepsis was extracted based on the ICU admission diagnosis, which is coded using the APACHE IV diagnosis system [147] and routinely recorded in the eICU database. From here, the following septic groups were identified: pulmonary, abdominal, and renal/ urinary tract infection (UTI). Other smaller cohorts of septic patient groups were excluded either because of a lack of clarity regarding their clinical source (e.g., those encoded as "unknown" or "others") or because of their considerably smaller number of cases (e.g., gynaecologic sepsis with less than 20 admissions). The prevalence in these groups was also reviewed against the encoded ICD codes for these patients to ensure that the relevant cohorts were well-defined.

4.4 Univariable Analysis

We perform univariable analyses on the individual variables in the dataset to understand the distribution of the data values appropriately. The most common way to perform a univariate analysis is to describe a variable using summary statistics. We used nonparametric statistical tests for continuous and categorical variables for univariate analysis of the three main groups of sepsis. The univariate analysis aims to compare variable distributions for significant differences amongst the sepsis groups. The Kruskal-Wallis test was applied to assess the differences among the sepsis groups for all continuous variables. Similarly, Pearson's Chi-Square was used to assess differences for all categorical variables. P-values < 0.05 were considered statistically significant. However, comparing characteristics using p-values can be problematic in large databases of administrative data, especially when dealing with large samples[148]. One issue with using p-values in large samples is that any slight difference between groups can result in statistical significance, even if the difference is not practically meaningful [149]. Therefore, when comparing characteristics in large datasets, it is crucial to not only rely on statistical significance but also consider effect sizes and practical significance[150].

Table 9: Demographics, comorbidities, vital signs, and routine prognostic scores used for modelling. The first column displays the data characteristics (variables). Columns second to fourth show summary statistics of all the variables for each sepsis group. Sepsis group cohort sizes are reported under the group name. Numeric variables are reported with the median and IQR (in parentheses), while categorical variables are reported with the frequency and proportion (in parenthesis). The resulting statistical tests are reported in the fifth column in the form of p-values. Any p-value smaller than 0.001 was indicated as "<0.001".

	Abdominal (N=544)	**Pulmonary** (N=2392)	**Renal/UTI** (N=1022)	**P-Value**
Outcome				
In-hospital Mortality	103 (18.9%)	461 (19.3%)	131 (12.8%)	< 0.001
Demographics				
Age	67.0 (56.0, 76.0)	67.0 (56.0, 77.0)	71.0 (60.0, 81.0)	< 0.001
Gender (Male)	276 (50.7%)	1281 (53.6%)	437 (42.8%)	< 0.001
Comorbidities				

Myocardial Infarction	45 (8.3%)	184 (7.7%)	85 (8.3%)	0.7862
CHF	85 (15.6%)	461 (19.3%)	204 (20.0%)	0.0932
PVD	27 (5.0%)	116 (4.8%)	53 (5.2%)	0.9172
Dementia	14 (2.6%)	166 (6.9%)	104 (10.2%)	< 0.001
COPD	81 (14.9%)	600 (25.1%)	136 (13.3%)	< 0.001
CTD	16 (2.9%)	70 (2.9%)	35 (3.4%)	0.7302
Peptic Ulcer Disease	14 (2.6%)	75 (3.1%)	35 (3.4%)	0.6552
Mild Liver Disease	31 (5.7%)	55 (2.3%)	26 (2.5%)	< 0.001
Uncomplicated DM	146 (26.8%)	713 (29.8%)	407 (39.8%)	< 0.001
Renal Disease	94 (17.3%)	334 (14.0%)	165 (16.1%)	0.0712
Hemiplegia	45 (8.3%)	246 (10.3%)	146 (14.3%)	< 0.001
Severe Liver Disease	32 (5.9%)	49 (2.0%)	18 (1.8%)	< 0.001
Hypertension	269 (49.4%)	1143 (47.8%)	564 (55.2%)	< 0.001
Hypothyroidism	16 (2.9%)	100 (4.2%)	43 (4.2%)	0.3882
Atrial Fibrillation	70 (12.9%)	307 (12.8%)	144 (14.1%)	0.5962
Asthma	38 (7.0%)	219 (9.2%)	70 (6.8%)	0.0412
Seizures	32 (5.9%)	166 (6.9%)	83 (8.1%)	0.2312
Respiratory Failure	10 (1.8%)	126 (5.3%)	46 (4.5%)	0.0032
CABG	25 (4.6%)	139 (5.8%)	46 (4.5%)	0.2142
Cancer	116 (21.3%)	422 (17.6%)	169 (16.5%)	0.0572
Admission diagnosis				
Pulmonary	181 (33.3%)	2109 (88.2%)	350 (34.2%)	< 0.001
Cardiovascular	423 (77.8%)	1788 (74.7%)	787 (77.0%)	0.1852
Infectious Diseases	165 (30.3%)	569 (23.8%)	361 (35.3%)	< 0.001
Renal	205 (37.7%)	730 (30.5%)	662 (64.8%)	< 0.001
Gastrointestinal	323 (59.4%)	211 (8.8%)	91 (8.9%)	< 0.001
Oncology	20 (3.7%)	114 (4.8%)	24 (2.3%)	0.0042
Neurologic	85 (15.6%)	443 (18.5%)	270 (26.4%)	< 0.001
Endocrine	63 (11.6%)	330 (13.8%)	169 (16.5%)	0.0192
Vitals				
Avg Heart Rate	94.0 (81.9, 105.0)	90.2 (79.8, 100.8)	89.0 (78.1, 98.7)	< 0.001
Heart Rate Var	9.5 (6.9, 12.6)	10.0 (7.3, 13.5)	9.7 (7.1, 13.4)	0.0201
Avg SaO2	96.6 (95.3, 98.2)	96.6 (95.1, 98.0)	97.1 (95.9, 98.5)	< 0.001
SaO2 Var	1.9 (1.4, 2.5)	2.1 (1.6, 2.7)	1.8 (1.3, 2.5)	< 0.001
Avg GCS Total	13.8 (10.5, 14.9)	11.3 (9.0, 14.3)	13.6 (10.0, 14.8)	< 0.001
GCS Total Var	0.7 (0.2, 1.7)	0.9 (0.4, 1.9)	0.6 (0.3, 1.5)	< 0.001
Avg Respiratory Rate	20.5 (18.0, 23.9)	21.4 (18.6, 24.8)	20.1 (17.6, 23.3)	< 0.001
Respiratory Rate Var	3.8 (2.9, 5.0)	4.0 (2.9, 5.2)	3.7 (2.9, 4.9)	0.0171
Avg Temperature °C	36.8 (36.6, 37.2)	36.9 (36.6, 37.2)	36.8 (36.6, 37.2)	0.0431
Temperature °C Var	0.4 (0.3, 0.6)	0.4 (0.3, 0.6)	0.4 (0.3, 0.6)	0.0411
Avg MAP	76.8 (72.4, 84.3)	80.1 (74.4, 87.6)	78.6 (73.2, 86.4)	< 0.001
MAP Var	9.1 (7.3, 11.6)	9.6 (7.5, 12.1)	9.9 (7.9, 12.5)	< 0.001
Laboratory				
Avg Lymphs	7.0 (4.0, 10.0)	7.0 (4.3, 11.5)	8.0 (4.5, 12.1)	0.0101
Lymphs Var	2.1 (1.3, 3.8)	2.1 (1.0, 3.5)	2.1 (1.1, 3.5)	0.1171
Avg WBC	13.5 (9.3, 19.1)	12.2 (8.6, 16.9)	12.6 (8.7, 18.0)	< 0.001

WBC Var	2.5 (1.3, 4.5)	2.0 (1.1, 3.6)	2.3 (1.1, 4.2)	< 0.001
Avg Albumin	2.3 (1.9, 2.6)	2.3 (2.0, 2.7)	2.3 (2.0, 2.7)	0.0161
Albumin Var	0.2 (0.1, 0.3)	0.1 (0.1, 0.3)	0.1 (0.1, 0.2)	0.0041
Avg Platelets	163.8 (100.7, 239.8)	180.0 (124.0, 249.0)	164.6 (106.0, 230.8)	< 0.001
Platelets Var	21.2 (12.0, 37.1)	18.6 (9.2, 31.8)	17.7 (9.2, 29.8)	0.0071
Avg PaO2	92.4 (75.9, 115.4)	91.0 (75.8, 113.1)	97.0 (79.4, 120.0)	0.0181
PaO2 Var	20.6 (11.2, 43.4)	20.6 (11.1, 37.8)	19.3 (9.2, 34.3)	0.3321
Avg PaCO2	36.3 (31.6, 42.0)	39.3 (34.0, 46.3)	35.8 (30.2, 41.2)	< 0.001
PaCO2 Var	4.1 (2.6, 6.4)	4.0 (2.2, 7.1)	3.5 (2.1, 5.8)	0.0821
Avg FiO2	43.0 (35.0, 60.0)	50.0 (40.0, 70.0)	40.0 (33.3, 53.6)	< 0.001
FiO2 Var	7.5 (0.0, 17.9)	9.5 (3.5, 18.3)	7.1 (0.7, 15.2)	0.1121
Avg Total Bilirubin	0.9 (0.5, 2.3)	0.6 (0.4, 1.0)	0.6 (0.4, 1.2)	< 0.001
Total Bilirubin Var	0.2 (0.1, 0.5)	0.1 (0.1, 0.3)	0.1 (0.1, 0.3)	< 0.001
Avg Creatinine	1.4 (0.9, 2.6)	1.0 (0.7, 1.8)	1.4 (0.9, 2.3)	< 0.001
Creatinine Var	0.2 (0.1, 0.4)	0.1 (0.1, 0.3)	0.2 (0.1, 0.4)	< 0.001
Avg BUN	29.6 (17.7, 49.8)	25.5 (16.0, 41.0)	31.0 (18.3, 48.9)	< 0.001
BUN Var	4.8 (2.4, 8.8)	4.0 (2.1, 7.3)	4.2 (2.1, 8.6)	< 0.001
Avg PH	7.4 (7.3, 7.4)	7.4 (7.3, 7.4)	7.4 (7.3, 7.4)	< 0.001
pH Var	0.0 (0.0, 0.1)	0.0 (0.0, 0.1)	0.0 (0.0, 0.1)	0.0681
Avg Sodium	139.0 (136.0, 142.7)	139.8 (136.7, 143.0)	140.0 (136.9, 144.0)	< 0.001
Sodium Var	1.8 (1.2, 3.1)	1.9 (1.2, 2.8)	2.1 (1.3, 3.1)	0.0171
Avg Glucose	130.8 (110.0, 161.5)	141.0 (114.6, 170.6)	139.2 (115.8, 170.5)	< 0.001
Glucose Var	24.7 (15.3, 37.7)	26.8 (17.1, 41.4)	29.8 (19.5, 45.8)	< 0.001
Avg Hematocrit	28.9 (25.6, 32.8)	29.9 (26.5, 34.0)	29.5 (26.5, 33.3)	< 0.001
Hematocrit Var	2.1 (1.2, 3.1)	1.6 (0.9, 2.6)	1.5 (0.9, 2.5)	< 0.001
Avg Urine	161.1 (68.8, 364.1)	226.2 (96.3, 475.0)	224.2 (91.4, 551.0)	< 0.001
Urine Var	70.6 (33.3, 158.9)	108.9 (54.5, 208.9)	106.1 (50.3, 226.3)	< 0.001
Respiration				
Intubated	289 (53.1%)	1914 (80.0%)	486 (47.6%)	< 0.001
Drugs				
Norepinephrine	241 (44.3%)	861 (36.0%)	436 (42.7%)	< 0.001
Vasopressin	80 (14.7%)	225 (9.4%)	111 (10.9%)	0.0012
Phenylephrine	56 (10.3%)	147 (6.1%)	60 (5.9%)	0.0012
Dopamine	18 (3.3%)	60 (2.5%)	44 (4.3%)	0.0202
Epinephrine	15 (2.8%)	36 (1.5%)	15 (1.5%)	0.1022
Dobutamine	16 (2.9%)	43 (1.8%)	24 (2.3%)	0.1972
Scores				
Charlson CI	2.0 (0.0, 3.0)	2.0 (0.0, 3.0)	2.0 (1.0, 3.0)	0.4031
SOFA	4.0 (1.0, 7.0)	4.0 (2.0, 7.0)	3.0 (1.0, 6.0)	< 0.001
APACHE IV	73.0 (61.0, 88.0)	73.0 (58.0, 89.0)	73.0 (62.0, 87.0)	0.8951
SIRS	2.0 (1.0, 2.0)	2.0 (1.0, 2.0)	1.0 (1.0, 2.0)	< 0.001
qSOFA	1.0 (1.0, 2.0)	1.0 (1.0, 2.0)	1.0 (1.0, 2.0)	< 0.001
Unit Stay Type				< 0.001
Admit	453 (83.3%)	1991 (83.2%)	863 (84.4%)	

Other/Stepdown/Transfer	67 (12.3%)	277 (11.6%)	138 (13.5%)	
Readmit	24 (4.4%)	124 (5.2%)	21 (2.1%)	
Unit Type				< 0.001
Med-Surg ICU	386 (71.0%)	1830 (76.5%)	785 (76.8%)	
MICU	104 (19.1%)	440 (18.4%)	197 (19.3%)	
SICU	54 (9.9%)	122 (5.1%)	40 (3.9%)	
Admission Duration				
Hospital LOS	287.3 (190.7, 470.7)	264.4 (172.7, 400.2)	222.7 (159.6, 343.7)	< 0.001
ICU LOS	125.9 (92.1, 209.0)	140.7 (97.7, 228.8)	112.1 (87.5, 159.3)	< 0.001

Abbreviations: CHF: congestive heart failure; PVD: Peripheral vascular disease; COPD: Chronic obstructive pulmonary disease; CTD: Connective tissue diseases; DM: diabetes mellitus; CABG: Coronary artery bypass graft surgery; SaO2: oxygen saturation; GCS: Glasgow coma scale; MAP: Mean Arterial Pressure; WBC: white blood cells count; PaO2: partial pressure of oxygen; FiO2: Fraction of Inspired Oxygen; BUN: blood urea nitrogen; SOFA: Sequential Organ Failure Assessment; qSOFA: quick SOFA; APACHE: Acute Physiology And Chronic Health Evaluation; SIRS: Systemic Inflammatory Response Syndrome; ICU: Intensive Care Unit; Med-Surg ICU: medical-surgical ICU; MICU: medical ICU, SICU: surgical ICU; LOS: length of stay; Avg: average (mean); Var: variation (standard deviation).

4.5 Variable Selection and Cross-Validation

For LR, input variables were automatically selected using a sequential forward search algorithm over 10 dataset instances (10-fold cross-validation). For each iteration, an inner cycle of 5-fold cross-validation was used to select relevant variables. Collectively this is referred to as nested cross-validation (figure S1 in the supplementary materials). The selection algorithm starts with a baseline model (i.e., all coefficients but the intercept set to zero, $\beta_{k\neq 0} = 0$), and in each step, the variable which most improves the performance on the validation set is added [129]. Performance was defined for the forward search algorithm using the AUC metric and evaluated based on the inner 5-fold cross-validation cycles, with each fold used exactly once as the test data. The five results from the five folds then can be averaged to produce a single estimation of performance [151].

4.6 Clinical Relevance

In this Chapter, we address the knowledge gap of a) comparing different risk factors for each sepsis type and b) highlighting specific factors associated with mortality and LOS in the distinct sepsis groups, depending on the origin of the underlying infection. c) we compare sepsis-type features in a univariable analysis to investigate feature level differences. This research may help address the heterogeneity of the sepsis patient population, define discrete patient populations to guide the development of effective therapies and identify cohorts that benefit from specific interventions.

4.7 In-Hospital Mortality of Sepsis Differs Depending on The Origin of Infection: An Investigation of Predisposing Factors.

4.7.1 Study aim

The aims of this study were 1. to define in-hospital mortality depending on the origin of infection and 2. To investigate predictors of in-hospital mortality for each of the most common types of sepsis: abdominal, urinary and chest sepsis.

4.7.2 Outcome

The primary outcome was In-Hospital Mortality, which was coded as a binary variable to indicate whether the patient was dead (‘1’) or alive (‘0’). Model performance was measured using the area under the receiver operator characteristic (AUC) curve. AUC means and confidence intervals (CI) were calculated for each sepsis type.

4.7.3 Multiple Logistic Regression

Multiple logistic regression (LR) was used throughout the experiments. LR models the outcome probability or risk to be '1' (positive class) as $P(Y = 1) = 1/\left(1 + \exp\left[-\sum_{k=0}^{K} \beta_k X_k\right]\right)$, where $\{\beta_0, \ldots, \beta_K\}$ are the model coefficients which are estimated by maximum likelihood [126]. The LR coefficients are the logarithm odds ratios (OR) between the factors and the outcome. If a factor increased by one unit, its coefficient measures how much the outcome odd would increase or decrease, depending on whether the coefficient is positive or negative.

4.7.4 Model Explainability

To provide model explainability, we developed a forest plot for each sepsis type and a Sankey network diagram. The forest plots display the ORs, and CIs associated with each clinical feature relevant to the developed LR models. The Sankey network diagram was used in a novel way to visualise the interactions between the significant clinical features and sepsis groups. For this, we selected the significant variables ($P < 0.05$) from the LR models (nodes on the left-hand side of the diagram) and generated links between them and the sepsis groups (nodes on the right-hand side of the diagram). Additionally, the absolute value of the OR interactions between clinical features and sepsis groups was represented by the height of the nodes, to provide further information regarding the relevance of each clinical feature.

4.7.5 Comparisons of the Novel Models Against Established Critical Care Deterioration Scores

We compared the performance of two commonly used clinical scoring systems, the APACHE IV and SOFA score, which are typically used to predict in-hospital mortality for patients in critical care. We used the SOFA and APACHE IV scores as independent variables in a univariate LR model to produce the mortality risk estimate for the outcome. The purpose was to allow for a fair comparison between the developed models and the scores using the same methodology to evaluate how well each of them can predict the outcome.

The APACHE IV and SOFA scores are readily available in the eICU database. The APACHE IV scores were calculated based upon data collected on admission to the ICU, these values were available and listed in the eICU table 'apachePatientResults'. Individual components of the SOFA score were calculated [152] for the first 3 days and then averaged. qSOFA scores were calculated by assigning points for 1. altered mental state (<15 in the Glasgow Coma Scale), 2. Fast respiratory rate ($\geq$22 breaths per minute) 3. Low blood pressure (systolic blood pressure $\leq$ 100 mmHg). Further definitions of the acuity score's weighting can be found in the supplementary materials.

4.7.6 Results

4.7.6.1 Sepsis Groups

A total of 3,958 ICU admissions were analysed. 2393 patients were admitted with pulmonary sepsis, 1044 with urinary sepsis and 544 with abdominal sepsis (Figure 11). Unadjusted statistical comparisons between the three sepsis groups are displayed in Table 9. Patients with urinary sepsis were older than patients with pulmonary and abdominal sepsis.

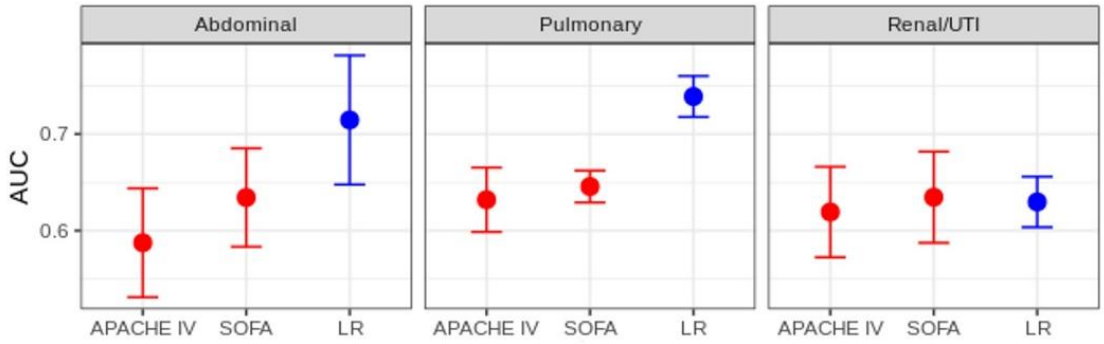
Except for hypertension, there were no significant differences in cardiovascular comorbidities between the groups. We found statistically significant group differences (p-value<0.05) for comorbidities such as mild and severe liver disease, dementia and respiratory diseases (COPD, asthma). We also observed significant group differences in vital signs (average heart rate, average mean arterial pressure, average saturation, average respiratory rate and average temperature) and blood counts (average lymphocyte count, average white blood cell count, average platelet count and haematocrit). Blood gas results differed between groups with regard to average pH, average PaO2 and average PaCO2. Liver and

kidney function was also significantly different between groups. Compared to patients with pulmonary or abdominal sepsis, a smaller proportion of patients with urinary sepsis required inotropes during their stay.

While there was a significant difference between SOFA and qSOFA scores between the groups, the Charlson comorbidity index and APACHE IV score were comparable between abdominal, urinary and pulmonary sepsis.

4.7.6.2 Evaluation of Model Performances

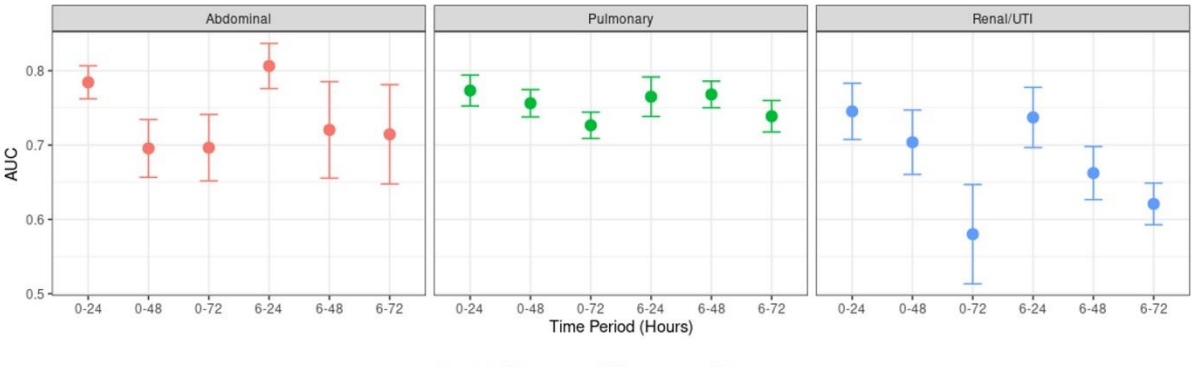
Figure 12 displays the results of the comparison between the developed multivariate models and the APACHE IV and SOFA scores. These AUC results show that, for pulmonary and abdominal sepsis, the novel models outperformed APACHE IV and SOFA scores (AUC 0.74 and 0.71, respectively), but were not superior in urinary sepsis (AUC 0.63). The AUC means and confidence intervals were approximated from the outer 10-folds of the nested cross-validation method implemented.



Sepsis Type	Algorithms	AUC	Sensitivity	Specificity
Abdominal	APACHE IV	0.59 (0.53-0.64)	0.54 (0.45 0.63)	0.65 (0.60-0.70)
	SOFA	0.63 (0.58-0.69)	0.57 (0.44-0.69)	0.61 (0.56-0.66)
	LR	0.71 (0.65-0.78)	0.54 (0.43-0.65)	0.73 (0.68-0.79)
Pulmonary	APACHE IV	0.63 (0.60-0.67)	0.54 (0.48-0.59)	0.66 (0.64-0.68)
	SOFA	0.65 (0.63-0.66)	0.60 (0.57-0.64)	0.60 (0.59-0.61)
	LR	0.74 (0.72-0.76)	0.65 (0.61-0.70)	0.70 (0.67-0.73)
Renal/UTI	APACHE IV	0.62 (0.57-0.67)	0.53 (0.45-0.61)	0.65 (0.63-0.66)
	SOFA	0.63 (0.59-0.68)	0.54 (0.45-0.63)	0.69 (0.67-0.72)
	LR	0.63 (0.60-0.66)	0.48 (0.41-0.55)	0.70 (0.66-0.73)

Figure 12: Model performance comparisons. Top) Area under the ROC curve (AUC) for each sepsis group. Average AUC (filled circles) and confidence intervals (vertical bars) were estimated after the 10 repetitions of the outer cross-validation. Deterioration scores (APACHE IV and SOFA) models are represented in red, LR models in blue. Bottom) Detailed comparison, also including sensitivity and specificity. Acronyms used: APACHE IV: Acute Physiology and Chronic Health Evaluation IV, SOFA: Sequential Organ Failure Assessment, LR: multiple logistic regression.

Comparisons using different time windows for data extraction was performed to assess a) how this decision impacts model performances, and b) how our analysis compares to previous studies. Figure 3 compiles the results obtained for the first 24, 48 and 72 hours, with or without the inclusion of the first 6 hours. The best results were obtained when using the first 24 hours, where the cohort sizes were generally twice the size of those at 72 hours (see the bottom of Figure 13), as a great proportion of patients either died or were discharged between 24 and 72 hours after ICU admission. Bootstrapping was not considered as the method to provide parameter estimates due to computational costs, and it may not be the most appropriate estimator of the mean values regarding performance [153].



Sepsis Type	Time Split	Cohort Size	Discharged	Died	Prevalence	Lost in the first 24h	Discharged	Died	Lost in the first 48h	Discharged	Died
Pulmonary	0-24	4,856	4046 [83%]	810 [17%]	16.7	-	-	-	-	-	-
	0-48	3,442	2833 [82%]	609 [18%]	17.7	1,414	1213 [86%]	201 [14%]	-	-	-
	0-72	2,414	1948 [81%]	466 [19%]	19.3	-	-	-	2,442	2098 [86%]	344[14%]
	6-24	4,475	3715 [83%]	760 [17%]	17.0	-	-	-	-	-	-
	6-48	3,393	2790 [82%]	603 [18%]	17.8	1082	925 [85%]	157 [15%]	-	-	-
	6-72	2,392	1931 [81%]	461 [19%]	19.3	-	-	-	2083	1784 [86%]	299 [14%]
Renal	0-24	2,772	2511 [91%]	261 [9%]	9.4	-	-	-	-	-	-
	0-48	1,713	1535 [90%]	178 [10%]	10.4	1059	976 [92%]	83 [8%]	-	-	-
	0-72	1,028	897 [87%]	131 [13%]	12.7	-	-	-	1744	1614 [93%]	130 [7%]
	6-24	2,546	2295 [90%]	251 [10%]	9.9	-	-	-	-	-	-
	6-48	1,692	1514 [89%]	178 [11%]	10.5	854	781 [91%]	73 [9%]	-	-	-
	6-72	1,022	891 [87%]	131 [13%]	12.8	-	-	-	1524	1404 [92%]	120 [8%]
Abdominal	0-24	1,408	1156 [82%]	252 [18%]	17.9	-	-	-	-	-	-
	0-48	884	730 [83%]	154 [17%]	17.4	524	426 [81%]	98 [19%]	-	-	-
	0-72	546	443 [81%]	103 [19%]	18.9	-	-	-	862	713 [83%]	149 [17%]
	6-24	1,344	1101 [82%]	243 [18%]	18.1	-	-	-	-	-	-
	6-48	876	723 [83%]	153 [17%]	17.5	468	378 [81%]	90 [19%]	-	-	-
	6-72	544	441 [81%]	103 [19%]	18.9	-	-	-	800	660 [83%]	140 [17%]

Figure 13: Model performance measures on several time windows. Top) Model performance comparisons as measured using the AUC for each sepsis group at several time intervals. The figure shows AUC means and confidence intervals estimated after the 10 repetitions of the outer cross-validation with logistic regression. Bottom) Effects of different time windows on cohort size and mortality rates.

4.7.6.3 Explanatory Analysis

Figure 14 displays the ORs for risk factors in the three sepsis groups as estimated across the 10 dataset instances. Higher age and higher average heart rate were associated with increased mortality risk. Increased values in average Mean Arterial Pressure (MAP) were associated with a reduced mortality risk across all sepsis groups. Our LR models identified significant factors that were relevant only for certain sepsis groups. For instance, atrial fibrillation and cancer were associated with an increased mortality risk only in pulmonary sepsis, but not in urinary or abdominal sepsis. Contrastingly, in abdominal sepsis hypertension represented a relevant risk factor of mortality. Interestingly, abdominal sepsis was the only group for which uncomplicated diabetes represented a significant protective factor regarding mortality risk.

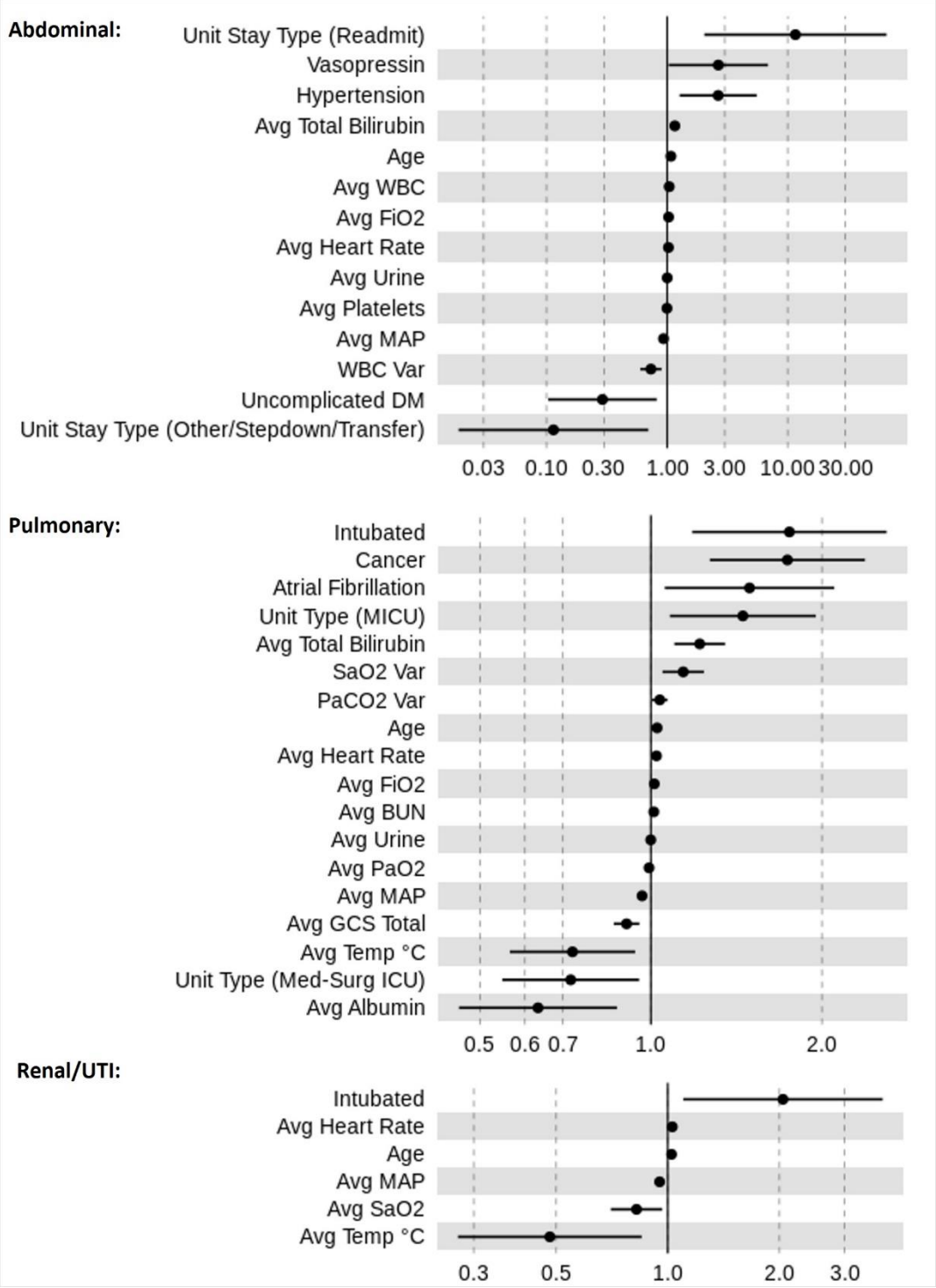


Figure 14: Odds ratio (OR) estimates for LR. The figure displays the pooled ORs average (filled circles) and confidence intervals (vertical bars) for all significant features ($p < 0.05$) selected by the feature selection algorithms for the sepsis groups: pulmonary, abdominal, and renal/UTI. An OR of 1 represents a baseline risk, with values < 1 indicating a reduction in risk for the outcome, and > 1 indicating an increased risk in relation to the outcome.

Several factors were relevant to more than one sepsis group. For instance, the most influential factor for increased mortality risk was "intubation" for urinary and pulmonary sepsis groups, however, in abdominal sepsis "readmitted to ICU" represented the most important factor. A rise in risk was associated with higher "average FiO2" and "average total bilirubin" values in both abdominal and pulmonary sepsis, but not in urinary sepsis. Distinctively, in pulmonary and renal sepsis lower average temperature was indicative of reduced mortality risk. The average albumin was associated with the

greatest risk reduction in pulmonary sepsis, whereas in renal and abdominal sepsis "average temperature" and "unit stay type (other/stepdown/transfer)" represented important variables.

Moreover, results illustrated that the average value for certain parameters was relevant while for other variables, the average variation played a greater role in mortality risk prediction. For instance, mortality risk reduces in renal sepsis when there is an increase in "average SaO2". This is dissimilar to pulmonary sepsis, for which higher "SaO2 variation" increased the risk of mortality.

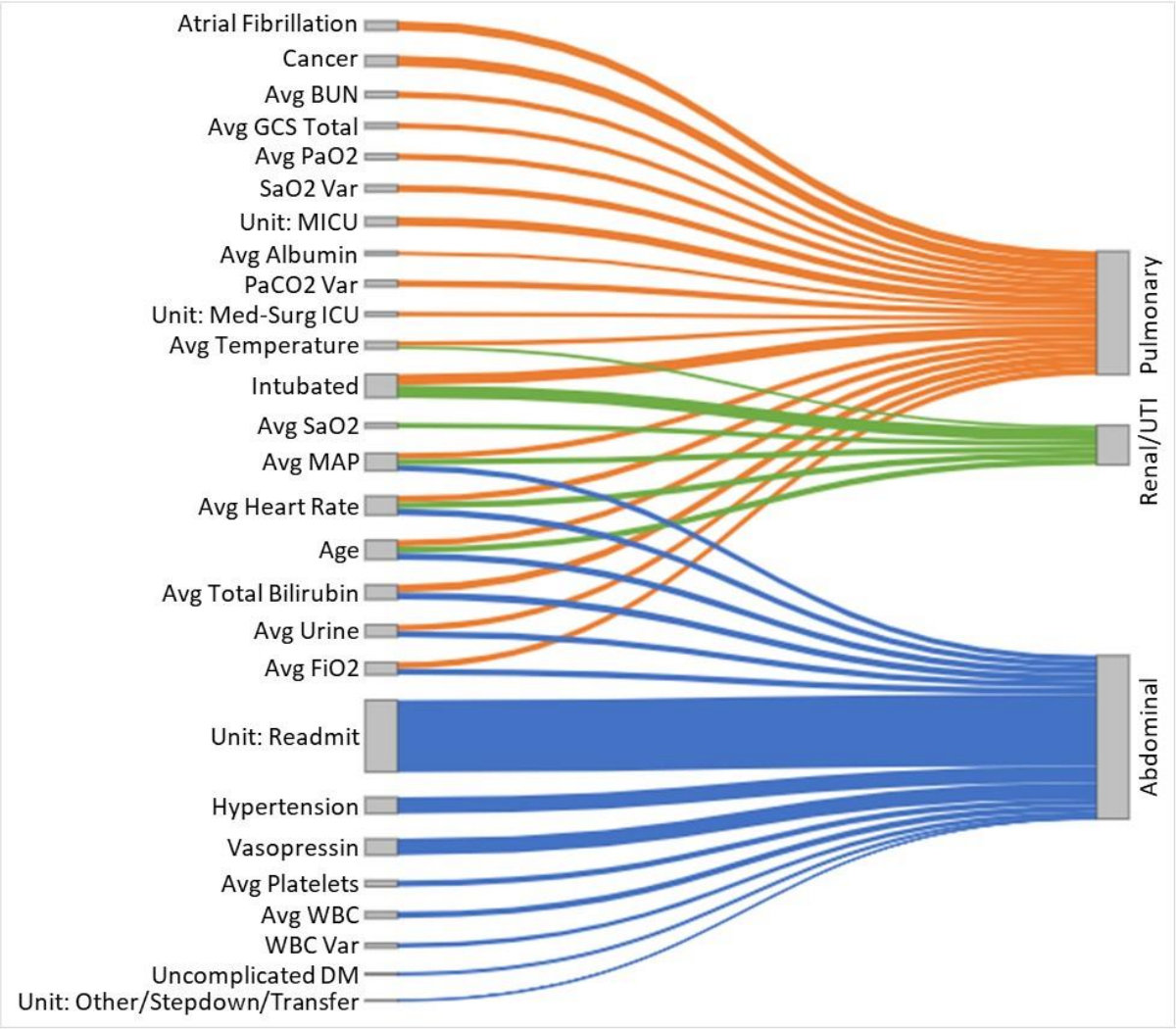


Figure 15: A Sankey diagram representing the relationship between several clinical features (nodes on the left-hand side) and the sepsis groups (nodes on the right-hand side), with the link widths representing the absolute ORs proportional to the risk of in-hospital mortality for each of the sepsis groups.

Figure 15 presents a Sankey network diagram displaying the relationship between several clinical features and the sepsis groups. It shows that "intubation", "average total bilirubin", "average FiO2", "average urine output", "average heart rate" and "average MAP" had the greatest overlap between sepsis groups. In abdominal sepsis, readmission had the greatest influence on the risk of in-hospital mortality compared to any other variables included in the model.

4.7.7 Discussion

In this study, we conduct a LR analysis of several types of sepsis based on the origin of infection. Our results showed that using LR as a relatively simple approach to ML was sufficient to obtain good to very good models for renal, abdominal and pulmonary sepsis that consistently outperformed the established risk scores for predicting in-hospital mortality. Biomedical and social scientists are usually

familiar with the results provided by LR models, hence their great popularity. The major drawbacks of LR are the linearity and normality assumptions of the data which could yield biased models.

Traditionally, outcome prediction in sepsis is based on clinical scores, such as SOFA, APACHE or SAPS. Such mortality prediction scores for critically ill patients are used worldwide and have been extensively validated [154]. These models, however, may not be ideal for routine clinical use as they lack granularity and are designed for use at ICU admission, thus neglecting the change of physiological parameters over time. So far, only a limited number of studies describe prognostication for in-hospital mortality in patients with sepsis comparing different sources of infection as an independent factor [155]. In this study, we address this knowledge gap by a) comparing different risk factors for each sepsis type and b) highlighting specific factors associated with in-hospital mortality in the distinct sepsis groups, depending on the origin of the underlying infection. This approach may help to address the heterogeneity of the patient population with sepsis, to define discrete patient populations to guide the development of effective therapies and identify cohorts that benefit from certain interventions.

A fundamental difference between our models and existing ones for outcome prediction is that we include data from a longer observation period. For frequently measured variables such as vital signs, up to 72-hours' worth of data points were used, with measurements recorded every hour. We extracted the mean and the standard deviation of all data points available to factor in change over time, with the former indicating the average values for each patient, and the latter indicating the range of variation in those values, e.g., a high heart rate variation may be indicative of some form of haemodynamic instability. However, the mean and the standard deviation represent a crude representation of change over time, and further research is required to investigate and define the best mathematical approach to reflect the variation of variables, particularly those with frequent measurement, e.g., heart rate or blood pressure.

We performed outcome prediction at various time points during the early phase of sepsis. Our results demonstrate that the performance of ML models drops over the first 72h after ICU admission in all the types of sepsis studied. Model performance is best maintained in pulmonary sepsis, while loss of performance is greatest in urinary sepsis. A possible explanation is that the causes of death from sepsis vary over time. While early deaths occur in about a third of septic patients and are mainly attributable to multiple organ failures caused by the primary infection, late deaths are influenced by end-of-life decisions and often relate to recurrent or late infections [156].

Early deaths in sepsis are typically associated with a hyperinflammatory 'cytokine storm' response with fever, refractory shock, acidosis and hypercatabolism [157]. If regulation of the immune response from hyperinflammation to normal activity fails after the acute phase, patients enter a marked immunosuppressive state. Later deaths after the acute phase occur due to an inability to clear primary infections and the development of secondary infections [157]. Taking into account the biphasic or even polyphasic course of sepsis, mortality prediction in the acute phase will differ from models predicting later mortality. Hence models that only include admission data are likely to disproportionately focus on early death occurring in the first 24h of admission. Whilst optimising data collection periods may improve outcome prediction, the ideal model should reflect dynamic changes and risk the profile throughout the Intensive Care admission.

Our results indicate that prediction after the acute phase of sepsis is more complex and not well described in existing prognostication models. In addition, outcome prognostication is often performed early during the ICU stay, and many scores such as APACHE IV, are only validated for use on admission to Critical Care. The degree of organ failure associated with the type of sepsis and the early progression of the disease varies between sepsis groups and may be influenced differently in each group by early deaths and vice versa, early recovery and discharge alive. This assumption is supported by the higher dropout of cases in the urinary sepsis group compared to other sepsis caused by abdominal and chest infections.

Sepsis is not a uniform disease, but a syndrome characterised by the striking variation of biological features [158]. Systematic analysis of these features, using data mining and advanced statistical methods or machine learning, may allow the identification of types of sepsis with different risk profiles and responses to treatment. In an attempt to classify different types of sepsis, several approaches have been chosen [159]. More sophisticated definitions of distinct molecular endotypes are based on leukocyte genome-wide expression profiles from samples collected on ICU admission [160]–[162]. However, the implementation of these complex prognostic and predictive strategies at the bedside of patients is limited [163] due to the need for expensive laboratory analysis, which is not routinely available and is often too time-consuming to allow clinical decision-making. Different statistical methods, including latent class analysis [164], [165], group-based trajectory modelling [166] and various machine learning algorithms [167] have been applied to large clinical data sets.

Clinicians instantaneously recognise that bacterial sepsis in young otherwise healthy patients carries a better prognosis than fungal sepsis in an elderly haematology patient. Similarly, urinary sepsis is commonly perceived as less fatal than chest or intraabdominal sepsis. A systematic review which addressed the impact of the source of infection on mortality [168], identified several studies in which lower in-hospital mortality was observed for urinary sepsis compared to respiratory sepsis. This observation was independent of the stage of sepsis with lower mortality observed in sepsis, severe sepsis and septic shock. Our results confirm the observation that in-hospital mortality is lower in critically ill patients with urinary sepsis compared to abdominal and respiratory sepsis. Factors influencing mortality differed between sepsis groups in our research, e. g. ICU readmission was a significant risk factor in abdominal sepsis, but played no role in pulmonary or urinary sepsis, indicating that the numerous ICU stays required for complex abdominal sepsis are associated with a worsening prognosis. In contrast, for pulmonary and urinary sepsis, the need for invasive ventilation was a significant risk factor for mortality. The origin of infection is often known to treating physicians early in the clinical course and as such, outcome prediction based on the type of causative infection using clinical data only, may be easier to implement than models relying on complex combinations of clinical data and biomarkers, which are often not readily available at the bedside. Modern monitoring devices allow the integration of such prognostic algorithms into their software package and facilitate easy clinical implementations for all patients requiring regular monitoring.

The strength of this study is that we used the eICU database, a public database containing a large number of datasets for critically ill patients to generate our models. Moreover, we included time series for vital signs and laboratory tests for up to 72h after admission to ICU in patients with different origins of sepsis and demonstrated that the models outperformed existing prediction tools. However, our study also has limitations. External validation and comparison with other machine learning approaches are required to explore the transferability and generalisability of our models in different critical care settings. Furthermore, the combination of molecular diagnostics such as transcriptomics and genomics with the routinely available clinical data used in our model may further improve the performance.

4.7.8 Conclusions

We present a logistic regression framework for different types of sepsis which are defined by their origin of infection using routinely available clinical data from a large publicly available dataset. Our model outperforms routinely used prediction tools such as SOFA and APACHE scores and considers changes in parameters over a 72h period. We demonstrate that factors of importance show considerable heterogeneity depending on the source of infection. Beyond outcome prediction, our results are important for treatment decisions and the planning of research studies. The observation that factors influencing outcome vary depending on the source of sepsis may explain why most sepsis trials have failed to identify an effective treatment.

4.8 Machine Learning Methods For Analysing Sepsis Depending On The Origin Of Infection: An Investigation Of Predisposing Factors For Mortality and LOS.

4.8.1 Study aim

The aims of this study were 1.) To investigate several outcomes and predictors for in-hospital, in-ICU mortality, hospital-LOS and ICU LOS for each of the most common types of sepsis: abdominal, urinary and chest sepsis, 2.) To build onward from the last experiments and investigate a range of ML algorithms concerning performance and interpretability for the listed outcomes.

4.8.2 Outcome

The primary outcome was in-hospital mortality, in-ICU mortality, and hospital and ICU LOS, coded as a binary variable to indicate whether the patient was dead ('1') or alive ('0') for mortality classification. For LOS prediction, we modelled the duration in hours for both outcomes. Model performance was measured using the area under the receiver operator characteristic (AUC) curve. AUC means and confidence intervals (CI) were calculated for each sepsis type. The primary model performance for LOS predictions was measured using MSE.

4.8.3 Machine Learning Algorithms

We compare four commonly deployed ML methods, logistic and linear regression, with forward step-wise feature selection (LR), in addition to a random forest (RF) and gradient-boosted machines (GBM). The logistic regression coefficients are the logarithm odds ratios (LogORs) between the factors and the outcome, which makes them useful for explanatory analysis. Thus, if a factor increased one unit, its LogOR measures how much the outcome odd would increase or decrease, depending on whether the coefficient is positive or negative, respectively. Multiple linear regression models the outcome Y (continuous value) to be $\sum_{k=0}^{K} \beta_k X_k$ where $\{\beta_0, \dots, \beta_K\}$ are the model coefficients that estimate each factor's impact similarly to logistic regression. Thus, if a factor increased by one unit, the outcome Y measures how much the outcome would increase or decrease, again depending on whether the coefficient is positive or negative. RF repeatedly fits induction trees to several subsets of random samples with replacements extracted from the training set. RF predicts a new outcome in classification tasks by taking the majority vote. In a regression task, RF and GBM predicts a new outcome by averaging across all of the predicted Y values. GBM also uses many decision trees to make predictions, although any ML algorithm could be used. Unlike RF, GBM implements an iterative learning algorithm such that a new tree model is fitted on the cases that a previous tree performed inadequately. RF and GBM can also be used for explanatory analysis as they rank the input variables based on their relevance to the model predictions (variable importance).

4.8.4 Variable Selection and Hyperparameter Tuning

For Linear and logistic regression, input variables were automatically selected using a sequential forward search algorithm over 10 dataset instances (10-fold cross-validation). An inner cycle of 5-fold cross-validation was used for each iteration to select relevant variables. The selection algorithm starts with a baseline model (i.e., all coefficients but the intercept set to zero, $\beta_{k\neq 0} = 0$), and in each step, the variable which most improves the performance on the validation set is added. Several RF and GBM hyperparameters were tuned using the same validation splits as for LR. For the RF models, we tested a range of number of variables randomly sampled between 3 and 30, and a range of minimum node size (which controls the depth of the trees) between 3 and 30. For GBM models, a range of shrinkage values (which controls the impact of each additional fitted tree) from 10^{-5} to 10^{1} and a range of minimum number of observations in a node from 3 to 30 were tested. The RF and GBM variable importance of the models with the best hyperparameter set for each cross-validation cycle was also estimated.

4.8.5 Model Explanation

By explaining a model and providing insight, there is more chance of trust and usability in the model created. We used SHAP (SHapley Additive exPlanations) to explain the logic behind the predicted risk

values of our implemented RF and GBM models. Widely used in game theory [169], SHAP assumes that feature values of a data instance act like players in a coalition game and estimates their Shapley values, which would inform how to distribute the importance among the features fairly. SHAP uses the estimated Shapley values to score the association strength between feature values with the prediction [170]. They are implicitly normalised, which makes them easier to interpret and compare. Shapley values close to 1 or -1 indicate strong positive or negative associations with the outcome, respectively. Likewise, a weak association is represented by Shapley values close to 0. The SHAP results are typically summarised in a plot in which factors are listed on the vertical axis and sorted by variable importance (if a tree-based algorithm is used, like RF and GBM), their values, normalised and represented with colour codes, and their Shapley values, in the horizontal axis.

Additionally, we explored partial dependencies, which can be insightful in exploring how a particular feature variable relates to the outcome. Since each of the predictions is made using all information in all the other predictors of an observation, the prediction obtained from the partial dependence algorithms also contains this information. This means that the relationship displayed in a partial dependence plot contains all the relations between x_j and y, including the averaged effects of all interactions of x_j with all the other predictors X_j which is why this method gives the partial dependence rather than marginal dependence[171]. This method allows for a visualisation of the feature values concerning the outcome providing more granular insight into the rationale of the model's predictions

4.8.6 Results

4.8.6.1 Evaluation of Model Performances

The predictive performance is displayed in Figure 16 for all outcomes for the considered sepsis groups and ML algorithms. Collectively, ML algorithms outperformed traditional ICU algorithms (i.e., Apache-IV and SOFA). Additionally, hospital outcomes such as mortality and LOS generally show lower performance scores than ICU outcomes. This is particularly apparent with sepsis groups pulmonary and renal sepsis. RF is displayed as the most suitable algorithmic choice for collectively modelling ICU outcomes. Although it did not score the best predictive performance for every outcome for each sepsis group, it was shown 75% of the time to be the best-performing model in all cases. For ICU mortality, GBM showed superior performance for pulmonary and renal sepsis achieving AUC performances of 75% and 65%, respectively. Abdominal sepsis achieved an AUC of 73% for in-hospital mortality with RF, which outperformed GBM. Slightly superior AUC scores were achieved for in-ICU mortality for abdominal and renal sepsis compared to in-hospital, both achieving AUC scores of 74%. Pulmonary sepsis achieved a similar score of 75 % for in-ICU mortality, however, with a slightly reduced standard error. For hospital LOS, RF showed marginally superior performance for all sepsis groups compared to all methods implemented. Renal sepsis displayed the lowest MSE of 1.39 hours compared to pulmonary and abdominal sepsis, which received an MSE of 1.40 hours and 1.48 hours. For ICU LOS, RF displayed the best performance for pulmonary achieving an MSE of 1.28 hours, and abdominal achieving an MSE of 1.32 hours with GBM. However, for renal sepsis, all ML models scored an MSE for ICU LOS of 1.22 hours compared to the ICU models, which generated performances of 1.23 hours and 1.24 hours for SOFA and APACHE-IV.

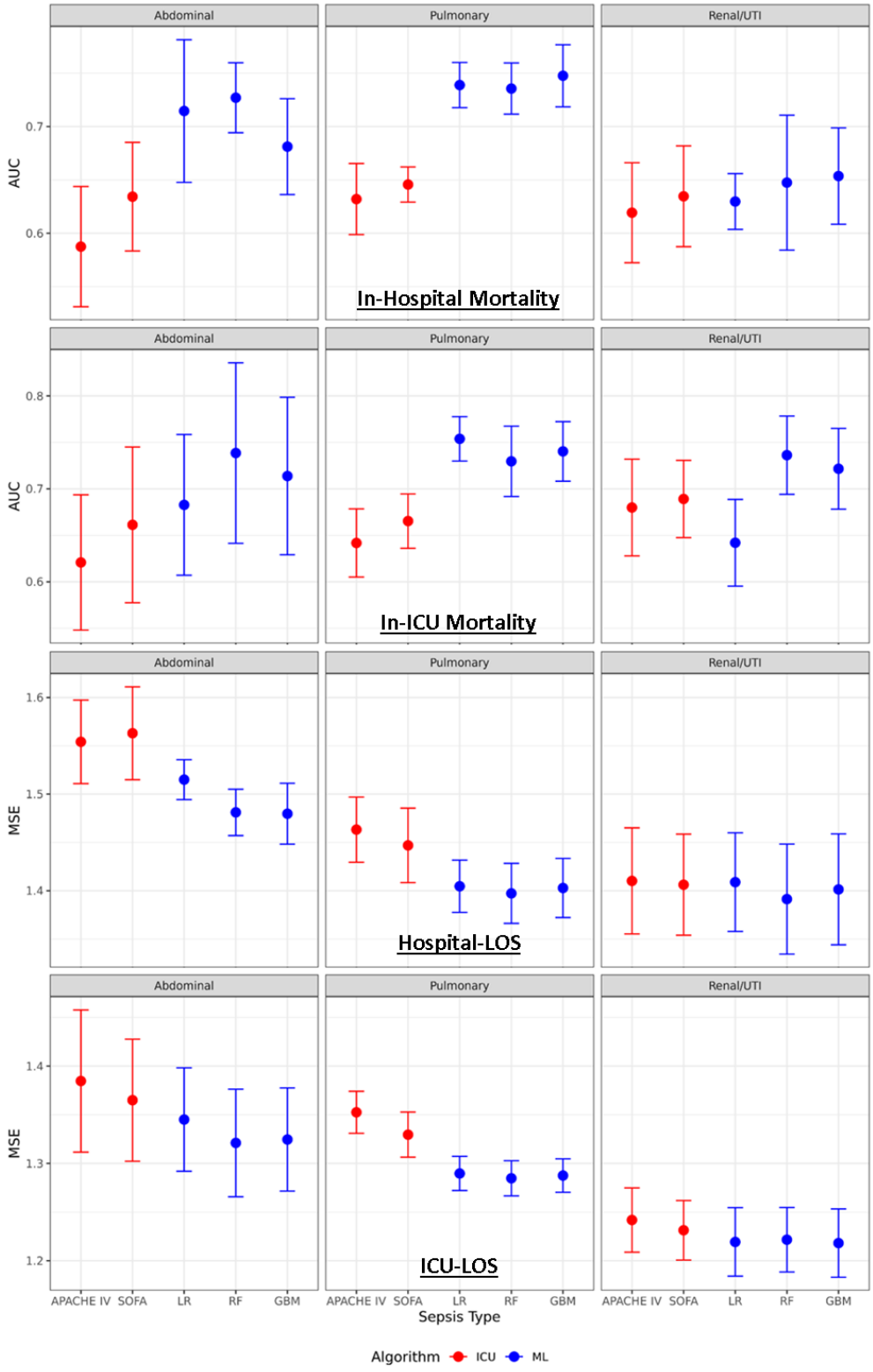


Figure 16: Model performance comparisons as measured using the area under the ROC curve (AUC) and mean square error (MSE) for each sepsis group. The figure shows AUC and MSE means (filled circles), and confidence intervals (vertical bars) estimated after the 10 repetitions of the outer cross-validation. The red are traditional ICU approaches to estimate In-hospital mortality, and blue ML methods applied. Acronyms used: APACHE IV: Acute Physiology and Chronic Health Evaluation IV, SOFA: Sequential Organ Failure Assessment, LR: multiple logistic regression, LR: multiple linear regression, RF: random forest, GBM: gradient boosted machines.

Table 10: Model performance measures the mean and confidence intervals for the area under the ROC curve (AUC), sensitivity (true positive rate), and specificity (true negative rate) for mortality classification. For LOS predictions, performance measures the mean and confidence intervals for means square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). The calibration value or class threshold was fixed to reflect the sepsis group mortality prevalence. For abdominal sepsis, this value was 18.93%, respectfully.

Abdominal Sepsis			
Hospital Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.59 (0.53-0.64)	0.54 (0.45 0.63)	0.65 (0.60-0.70)
SOFA	0.63 (0.58-0.69)	0.57 (0.44-0.69)	0.61 (0.56-0.66)
LR	0.71 (0.65-0.78)	0.54 (0.43-0.65)	0.73 (0.68-0.79)
RF	0.73 (0.69-0.76)	0.79 (0.73-0.84)	0.54 (0.52-0.57)
GBM	0.68 (0.64-0.73)	0.54 (0.46-0.61)	0.71 (0.66-0.77)
ICU Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.62 (0.58-0.66)	0.55 (0.48-0.63)	0.66 (0.64-0.69)
GBM	0.71 (0.67-0.76)	0.52 (0.45-0.58)	0.77 (0.74-0.79)
LR	0.68 (0.64-0.72)	0.36 (0.29-0.43)	0.83 (0.81-0.85)
RF	0.74 (0.69-0.79)	0.78 (0.71-0.85)	0.58 (0.55-0.6)
SOFA	0.66 (0.62-0.7)	0.55 (0.48-0.62)	0.66 (0.64-0.68)
Hospital LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.55 (1.53-1.58)	1.70 (1.69-1.71)	1.94 (1.92-1.96)
GBM	1.48 (1.46-1.50)	1.65 (1.63-1.66)	1.87 (1.85-1.89)
LR	1.51 (1.50-1.53)	1.69 (1.68-1.70)	1.90 (1.89-1.91)
RF	1.48 (1.47-1.49)	1.64 (1.63-1.66)	1.87 (1.86-1.88)
SOFA	1.56 (1.54-1.59)	1.7 (1.69-1.72)	1.95 (1.93-1.97)
ICU LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.38 (1.35-1.42)	1.58 (1.56-1.6)	1.76 (1.72-1.8)
GBM	1.32 (1.3-1.35)	1.52 (1.5-1.54)	1.69 (1.66-1.73)
LR	1.35 (1.32-1.37)	1.54 (1.52-1.56)	1.72 (1.69-1.75)
RF	1.32 (1.29-1.35)	1.52 (1.5-1.54)	1.69 (1.65-1.72)
SOFA	1.36 (1.33-1.4)	1.56 (1.54-1.58)	1.74 (1.7-1.78)

Table 11: Model performance measures the mean and confidence intervals for the area under the ROC curve (AUC), sensitivity (true positive rate), and specificity (true negative rate) for mortality classification. For LOS predictions, performance measures the mean and confidence intervals for means square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). The calibration value or class threshold was fixed to reflect the sepsis group mortality prevalence: for pulmonary sepsis, this value was 19.27%, respectively.

Pulmonary Sepsis			
Hospital Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.63 (0.60-0.67)	0.54 (0.48-0.59)	0.66 (0.64-0.68)
SOFA	0.65 (0.63-0.66)	0.60 (0.57-0.64)	0.60 (0.59-0.61)
LR	0.74 (0.72-0.76)	0.65 (0.61-0.70)	0.70 (0.67-0.73)
RF	0.74 (0.71-0.76)	0.76 (0.72-0.80)	0.59 (0.57-0.61)
GBM	0.75 (0.72-0.78)	0.65 (0.60-0.70)	0.71 (0.69-0.74)
ICU Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.64 (0.62-0.66)	0.54 (0.5-0.57)	0.66 (0.65-0.67)
GBM	0.74 (0.72-0.76)	0.62 (0.58-0.66)	0.73 (0.72-0.74)
LR	0.75 (0.74-0.77)	0.67 (0.64-0.7)	0.71 (0.7-0.72)
RF	0.73 (0.71-0.75)	0.73 (0.69-0.76)	0.58 (0.57-0.59)
SOFA	0.67 (0.65-0.68)	0.56 (0.54-0.59)	0.68 (0.68-0.69)
Hospital LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.46 (1.45-1.48)	1.64 (1.62-1.65)	1.85 (1.83-1.87)
GBM	1.40 (1.39-1.42)	1.59 (1.57-1.60)	1.79 (1.77-1.81)
LR	1.40 (1.39-1.42)	1.59 (1.57-1.60)	1.79 (1.77-1.81)
RF	1.40 (1.38-1.41)	1.58 (1.57-1.59)	1.78 (1.76-1.80)
SOFA	1.45 (1.43-1.47)	1.62 (1.6-1.64)	1.83 (1.81-1.86)
ICU LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.35 (1.34-1.36)	1.57 (1.56-1.58)	1.73 (1.72-1.74)
GBM	1.29 (1.28-1.3)	1.5 (1.49-1.51)	1.65 (1.64-1.66)
LR	1.29 (1.28-1.3)	1.5 (1.49-1.51)	1.66 (1.64-1.67)
RF	1.28 (1.28-1.29)	1.5 (1.49-1.51)	1.65 (1.64-1.66)
SOFA	1.33 (1.32-1.34)	1.54 (1.53-1.55)	1.7 (1.69-1.72)

Table 12: Model performance measures the mean and confidence intervals for the area under the ROC curve (AUC), sensitivity (true positive rate), and specificity (true negative rate) for mortality classification. For LOS predictions, performance measures the mean and confidence intervals for means square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). The calibration value or class threshold reflected the sepsis group mortality prevalence. For renal sepsis, this value was 12.81%, respectfully.

Renal/UTI Sepsis			
Hospital Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.62 (0.57-0.67)	0.53 (0.45-0.61)	0.65 (0.63-0.66)
SOFA	0.63 (0.59-0.68)	0.54 (0.45-0.63)	0.69 (0.67-0.72)
LR	0.63 (0.60-0.66)	0.48 (0.41-0.55)	0.70 (0.66-0.73)
RF	0.65 (0.58-0.71)	0.68 (0.57-0.79)	0.54 (0.52-0.57)
GBM	0.65 (0.61-0.70)	0.54 (0.43-0.64)	0.67 (0.59-0.76)
ICU Mortality			
Algorithms	AUC*	Sensitivity	Specificity
APACHE IV	0.68 (0.65-0.71)	0.59 (0.55-0.63)	0.68 (0.66-0.69)
GBM	0.72 (0.7-0.74)	0.54 (0.48-0.6)	0.76 (0.74-0.77)
LR	0.64 (0.62-0.67)	0.44 (0.39-0.49)	0.77 (0.76-0.78)
RF	0.74 (0.71-0.76)	0.76 (0.72-0.81)	0.57 (0.55-0.58)
SOFA	0.69 (0.67-0.71)	0.63 (0.57-0.68)	0.69 (0.68-0.7)
Hospital LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.41 (1.38-1.44)	1.59 (1.57-1.61)	1.79 (1.76-1.82)
GBM	1.40 (1.37-1.43)	1.58 (1.56-1.60)	1.78 (1.75-1.81)
LR	1.41 (1.38-1.43)	1.59 (1.57-1.60)	1.79 (1.76-1.82)
RF	1.39 (1.36-1.42)	1.57 (1.55-1.59)	1.77 (1.74-1.80)
SOFA	1.41 (1.38-1.43)	1.58 (1.57-1.60)	1.79 (1.76-1.82)
ICU LOS			
Algorithms	MSE*	MAE	RMSE
APACHE IV	1.24 (1.23-1.26)	1.44 (1.43-1.46)	1.59 (1.57-1.61)
GBM	1.22 (1.2-1.24)	1.41 (1.4-1.43)	1.56 (1.53-1.58)
LR	1.22 (1.2-1.24)	1.41 (1.4-1.43)	1.56 (1.53-1.58)
RF	1.22 (1.2-1.24)	1.42 (1.41-1.43)	1.56 (1.54-1.58)
SOFA	1.23 (1.22-1.25)	1.43 (1.42-1.44)	1.58 (1.55-1.6)

4.8.6.2 SHARP Analysis with the Optimal Machine Learning Model for In-Hospital Mortality using GBM

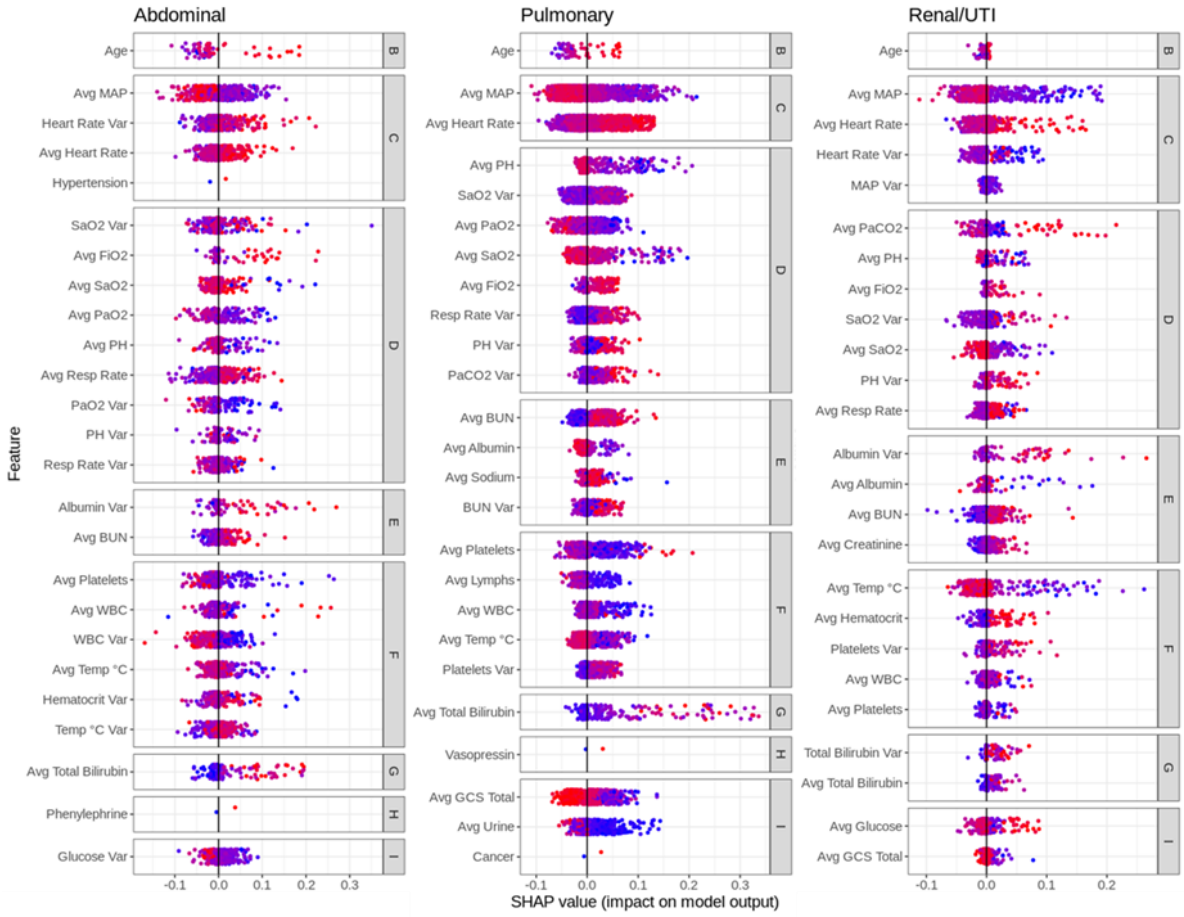


Figure 17: Summary of the SHAP values calculated by the RF models for Pulmonary, Abdominal and Renal sepsis. The colour represents the value of the feature from low to high. Variables with the highest importance are displayed, organised by clinical groups: A, Admission diagnosis; B, Demographics; C, Cardiovascular; D, Respiratory; E, Renal; F, Immune response; G, Liver; H, Drugs; and Unit Type, K, Unit Stay Type, L, task), Acronyms and short names used.

Figure 17 displays SHAP results for GBM for all sepsis groups for in-hospital mortality. It displays the top 25 features selected by GBM for each sepsis group, aggregated by the clinical groups assigned to each variable. The results show unique differences among the variables selected for each sepsis group. Collectively, categorical features were only displayed as necessary for abdominal and pulmonary sepsis compared to renal sepsis. It can be seen that some features such as age, MAP and total bilirubin, amongst others, are essential across all groups, though individualities remain present. For instance, RF's results for respiratory factors displayed the average pH most relevant to the pulmonary sepsis group. In contrast, SaO2 variation and average PaCO2 were ranked more important for renal and abdominal sepsis. For the cardiovascular features used, all sepsis groups ranked average MAP as the most impactful factor in addition to selecting heart rate, both average and variation depending on the sepsis group. For the renal features selected, abdominal and renal both rank albumin variation as the most important factors compared to pulmonary, which ranked average BUN as the most important. For features in group immune response, average palettes were ranked as the most important for abdominal and pulmonary sepsis compared to renal, which ranked this variable as the least important in this category and ranked average temperature as the most important factor. For liver features, average total bilirubin was selected as an important feature across all sepsis groups however, in renal sepsis, the variation is also listed as a significant factor, unlike in other sepsis groups. Only pulmonary and abdominal sepsis selected drugs as important when modelling the sepsis groups for drugs administered. For pulmonary and abdominal sepsis, vasopressin and phenylephrine were selected, and both displayed an increased mortality risk for each sepsis group. For group other features, each sepsis group ranked

the clinical features uniquely, with abdominal and renal sepsis selecting glucose variation and average glucose. In contrast, pulmonary sepsis ranked avg GCS total as the most important for this group.

4.8.6.3 Partial Dependencies Analysis using Optimal Machine learning model for In-hospital Mortality.

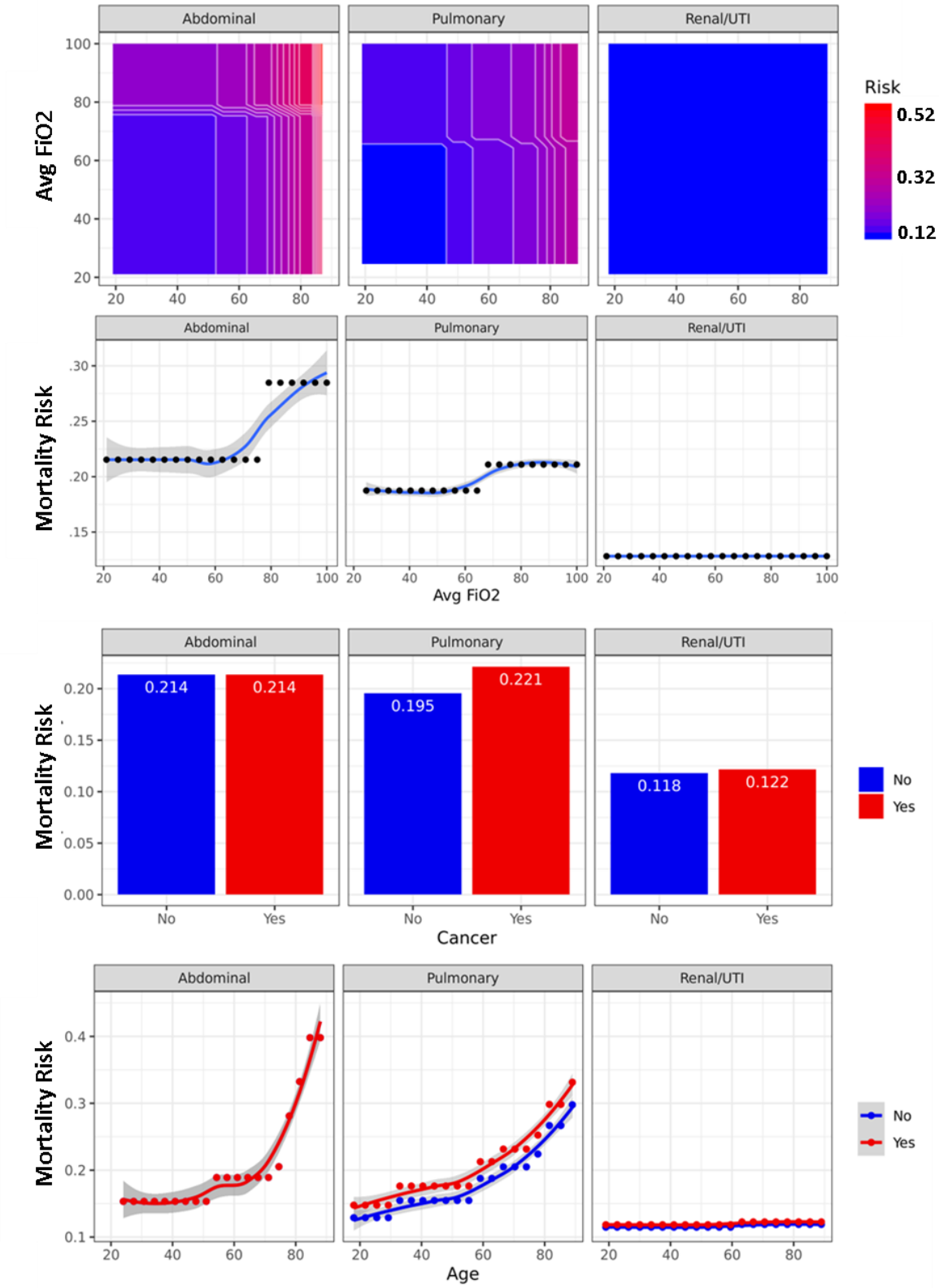


Figure 18: Partial dependencies analysis of cancer and avg Fi02 for pulmonary, abdominal, and renal sepsis.

Figure 18 displays the partial dependencies plots for GBM for in-hospital mortality. It displays the impact of risk in association with a feature class or value. The first figure displays two features as interaction terms (age and average FiO2) concerning the risk of hospital mortality. The second figure shows the risk associated with a range of FiO2 Values. The third plot displays the risk associated with feature cancer across the three sepsis groups. For abdominal sepsis, the risk associated with both classes is the same; however, for renal, the risk associated with the positive class increases slightly, similar to pulmonary sepsis. Finally, the last plot in the figure displays the sepsis group's interaction with age and cancer. The results show a linear relationship with risk as age increases for pulmonary and renal sepsis. However, abdominal showed no difference in risk. Similarly, the partial dependencies visualisations displayed unique risk profiles associated with the outcome for all sepsis groups.

4.8.6.4 SHARP Analysis using the Optimal Machine learning Model for In-hospital LOS with RF

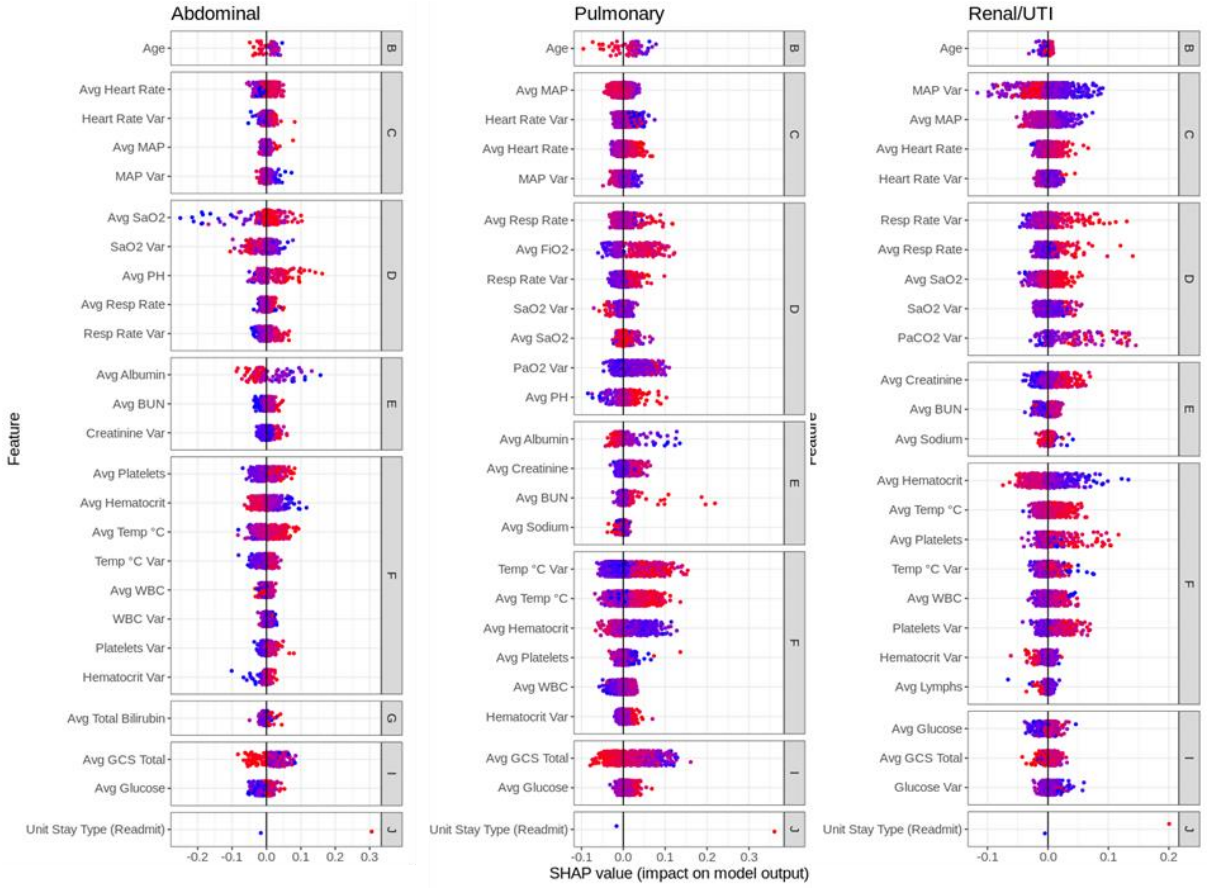


Figure 19: Summary of the SHAP values calculated by the RF models for Pulmonary, Abdominal and Renal sepsis. The colour represents the value of the feature from low to high. Variables with the highest importance are displayed, organised by clinical groups: A, Admission diagnosis; B, Demographics; C, Cardiovascular; D, Respiratory; E, Renal; F, Immune response; G, Liver; H, Drugs; and Unit Type, K, Unit Stay Type, L, task. Acronyms and short names are used.

Figure 19 displays SHAP results for RF for all sepsis groups for hospital-LOS. It displays the top 25 features selected by RF for each sepsis group, aggregated by the clinical groups assigned to each variable, similar to the previous analysis. The results show unique differences among the variables selected for each sepsis group. Only a single categorical feature was selected for all sepsis groups, which was unit stay type (readmit), which showed a significant impact on all sepsis groups regarding hospital LOS. Similar to the results generated from the in-hospital mortality features selected by GBM, some features, such as average GCS total, unit stay type, and age, are essential across all sepsis groups. However, individuality still remains regarding hospital LOS among the sepsis groups.

For instance, RF's results for respiratory factors displayed the average respiratory rate as most relevant to the pulmonary sepsis group. In contrast, the average SaO2 and respiratory rate variation were ranked more critical for abdominal and renal sepsis. For the cardiovascular features, unlike results for in-hospital mortality, each sepsis group had unique rankings; average heart rate, average MAP and MAP variation were selected for abdominal, pulmonary and renal sepsis. However, all sepsis groups selected the same features. For the renal features selected, abdominal and pulmonary both rank average albumin as essential factors compared to renal, which ranked average creatine as the most important from that group. Each sepsis group showed unique rankings for group immune response features, with average platelets and average albumin ranking as the most important for abdominal and pulmonary sepsis. In contrast, renal sepsis ranked average hematocrit as the most important in this category, whereas the other sepsis groups ranked this as the least important in that category. Only abdominal sepsis selected features from this class (i.e., average total bilirubin) for liver features, compared to the other sepsis types. No drugs were displayed as important for all sepsis groups. For other group features, each sepsis group ranked the clinical features uniquely, with abdominal and pulmonary sepsis selected average GSC total, compared to renal sepsis, which ranked average glaucous as the most important for this group.

4.8.6.5 Partial Dependencies Analysis using the Optimal Machine Learning Model for Hospital-LOS.

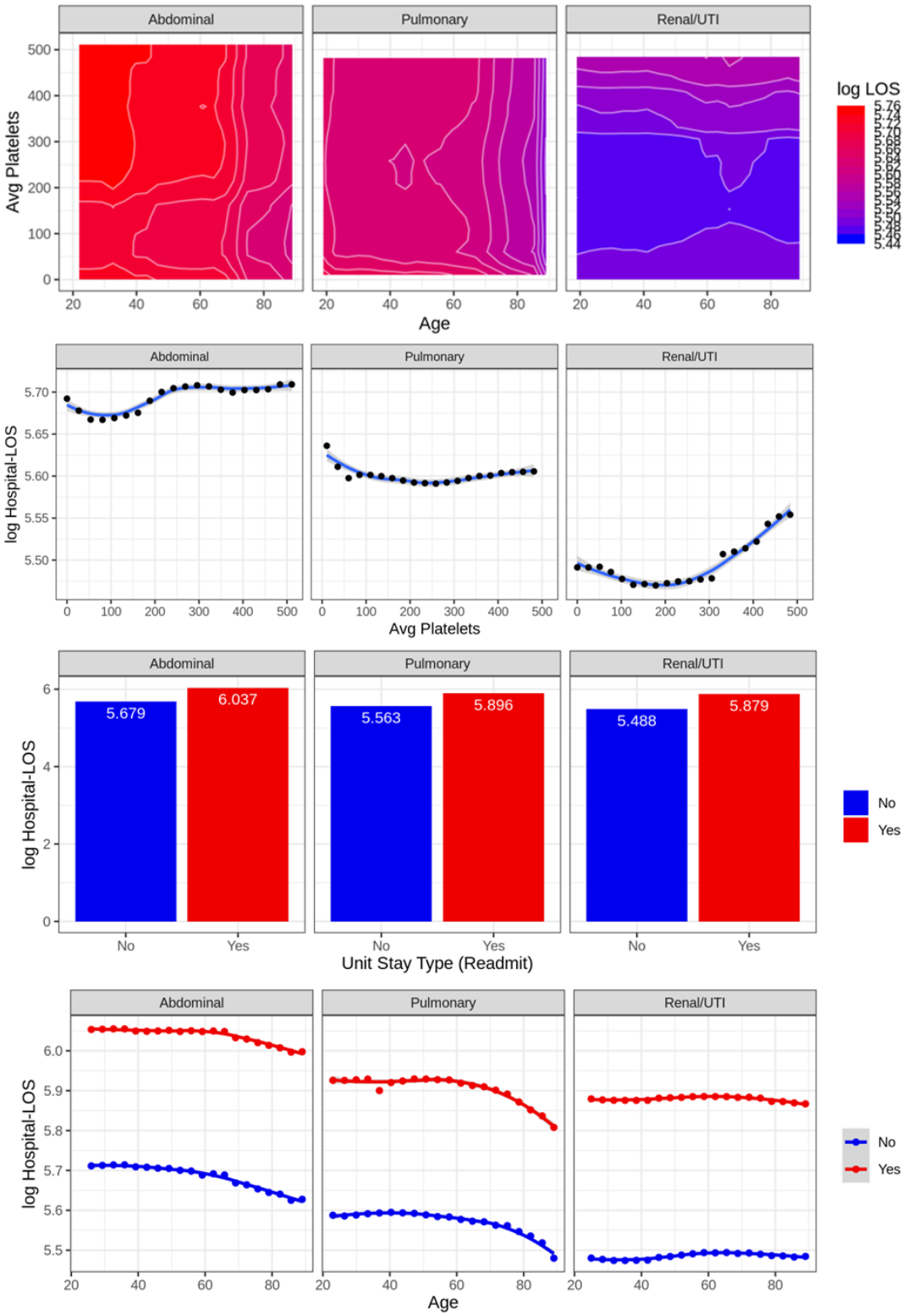


Figure 20: Partial dependencies analysis of unit stay type and avg platelets for pulmonary, abdominal and renal sepsis.

Figure 20 displays the partial dependencies plots for RF for hospital LOS displaying the impact of LOS in association with a feature class or value. The first figure displays two features as interaction terms (age and average platelets) in relation to hospital-LOS within a log scale. The second plot of the figure

displays the LOS associated with a range of platelet values. The third plot displays the LOS associated with feature unit stay type (readmit) across the three sepsis groups. The LOS associated with both classes is clearly unique for all sepsis types in relation to unit stay types. Lastly, the final plot in the figure displays the interaction with age and unit stay type (readmit) for the sepsis groups. The results show an unusual trend; as age increases, the LOS decrease for abdominal and pulmonary, whereas this is not as apparent with renal sepsis. A similar result is demonstrated where all sepsis groups display unique LOS dependencies associated with the hospital LOS.

4.8.6.6 Variable Importance Analysis with the Optimal ML Model for In-ICU Mortality using RF

Figure 21: Summary of the variable importance values calculated by the RF models for Pulmonary, Abdominal and Renal sepsis. Variables with the highest importance are displayed and organised by clinical groups: A, Admission diagnosis; B, Demographics; C, Cardiovascular; D, Respiratory; E, Renal; F, Immune response; G, Liver; H, Drugs; and Unit Type, K, Unit Stay Type, L, task. Acronyms and short names are used.

Figure 21 displays the variable importance for ICU mortality. A similar result was obtained with selected features compared to in-hospital mortality. Comparably to the results produced for in-hospital mortality, each of the sepsis groups listed a similar but unique range of features for each of the sepsis

cohorts, thus again emphasising the sepsis group's individuality, producing a similar yet unique list of features for ICU mortality prediction different for hospital mortality prediction.

4.8.6.7 Forest Plots and Odd Ratio Analysis of ICU LOS with Linear Regression

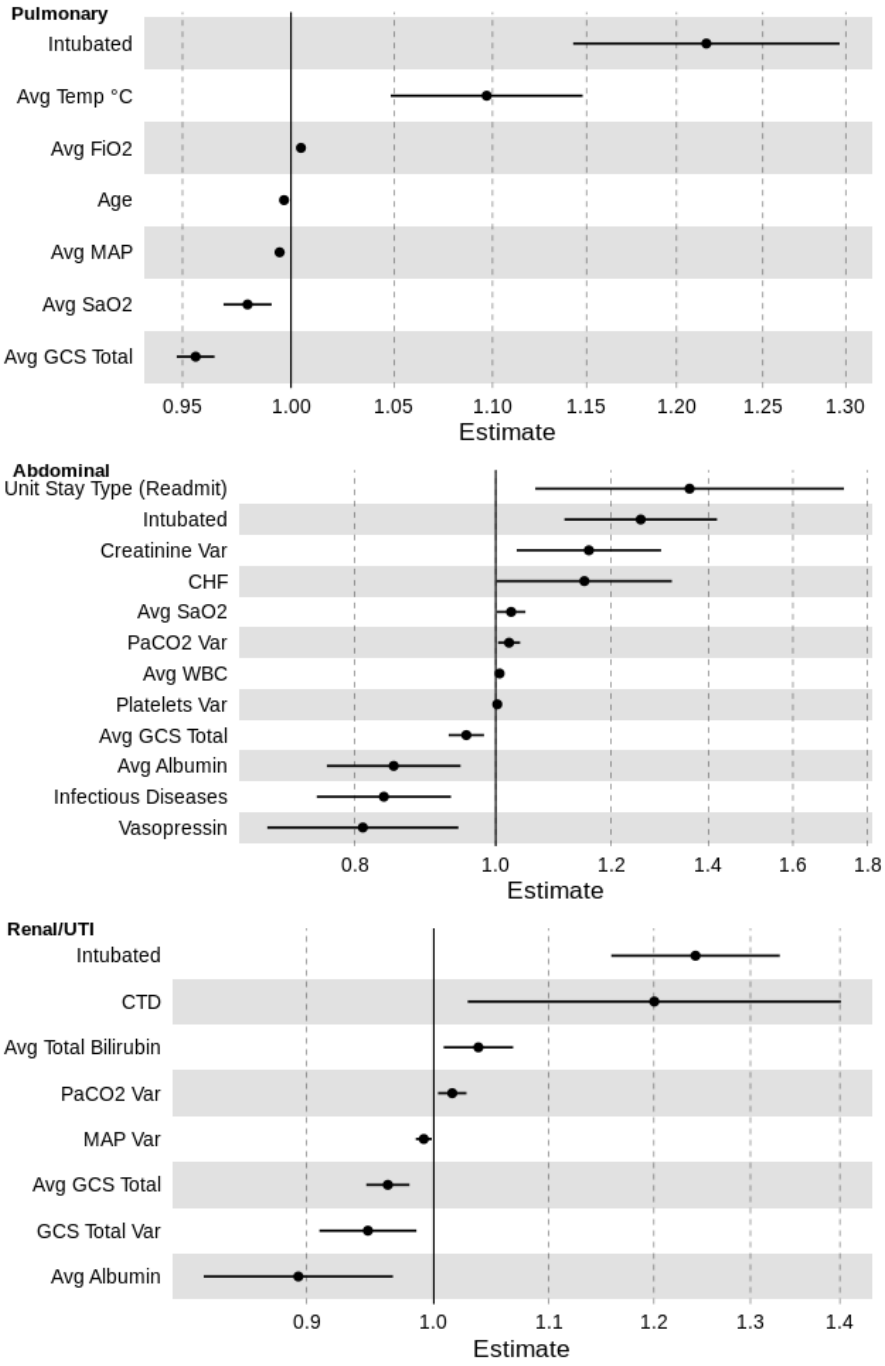


Figure 22: The beta coefficients estimate for Linear Regression. The figure displays the pooled beta values, the average (filled circles) and confidence intervals (vertical bars) for all significant features ($p < 0.05$) selected by the feature selection algorithms for the sepsis groups: pulmonary, abdominal, and renal/UTI. A beta value of 1 represents a baseline risk, with values < 1 indicating a reduction in LOS for the outcome and > 1 indicating an increased LOS concerning the outcome.

Figure 22 displays the significant coefficients of the linear regression model for modelling ICU LOS. Similarly to the results generated from hospital-LOS, the sepsis groups again display individuality in the features which were deemed necessary when modelling. Likewise, the sepsis groups displayed significant features for all groups, such as intubation and GCS total. However, it is clear that the

heterogeneity in the sepsis cohorts is present as each sepsis group resulted in a list of unique variables, similar to the results of hospital-LOS prediction.

4.8.7 Discussion

So far, only a limited number of studies describe prognostication of in-hospital mortality and other similar outcomes in patients with sepsis by comparing and optimising different ML techniques[172]. In this study, we address this knowledge gap not only by comparing different ML techniques but also by introducing specifically designed models to prognosticate a range of outcomes in distinct subgroups of sepsis depending on the origin of the underlying infection.

For future clinical use, ML offers the opportunity to extract common patterns in patients presenting different sepsis manifestations and transfer this information into group-specific algorithms. As a result, our ML approach has improved the often-moderate performance of existing in-hospital mortality or LOS prediction models in all sepsis subgroups, excluding renal sepsis, where traditional ICU modelling methods delivered a similar performance. Furthermore, the ML models unfold the homogeneities and inhomogeneities of different types of sepsis depending on the origin of the underlying infection. This information is not available in traditional ICU models and is beneficial not only for direct clinical care but also for research methodology, e.g., for the design of sepsis studies.

Nowadays, ML-based prediction models can be integrated into clinical monitoring systems to support clinical assessment and to allow early treatment decisions. Therefore, the integration of prognostic algorithms is crucial for developing ML-supported clinical decision-support tools. A recent systematic review demonstrates that 40% of ML-supported clinical decision-support tools have been developed using data from ICU databases[173]. However, different types of sepsis defined by the origin of infection in our methodological approach have not yet been considered in developing such ML-based decision-support tools. Furthermore, current approaches are limited, particularly when used for conditions representing complex syndromes with different pathogenetic origins, such as sepsis.

4.8.7.1 Choice of Machine Learning Algorithms

In principle, a range of ML algorithms could have been used to model the outcomes. In our selection, we favoured ML algorithms that could offer a significant level of interpretability by design: linear and logistic regression (LR), RF, and GBM. Logistic and linear regression are well-known linear models regarded as the gold standard in life and social sciences for multivariate statistical analysis of binary and continuous outcomes[174]. Logistic regression models the logarithm of the outcome odds as the linear combination of the input variables (or factors). Linear regression models the outcome value as a linear combination of the input variables, similar to logistic regression without the logit function. This allows for a direct interpretation of the learnt feature parameters, which correspond to the coefficients learnt when training allowing for a direct interpretation of these factors concerning the outcome. Logistic and linear regression has shown to be highly competitive when the data is not noisy, and variables are relatively independent and perform similarly in predictive performance compared to other ML approaches [175]. Furthermore, biomedical and social scientists are usually familiar with the results provided by logistic and linear regression models hence their great popularity. Though we must consider LR's significant drawbacks, these are the data's linearity and normality assumptions, which could yield biased models, especially when the data becomes more complex and non-linear.

RF and GBM are non-linear ML algorithms which account for these assumptions. In this study, we show that non-linear models, such as RF and GBM, predominately outperformed traditional methods, such as APACHE and SOFA, in addition to the linear models implemented. This may be due to the algorithmic design. They use ensemble learning to predict the outcome based on many decision trees (typically in the order of hundreds). An essential characteristic of those algorithms is that their decision trees use a random selection of the variables every time branch splits are completed. This allows for collecting information about which variables were more important to the predictions. As such, RF and GBM variable importance rankings are typically reported as a way to interpret their decision-making. The impact of the input variables on the predictions of RF and GBM can be further explored using the

SHAP analysis at a local and global level for each prediction. The results from the SHAP analysis and corresponding figures also reveal the homogeneities and inhomogeneities of different types of sepsis depending on the origin of the underlying infection. This was further emphasised with the partial dependency plots, which revealed differences at a feature level amongst the sepsis groups.

We acknowledge that more sophisticated ML methods could have been applied, such as neural networks (i.e., both shallow and deep feedforward neural nets) or methods such as support vector machines, which may or may not have achieved slightly better performance. However, although the same inference methods could be applied, such as SHAP or the investigation into partial dependencies, an inherent computational complexity is present. It takes considerable computational power to obtain the 'optimal' set of hyperparameters for these modelling approaches. This, combined with the ten-by-five-fold nested cross-validation method, makes the computational cost exponential and thus unfeasible for this analysis.

4.8.7.2 Machine Learning Interpretability

The increasing performance of ML models has led to them being encountered more frequently in daily life, including in clinical medicine [176]. Nowadays, complex ML models are outperforming traditional models in healthcare. However, those are often difficult to understand because of a lack of intuitiveness, difficult interpretation and lack of explanations of model predictions. ML models are often difficult to interpret due to the complexity and the lack of transparency in the process that was used to produce the final output. Furthermore, the lack of explanations regarding the decisions made by models represents a significant shortcoming in critical decision-making processes[177]. Therefore, ML models must have two main characteristics understandability and explainability. Understandability is related to the question of how the observer comprehends an explanation. Interpretability and explainability are similar concepts, often used interchangeably [178]. Interpretability can be identified as if the model's operations can be understood [179]. Interpretability of highly complex prediction models is needed in healthcare because of the nature of the work. To understand and accept or reject prediction, end-users and healthcare workers must understand the reasoning behind the prediction models[166]. Furthermore, the lack of interpretability is a crucial factor limiting wider ML adoption in healthcare. Subsequently, healthcare workers often find it challenging to trust complex ML models because the models are often designed and rigorously evaluated on specific diseases in a narrow environment and depending on one's technical knowledge of statistics and ML [177]. Moreover, most models in the literature focus on accuracy prediction and rarely explain their predictions in a meaningful way [180], [181]. This is especially problematic in healthcare applications where achieving high predictive accuracy is often as crucial as understanding the prediction. In addition to technical challenges related to the development of an interpretable model, we also need to address a myriad of ethical, legal and regulatory challenges, for example, the GDPRs right to explanation [182].

For these limitations currently in health care regarding ML applications, we have focused on delivering a range of approaches which aim to tackle these issues through the use of odd ratios, variable importance and SHAP. From a holistic viewpoint, SHAP provides a local and global interpretation of which features impact the prediction and is best suited to meet these requirements currently in healthcare. However, most practical applications of prediction models in healthcare are focused on the individual and would therefore require model-specific interpretability approaches to allow the highest possible level of interpretability [177]. To trust and maintain fairness and transparency of a specific model and its predictions, we must understand different approaches to model interpretability.

4.8.8 Conclusion

In this research, we conduct an ML analysis of several sepsis groups and outcomes. Our results showed that using relatively simple ML approaches was sufficient to obtain models that consistently outperformed the traditional ICU methods of measuring the risk of mortality and LOS in both the hospital and the ICU setting. Importantly, we showed that the proposed ML methods implemented can also be used for explanatory analysis giving granular insight into which features are essential and

influential when modelling the different types of sepsis, which, compared to the traditional approach such as APACHE and SOFA, may not have been feasible. Lastly, we displayed that non-linear models outperform linear approaches and current ICU modelling practices.

4.9 Multi-Task Learning for Model Optimisation: In-Hospital Mortality Analysis of Sepsis Patients

4.9.1 Introduction

Building on previous experiments regarding sepsis, we have postulated that depending on the source of the infection can lead to different manifestations and outcomes observed clinically. Therefore, identifying critical factors and biomarkers can be used to improve prognostic ML models. However, existing ML models do not consider sepsis's origin for mortality prediction, as previously discussed. In this research, we address the various manifestations of sepsis associated with the origin of infection and implement the use of multi-task learning (MTL)[183], [184]. MTL is a branch of ML that implements learning strategies where several tasks can be learned simultaneously. The rationale behind the MTL modelling of sepsis with different origins is that learning the individual functions, i.e., the learning of each individual sepsis manifestation, might benefit from sharing information between them, i.e., sharing information with the other sepsis groups. The aim is to increase the overall predictive performance for each sepsis group. Therefore, an additional sepsis group was constructed and was included in this analysis, which previously, due to their low sample size and their less defined characterisation (i.e., unknown, other, gynecologic, cutaneous), were excluded as we could not successfully analyse these sepsis groups using standard single task learning (STL). Thus, these cases were aggregated into a single group called 'Unknown/Other'.

We understand learning a task as the process of fitting a statistical or ML model. In this sense, a task encapsulates a dataset along with details about input variables and an outcome. Having enough data samples is typically enough for learning a task. However, it is frequent to encounter tasks that are very difficult for algorithms to learn. This is because they are not well defined either because there are multiple possible solutions or because a solution cannot be uniquely determined. If no constraints are imposed, the best solution is always to collect more data[185]. However, this is not always possible or not necessarily desired[186]. In this study, we show how MTL helps to effectively solve tasks that may be inadequately characterised, often due to reduced sample size. There are several approaches to the MTL idea, but in general, and in contrast to STL, MTL allows for several tasks to share information during the learning process to improve the model performance of all the individual tasks. STL consists of learning a map from one or several (input) variables to a target (or output) using a single dataset. Most ML algorithms are designed to work with STL. On the contrary, MTL is defined as the process of learning more than one task simultaneously. The general assumption in MTL is that when the involved tasks are close enough between themselves, then the individual model performance of each task will improve[184], [187].

MTL has proven to be successful in domains such as drug discovery[188], [189], genetics[190], and bioinformatics[191]. In healthcare, MTL is commonly used to account for group differences in electronic health records[192]–[195], focusing on improving model performance. However, less attention has been paid to the use of MTL in explanatory modelling, which is a frequent need in healthcare[196]. In this research, we selected this strategy to 1) improve the model performance of the individual sepsis groups and 2) exploit the ability of MTL to be used for the explanatory modelling of different sepsis groups.

4.9.1.1 Multi-task Learning Applications and Use Cases

Multi-task learning (MTL) is a ML strategy in which multiple related tasks are trained jointly instead of independently, with the overall goal of improving the generalisation and overall performance of the learning model. MTL improves generalisation by leveraging the domain-specific information contained in the training signals of the related tasks. This is done by training tasks in parallel while using a shared

representation, thus in effect, training signals for the extra task serve as an inductive bias [184]. There are several different branches of MTL, however, the underlying principle of each approach is to optimise the models performance by using available data from related tasks, which leads to an improved generalisation for the model, thus, improved performance. Transfer learning (TL) aims to produce an effective model for a target task with limited or no labelled training data by leveraging and exploiting knowledge from different but related source domains to predict the truth for an unseen target instance. By enhancing the training with supplementary labelled data from a related source domain, the model's ability to increase performance can be improved. However, the main challenge with TL, similarly to MTL, is distinguishing beneficial knowledge in the source input data [197][198].

4.9.2 Study Aim

This study aimed to build on previous experiments and to 1. To improve the model performance of the individual sepsis groups, and 2. To exploit the ability of MTL to be used for the explanatory modelling of different sepsis groups.

4.9.3 Outcome

The primary outcome was In-hospital mortality, coded as a binary variable to indicate whether the patient was dead ('1') or alive ('0') for mortality classification. Model performance was measured using the area under the receiver operator characteristic (AUC) curve. AUC means and confidence intervals (CI) were calculated for each sepsis type.

4.9.4 Single-Task and Multi-Task Learning Strategies

This section describes the proposed strategy for modelling sepsis groups using MTL. The primary motivation for this strategy is that relevant features to individual sepsis groups are identified during the learning process. As standard in ML, the STL strategy assumes that a model is fitted using an ML algorithm $\mathcal{A}$ and a dataset $\mathcal{D}$. Typically, $\mathcal{D}$ is randomly split into training, validation, and test subsets, $\mathcal{D}_{train}$, $\mathcal{D}_{val}$, and $\mathcal{D}_{test}$, respectively. The model is fitted using $\mathcal{D}_{train}$, whilst $\mathcal{D}_{val}$ is used to select the best possible set of parameters $\{\theta_i\}_{i=1\dots p}$. The test subset $\mathcal{D}_{test}$ is set aside and not used during the training process. It is only used to assess the performance of the fitted model. This strategy is known as holdout resampling and could be sufficient for quick model implementation. However, it is preferred to repeat this process several times either by using a k-fold cross-validation resampling strategy on the training and validation subsets while setting aside a test subset or by implementing a full nested cross-validation strategy: an inner loop is designed for hyperparameter selection and an outer loop for model evaluation. The entire nested cross-validation could also be repeated several times to reduce bias risk. We followed a repeated nested cross-validation strategy throughout the experiments. In particular, a ten-by-five nested cross-validation was implemented in all the experiments.

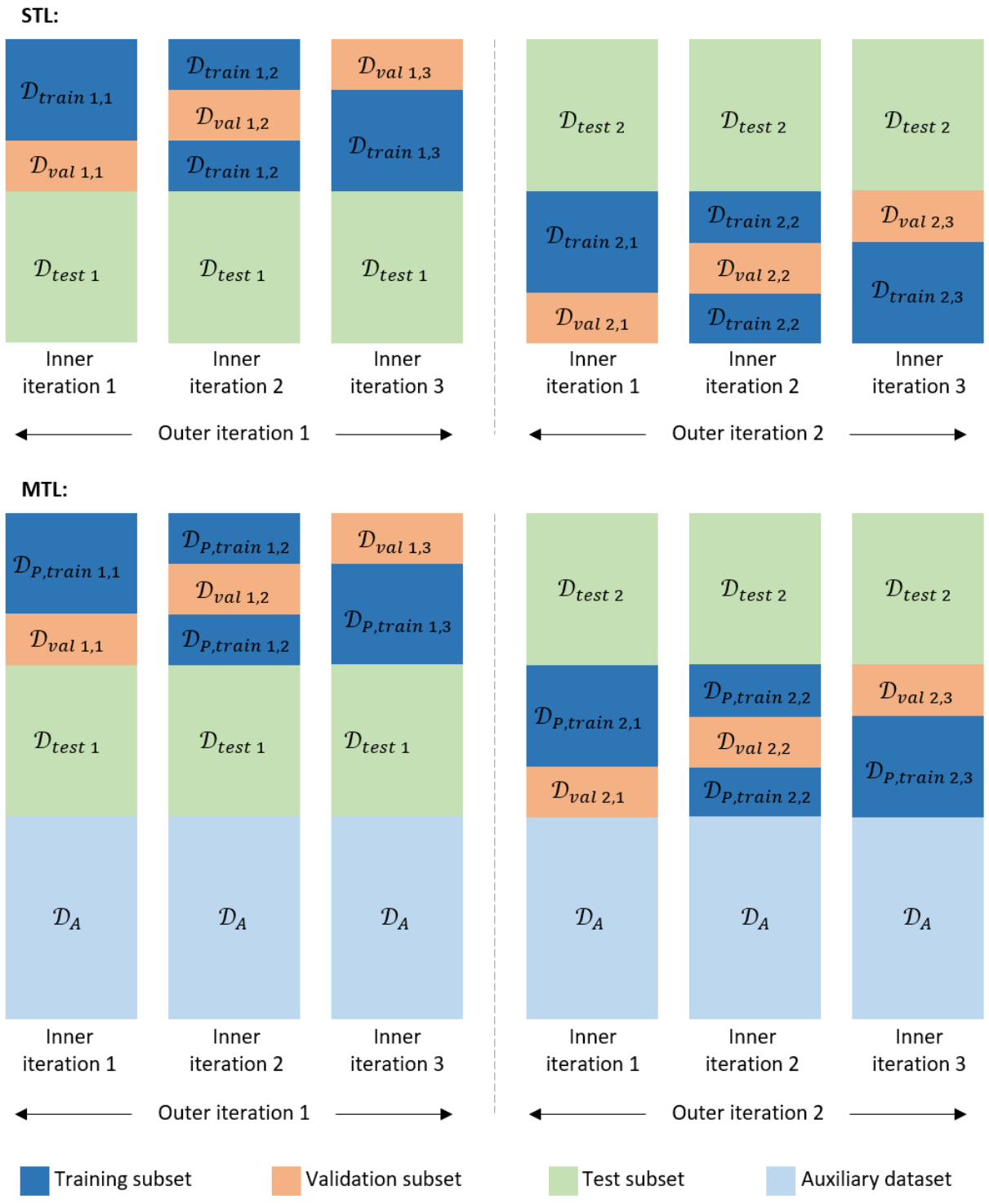


Figure 23: Single-task learning (STL) vs multi-task learning (MTL) validation strategies.

As mentioned before, the implemented MTL consists of primary and auxiliary tasks, with $\mathcal{D}_P$ and $\mathcal{D}_A$ datasets, respectively. The aim is to learn the primary task with the support of the auxiliary task, which is only used for training purposes. Therefore, we extend the STL strategy to MTL by appending the dataset of the auxiliary task to the training subset of the primary task. That is, $\mathcal{D}_{train} = \{\mathcal{D}_{P,train}, \mathcal{D}_A\}$. The validation and test subsets remain untouched. Figure 1 shows the differences between the implemented STL and MTL strategies. It is worth noting that the test and validation subsets in each iteration are only taken from the primary task. In this way, we can guarantee that the hyperparameter selection and the reported model performances are based only on the original (primary) task. In order

to perform a fair comparison between STL and MTL, we made sure that both strategies used the same data splits in all the experiments.

4.9.5 Model Performance

Model performances were measured using the area under the receiver operator characteristic (AUC) curve. We report AUC means and confidence intervals (CI) for each sepsis group, learning strategy, and ML algorithm from estimated AUCs using the test subsets. In addition, statistical t-test tests were performed to compare mean differences between learning strategies. Finally, reported p-values are corrected using Bonferroni to account for repeated experiments.

4.9.6 Results

4.10 Sepsis Groups

Following the aforementioned patient inclusion and exclusion criteria, 5,089 ICU admissions were eligible for the study, as displayed in Figure 24 This cohort includes 2,392 pulmonary, 1022 renal/UTI, 544 abdominal, and 1131 Unknown/Other sepsis admissions. The Uknown/Other sepsis group consists of 477 unknown sepsis admissions, 398 cutaneous sepsis admissions, 11 gynecologic sepsis admissions and 245 other sepsis admissions, respectively.

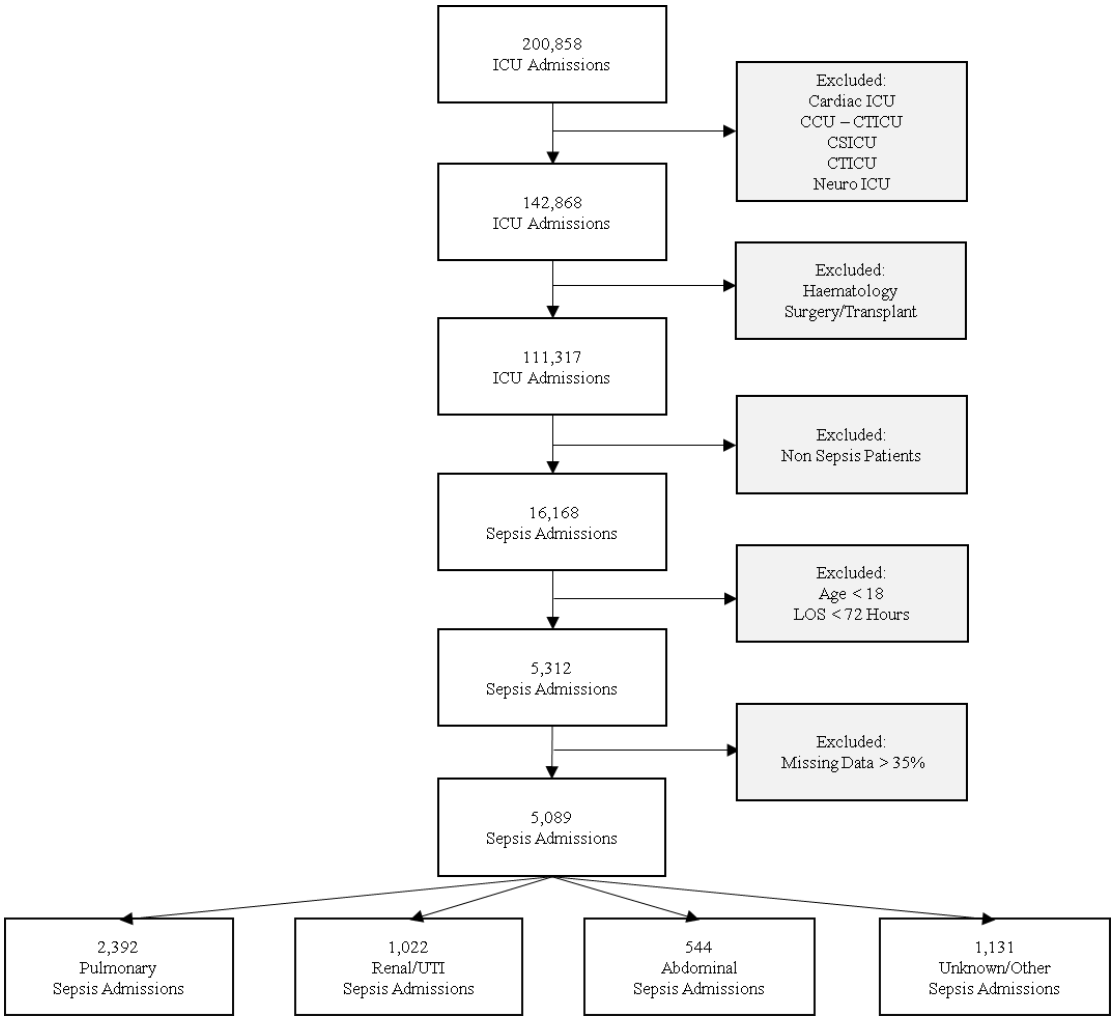


Figure 24: Flowchart of sepsis cohorts analysed showing the inclusion and exclusion criteria for the MTL sepsis study.

4.10.1.1 Evaluation of Model Performances

Figure 25 and Table 13 display model performance comparisons between the STL and MTL strategies across the considered sepsis groups and ML algorithms. Overall, MTL showed significantly better model performance, independently of the ML algorithm and data cohort used, with the pulmonary source modelling with RF as the only exception. It is also significant to note that STL never outperformed MTL in any of the head-to-head comparisons. Differences between strategies were more apparent in the renal/UTI tissue sepsis group but less evident in the pulmonary sepsis group. Statistical differences were observed for Abdominal and renal sepsis using GBM compared to STL vs MTL strategies, with renal additionally displaying a significant difference in performance with LR.

Table 13: Model performance comparisons as measured using the area under the ROC curve (AUC).

		AUC		
Sepsis	Algorithms	STL	MTL	P Value
Abdominal	APACHE IV	0.6 (0.56,0.64)	-	-
Abdominal	SOFA	0.61 (0.57,0.64)	-	-
Abdominal	LR	0.7 (0.63,0.77)	0.72 (0.66,0.78)	0.463
Abdominal	RF	0.72 (0.68,0.77)	0.72 (0.68,0.76)	0.901
Abdominal	GBM	0.66 (0.63,0.7)	0.77 (0.73,0.81)	0.00308
Pulmonary	APACHE IV	0.63 (0.6,0.67)	-	-
Pulmonary	SOFA	0.65 (0.63,0.66)	-	-
Pulmonary	LR	0.73 (0.71,0.76)	0.74 (0.72,0.76)	0.708
Pulmonary	RF	0.73 (0.71,0.75)	0.74 (0.72,0.77)	0.473
Pulmonary	GBM	0.75 (0.72,0.78)	0.76 (0.74,0.78)	0.572
Renal/UTI	APACHE IV	0.64 (0.59,0.69)	-	-
Renal/UTI	SOFA	0.64 (0.58,0.69)	-	-
Renal/UTI	LR	0.63 (0.58,0.68)	0.71 (0.67,0.75)	0.0248
Renal/UTI	RF	0.67 (0.63,0.72)	0.71 (0.67,0.76)	0.229
Renal/UTI	GBM	0.66 (0.63,0.69)	0.73 (0.68,0.78)	0.0412
Unknown/Other	APACHE IV	0.66 (0.6,0.73)	-	-
Unknown/Other	SOFA	0.67 (0.64,0.7)	-	-
Unknown/Other	LR	0.69 (0.65,0.74)	0.76 (0.73,0.79)	0.214
Unknown/Other	RF	0.73 (0.69,0.76)	0.73 (0.69,0.77)	0.328
Unknown/Other	GBM	0.73 (0.68,0.78)	0.78 (0.74,0.81)	0.156

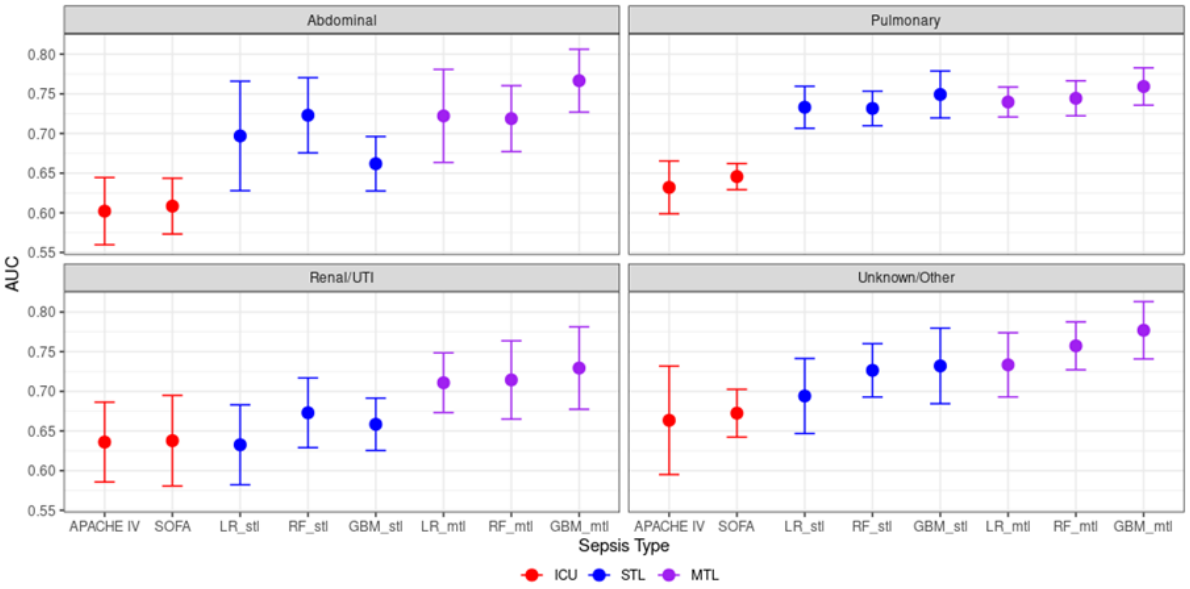


Figure 25: Model performance comparisons as measured using the area under the ROC curve (AUC) for each sepsis group. The figure shows AUC and MSE means (filled circles), and confidence intervals (vertical bars) estimated after the ten repetitions of the outer cross-validation. The red is traditional ICU approaches to estimate In-hospital mortality, and the blue ML methods applied, with purple highlighting the MTL strategy. Acronyms used: APACHE IV: Acute Physiology and Chronic Health Evaluation IV, SOFA: Sequential Organ Failure Assessment, LR: multiple logistic regression, LR: multiple linear regression, RF: random forest, GBM: gradient boosted machines. MTL: multi-task learning.

4.10.1.2 Evaluation of the SHAP Analysis Using the MTL-GBM Model

Figure 26 displays the SHAP values results generated by the MTL-GBM models for all the sepsis groups, aggregated by the clinical groups assigned to each variable regarding in-hospital mortality. In harmony with the previous analysis, the results show unique differences among the top 25 ranked variables by GBM for each sepsis group. We achieve greater predictive accuracy than the single-task method and the previous analysis undertaken. In addition, the MTL-GBM SHAP results show differences at the global and feature level regarding interpretability for the individual sepsis groups and the STL methods implemented in the previous analysis. Therefore, further displaying the heterogeneity between the sepsis types in the form of variable importance and value impact.

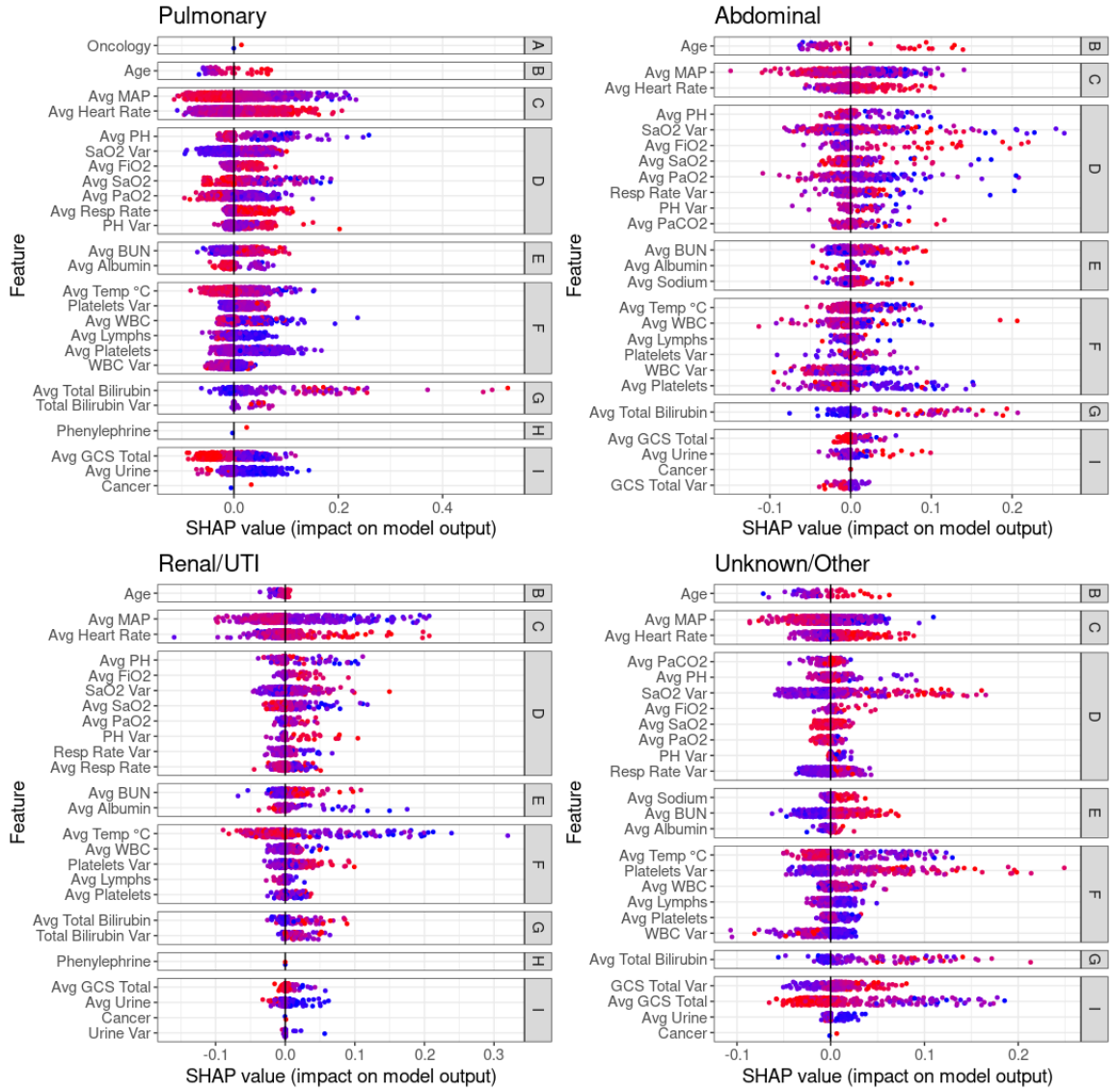


Figure 26: Summary of the top 25 SHAP values calculated by the MTL-GBM models for Pulmonary, Abdominal Renal and Uknown/Other sepsis. The colour represents the value of the feature from low to high. Variables with the highest importance are displayed, organised by clinical groups (A-K discussed in the results section). Acronyms and short names are used.

4.10.1.3 Evaluation of Odds Ratios using the MTL-LR Models

Figure 27 displays the MTL implementation of LR significant coefficients. Similarly to the results from GBM, the sepsis groups, in harmony with the previous experiments, displayed individuality in the significant features listed for each of the sepsis groups. In addition, compared to the single-task implementations, many more features were considered significant and impactful for each sepsis group. This may be due to feature contamination of the auxiliary tasks, which seems more visibly present with MTL-LR compared to the MTL-GBM algorithmic approach.

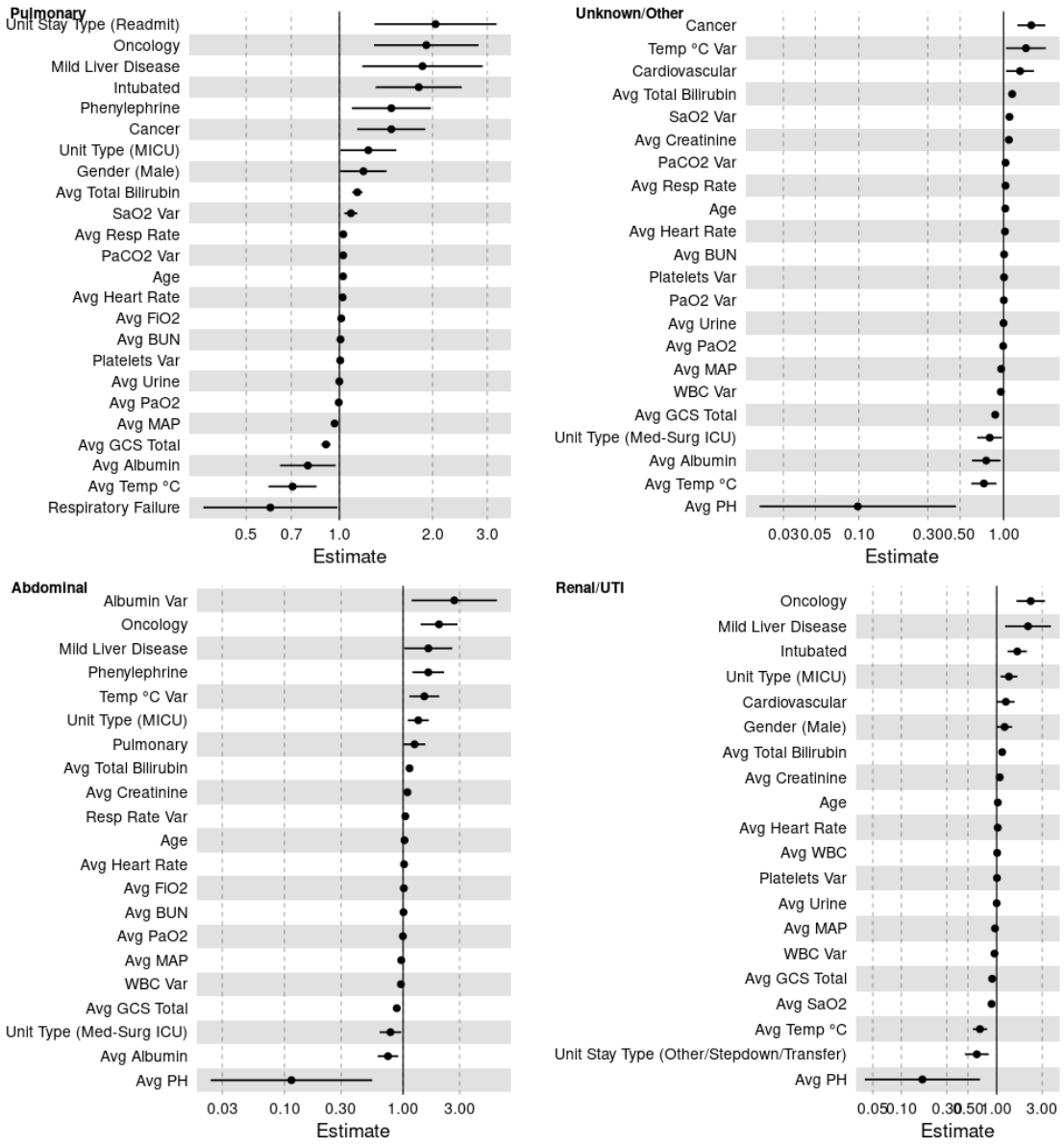


Figure 27: Odds ratio (OR) estimates for LR. The figure displays the pooled ORs average (filled circles) and confidence intervals (vertical bars) for all significant features ($p < 0.05$) selected by the feature selection algorithms for the sepsis groups: pulmonary, abdominal, renal/UTI, and 'Uknown/Other Sepsis. An OR of 1 represents a baseline risk, with values < 1 indicating a reduction in risk for the outcome and > 1 indicating an increased risk concerning the outcome.

4.10.2 Discussion

In this study, we compare different ML techniques by introducing MTL into various ML models specifically designed to prognosticate outcomes in distinct subgroups of sepsis depending on the origin of the underlying infection. For future clinical use, MTL offers the opportunity to extract common patterns in patients present with different patterns of sepsis and to transfer this information into group-specific algorithms. As a result, our approach in all sepsis subgroups has significantly improved the often-moderate performance of existing in-hospital mortality prediction models and commonly used ML approaches such as RF and GBM (i.e., single-task learning). Furthermore, the MTL approach unfolds the homogeneities and inhomogeneities of different types of sepsis depending on the origin of the underlying infection. This information is not available in traditional ML models and is beneficial not only for direct clinical care but also for research methodology, e.g., for the design of sepsis studies.

This research demonstrates how MTL can overcome these shortcomings, allowing for optimised predictive performance in conjunction with interpretability capabilities.

4.10.2.1 STL vs MTL

Results showed that MTL consistently outperformed STL independently of the ML algorithm implemented in this analysis. However, we should not assume this is always the case with MTL, as there is the risk of negative transfer between tasks, which may produce the opposite effect of model performance deterioration. The excellent results obtained by MTL in our analysis may be caused by the fact that the sepsis cohorts, although distinct, are sufficiently close, leading to a minimal risk of negative transfer. The major limitation of using STL to model the individual sepsis groups is the low sample size of some of them, such as abdominal and renal/UTI. It is worth mentioning that STL LR models of these two sepsis groups presented significant multi-collinearity issues[199], which are likely due to their smaller sample size. This was particularly critical with the abdominal sepsis group, as all the input variables were deemed linearly dependent by the LR algorithms when the STL strategy was applied.

4.10.2.2 MTL Limitations

MTL is a valuable tool to increase the predictive performance of an algorithm. However, it does come with shortcomings which must be considered. For example, when sharing tasks, the model does come at the risk of having task-specific features contaminating the results, lowering its performance or task-specific individuality feature traits. Thus, MTL can prevent the model's potential of specialising on its specific task due to other tasks altering or even overwriting features it has recorded in its connections to other tasks [200].

4.10.2.3 Explanatory Analysis with MTL

The MTL implementation of GBM and LR shows that the sepsis groups still display distinguished individualities. The number of significant features selected from MTL-LR is far greater than the single-task implementations. Differences in the ranking are present in the MTL vs STL when comparing both approaches. Although slight differences remain among the sepsis groups and feature meaningfulness and hierarchy, the presents of feature contamination may be present. Thus, although a superior predictive performance is achieved, the ability to trust the features selected and inference gains is diminished and uncertain using this approach of MTL.

4.10.2.4 Other MTL Approaches

MTL can be implemented in different forms, and many were reviewed by Zhang & Yang[201] and Thung & Wee[202]. In particular, several strategies could be used to select the data instances in the auxiliary task to reduce the risk of negative transfer. To keep the experiments as simple as possible, we opted for forming the auxiliary dataset using all the data available from other tasks. However, an alternative selection process can be implemented to select the samples from the auxiliary dataset. For instance, in Xu et al[203] and McCabe et al[204], auxiliary samples were chosen according to their similarities to samples in the training subset, while in Sadawi et al[205], a new metric was proposed to select entire datasets that should be included in the auxiliary task. Considering the differences in model performance between STL and MTL obtained in our analysis, we believe there would be little further benefit if a more sophisticated approach was used to form the auxiliary task.

4.10.3 Conclusion

In this research, we proposed the use of an MTL strategy to facilitate ML analysis of several sepsis groups, which can achieve superior predictive performance compared to traditional methods used to model these patients. Our results showed that using a relatively simple form of MTL was sufficient to obtain models that consistently outperformed the STL strategy. Furthermore, we displayed the ability to gain insight from the MTL models developed. Most importantly, we have shown that MTL can be applied and adopted in the ICU environment and further healthcare settings to achieve superior

predictive performance and gain insight from the developed models by leveraging available data to achieve model optimisation.

5 CHAPTER 5: Atrial Fibrillation Detection of Critical Care Patients in the ICU, using the MIMIC-III ICU Database.

5.1 Introduction

Sepsis is a life-threatening heterogeneous syndrome characterised by various clinical features, defined as a life-threatening organ dysfunction caused by a dysregulated host response to infection [206]. Atrial fibrillation (AF) is the most common cardiac arrhythmia among critically ill patients[207], with an exceptionally high incidence among patients with sepsis [208] [209]. Additionally, AF that occurs in the context of suspected infection may indicate acute cardiac dysfunction consistent with sepsis. AF during sepsis may contribute to the development of "post-ICU syndrome" and may represent an opportunity for interventions to improve long-term outcomes following critical illness[210]. Patients who experience newly diagnosed AF during severe sepsis have increased risks for in-hospital stroke and mortality[211][212]. Studies have shown that more than 50% of patients with newly diagnosed AF during severe sepsis do not survive to hospital discharge[213]. Additionally, studies have shown that 44% of septic shock patients were found to have new-onset AF in a medical ICU, with 34% of the patients studied would not having been diagnosed without the Holter ECG monitoring system [214]. Thus, emphasising the need for automated detection of AF in critical care, especially with septic patients.

Automated AF detection for critically ill patients has been studied very little, with most approaches utilising parameters from the ECG signal or rule-based methods to classify the ECGs. Most common AF detectors are developed and validated using the MIT-BIH atrial fibrillation database, where the patients were ambulatory and mostly paroxysmal AF was recorded [215]. However, AF detection in critically ill sepsis patients presents unique challenges, such as non-AF dysrhythmias and ectopic beats, which are familiar sources of false positives. Additionally, the longitudinal ECG waveform data from critically ill ICU patients can potentially suffer from unique noise artefacts [216][217]. In an ICU, the ECG is often severely corrupted by noise and motion artefacts and may drop data due to poor electrode contact with the skin. Both can reduce the diagnosis accuracy. Thus, the automated identification of poor-quality ECG signals is of paramount importance, especially when signal-processing algorithms are used to screen or monitor cardiac conditions [218]. Familiar sources of ECG noise include, among other things, lousy electrode contact, motion artefacts, electromyography noise and baseline wander [219][217]. Different approaches have been developed to detect and remove motion and noise artefacts from ECG signals using various filtering techniques [220][221]. The ECG is a simple and noninvasive procedure to assess the heart's electrical activity. However, arrhythmia diagnoses using a standard 12 lead ECG are complicated because specific arrhythmic beats can occur infrequently. Therefore, the traditional approach has been to perform offline processing of the ECG signal using signal processing algorithms that highlight potential anomaly sections for an ECG technician or physician to review [222].

Nevertheless, in recent years, deep learning (DL) models have been used to solve many problems in vision, sequence and speech tasks, showing a significant performance improvement compared to feature extraction-based methods[223]. The major drawback of these approaches is the black-box nature of the model, which does not allow for clinical interpretation of cardiac arrhythmias. While the black box approach of DL models might be adequate in many use cases, there is a need for interpretable models in sensitive domains such as medicine to understand model competency and potential failure cases. Furthermore, significant progress has been made in providing interpretability to CNN models by understanding the saliency of models through research such as[224] [225], which explores class activation maps and grad-CAM. Adapting these visualisation techniques to time series data such as ECGs would massively facilitate our understanding of DL model decisions behind the classification of the ECGs.

With interpretability in mind, here are the significant contributions of this work. First, we present an automated AF detection algorithm for septic ICU patients. Secondly, we extend this to non-sepsis ECG

cases to test generability. Thirdly, we introduced a novel visualisation pipeline utilising UMAP to explore the ECG embedding in conjunction with the Kmeans SeCo framework to identify ECG clusters. Fourthly, we explore saliency maps concerning the ECGs to understand the rationale behind the predictions. Moreover, the last contribution is that we trained, validated, and tested our algorithm using a large amount of data, using an independent blind test dataset containing ~274 hours of ECG recordings for all analytical approaches.

5.2 Study Aim

The aims of this chapter were; 1. To develop a framework to detect AF ECGs for septic ICU patients. 2. Test our implementation with non-sepsis ECG cases to test our frameworks generalizability. 3. To explore methods to gain inference behind the ECG predictions. This includes dimensionality reduction, clustering and saliency maps.

5.3 Outcome

The primary outcome was AF, coded as a binary variable to indicate whether an ECG contained AF ('1') or Non-AF ('0'). Model performance was measured using the area under the receiver operator characteristic (AUC) curve. AUC means, confidence intervals, or bootstrapping was not approximated due to computational time and the holdout validation method implemented. Cluster purities were evaluated by AF cluster accuracy.

5.4 Data Description

This study used a subset of the MIMIC-III dataset containing 45 subjects corresponding to 57 ECG records, with AF annotations provided by [216]. MIMIC III is an extensive open-source medical record database publicly available from PhysioNet [226]. MIMIC III links continuous ECG waveforms to many patients' time-varying clinical and hemodynamic data. The sampling frequency of the ECG recordings was 125 Hz, and the measuring unit was millivolts (mV). The annotations provided displayed specific times AF and Non-AF were present in the ECGs.

5.4.1 Pre-processing of the ECG Signal

The proposed ECG detection method consists of two phases. Firstly, the discrimination between segments containing ECG waveforms and those that do not. Secondly, motion and noise artefact detection from those data segments containing ECG waveforms. The MIMIC-III waveform data was recorded over long periods. Therefore, there were many practical problems in collecting and handling the data. These issues included missing values, stopped recordings, and bad electrode contact. As a result, there are cases where the ECG record exhibits no ECG for prolonged periods. Thus, these time periods must be identified and omitted from the analysis. To detect whether the ECG records contained analysable waveforms, each of the ECG records was divided into 30-second ECG segments, and the following steps were implemented. First, for each 30-second ECG record, upsampling from 125Hz to 300Hz, removal of any baseline wander, and normalisation between zero and one were applied. These steps were to aid in approximating the QRS complexes in each ECG and aid in the developed model's training. Next, we outline the pre-processing steps taken to determine an analysable ECG.

- Step 1: Detect missing values (denoted by Nan). If the ECG segment contained missing values, that segment was discarded, as ECG should not contain missing time points.
- Step 2: Detect consecutive values. If the ECG segment contained consecutive values for more than 600ms (6 seconds), this ECG was disregarded, as ECG should not have the same continuous values for more than a limited time.

- Step 3: Detect Heart rate in the ECG segment. An ECG segment is classified as analysable if it contains identifiable QRS and RR complexes. If the heart rate of the ECG cannot be approximated, these ECG segments are disregarded (Valid range: 20 <QRS < 80).

- Step 4: Detect uncertain/unidentifiable QRS and RR complexes and intervals. If more than ten uncertain/unidentifiable peaks were present in the ECG segments, these ECG segments were disregarded.

- Step 5: Detect unrealistic values. If the ECG segment metadata contained obscure values such as NA, Nan and Null, that segment was discarded to maintain data purity. Metadata features listed in Supplementary Table 21

These rules were implemented for the inclusion-exclusion criteria for each ECG used during the analysis. Table 15 shows the results from the preprocessing of the ECG signals. In addition, each ECG segment collected metadata containing a range of static values characterising the ECG. The list of metadata was compiled and calculated and is listed in Supplementary Table 21. The idea of the parameters collected in the metadata was to help supplement in determining analysable ECGs.

5.4.2 Clinically Reviewed Non-Sepsis ECGs Validation Data

We selected a small random sample of non-sepsis ECG records from the MIMIC-III database to test our model's generability capabilities. The aim was to test if our trained model was cohort-specific, i.e., for sepsis patients only, or if the model could capture AF ECGs in all ICU environments. As the chosen ECGs have no labels present in the database, we constructed a validation method using three clinicians to assign labels to the selected ECGs. As this is a labour-intensive task, only a modest size cohort was selected for review.

A random subset of 371 patients from the MIMIC-III was used, collectively containing a total count of 603 ECG records. Each ECG record was partitioned into 30-second segments. A random selection of three 30-second ECGs was selected from each record for clinical validation. Each clinician was assigned 1206 ECG segments and assigned 1 of 5 labels: ‘AF’, ‘Possible AF’, ‘NSR’, ‘Other Arrythmia’ and ‘Uninterpretable’. The labels are defined as the best possible representation of that ECG segment. The primary outcome was to assign labels to the ECG segments and review the clinicians’ classification errors. We implemented a 2/3 holdout validation method. Therefore, each ECG segment was reviewed by two clinicians independently. If the labels from the clinicians disagreed, a final senior cardiologist was used as the tiebreaker, with the majority vote taking president. Cases where three clinicians voted for a unique label, were classed as ‘Unknown’. Cases where clinicians voted AF and or Possible AF were aggregated collectively. Cases where multiple clinicians voted ‘Possible AF’ were aggregated with the ‘AF’ class. Cases where at least one clinician voted ‘Uninterpretable’ were taken as the primary class label, as in real-life circumstances, the ECG would likely have been discarded and or regenerated.

5.5 Methods

5.5.1 AF Detection & Classification

5.5.1.1 1D-CNN

Our proposed automated AF detection algorithm consists of several steps. Firstly, preprocessing is performed to discard non-analysable ECGs. Next, we develop a 1D-convolutional neural network (CNN), a deep learning approach from the ML domain.

Convolutional neural networks are mainly composed of two parts, feature extraction and classification. The feature extraction section is responsible for extracting useful features from the ECG signals automatically, while the part of classification oversees classifying signals accurately by making use of the extracted features. The 1D-CNN architecture has two distinct layer types: CNN and fully connected dense layers. The feature learning was proceeded in the CNNs layers by using convolution and sub-sampling (pooling) operations. The specific functions of the two layers are to reduce the complexity

and dimensions of the ECG signal by extracting features of the ECG. These features are then used in the classification task performed by the fully connected layer to calculate the ECG class. Thus, both feature engineering and ECG classification are integrated into one process allowing for a streamlined solution to improve performance [227]. Furthermore, the dimensionality reduction and clustering algorithms will later use these features to further extract insight and potential patterns from the ECG feature representations derived from the 1D-CNN model.

5.5.1.2 Parameter Optimisation

The final architecture and configuration of hyperparameters were obtained after approximately 100 interactions of the hyperparameter tuning. Tuning hyperparameters involves the procedures that determine the network configuration, contributing to providing more precise classification accuracy and better model performance [228]. Hyperparameter tuning is computationally expensive, particularly to test all the possible hyperparameters that may result in an acceptable performance range. In our implementation, we implemented a Bayesian optimisation (BO) approach when tuning the hyperparameters, as using a small number of samples can obtain the values of the optimum hyperparameters for optimisation of the loss function, unlike traditional methods [229]. However, hyperparameter tuning in deep neural networks still remains a bottleneck [230]. This is due to the time required to evaluate the validation error, for even a few hyperparameters can still be computationally expensive. BO is a sophisticated optimisation approach since it combines prior information about the unknown objective function with sample information of the hyperparameters to obtain posterior information of the function distribution. Then based on this posterior information from previous tuning runs, we can deduce where the function obtains the optimal values for the model's hyperparameters [231]. BO is a practical optimisation approach for the automatic configuration and selection of hyperparameters for a surrogate model of some unknown objective function that would otherwise be too expensive to compute. The tuned hyperparameters included the number of CNN blocks (containing multiple CNN layers), dense layers, filters, kernels, pooling, sampling rate, learning rate, batch size, and dropout. Every tuning run was replicated three times to control stability in the network's architecture with respect to the objective function. All tuning iterations were trained using an independent training and validation dataset.

Table 14: Hyperparameters search space for 1D-CNN.

Parameters	Search space
CNN blocks	{1:3}
Number of filters	{16:256}
Kernel size	{2:64}
Max pooling	{2:16}
Dense layers	{0:3}
Dense layer units	{4:256}
Dropout rate	{0:0.5}
Learning rate	{0.1,0.01,0.001,0.00001,0.000001}
Batch size	{32,64,128,256,512}

5.5.1.3 Model Configuration

The initial configurations of hyperparameters are randomly generated with shuffled training and validation data. The tuning hyperparameters were then selected based on the lowest validation error. A detailed list of Hyperparameters and search bounds are shown in Table 14. The proposed ensemble

architecture is a deep CNN architecture to classify AF from Non-AF cases to learn optimal representations of the ECG, which can distinguish both classes. First, we tuned CNN blocks which consisted of two CNN layers with fixed activation functions (Relu), each followed by batch normalisation to minimise overfitting. Next, a max pooling layer is added for dimensionality reduction, followed by dropout. Following the CNN layers, we tuned additional fully connected dense layers to measure the effects on model performance, followed by batch normalisation and dropout. Finally, a binary-cross entropy sigmoid activation function is used as the loss function with a single output unit representing the class probability. An optimisation method of Adam was used throughout. In addition, we applied early stop and reduced learning rate on a plateau to help with model regulation and reduce overfitting.

5.5.1.4 Model Validation & Performance

We applied a holdout validation method to validate our learning model, applying a 60:20:20 training, validation, and test data partitions, respectively. Our primary model performance metric was the area under the receiver operator characteristic (AUC) curve. However, other performance metrics such as sensitivity, specificity, accuracy, precision, recall and F-measures were collected. Finally, the calibration value selected was calculated to balance the sensitivity and specificity metrics based on the validation data partition. Confidence intervals could not be obtained as a cross-validation approach would be too computationally expensive to implement to estimate the uncertainty of the prediction performance.

5.5.1.5 Model Explainability

To provide model explainability, we developed saliency maps for the ECG segments. Saliency maps allow for salient regions of the ECG segment to be highlighted, allowing to visually identify regions in the ECG that most influence the outcome class label. Saliency maps were first introduced by [232], explaining the ideology of computing the gradient of the output class with respect to the input array, therefore informing how the output values change concerning small changes in the input. These gradients are then used to highlight input regions that cause the most change in the output, thus highlighting salient regions of the input array. Our goal is to generate a saliency map for each time step in the ECG signals, where higher values indicate stronger model dependence on the time step towards the outcome class. A good saliency map must accurately highlight the most relevant regions in the time series while remaining intuitive to the end user [233]. In our visualisation, regions of significant influence were highlighted in red, with more concentration reflecting a stronger dependency.

5.5.2 Dimensionality Reduction

We use dimensionality reduction (DR) techniques to provide an alternative view for users to visually analyse and explore the time-series data. The aim is to reduce the feature space to a two-dimensional latent space using dimensionality reduction methods. Several linear and nonlinear dimensionality reduction techniques have been proposed that aim to decrease the number of input features to describe the data in a lower-dimensional space [234]. The data attributes of the features in the lower-dimensional subspace are therefore approximated to the geometric attributes of the data in the original high-dimensional space. Our research applied different linear and nonlinear DR techniques, such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). The objective of using these techniques is to differentiate and visualise the high-dimensional data giving each data point a location in a two-dimensional map. Thus, different perceptions of the ECGs under consideration will be presented, which will help visualise, analyse, and facilitate the exploration of large ECG datasets. To overcome the complexity of the ECG waveform signals, we trained a 1D CNN model to automatically deduce the number of features learned. These features are then used to obtain a 2D visualisation of the univariate time series utilising and comparing PCA and UMAP. Therefore we aim to apply our DR methods to the ECGs feature matrix calculated by the 1D-CNN model.

5.5.2.1 UMAP

Uniform Manifold Approximation and Projection (UMAP) is a recently proposed manifold learning method grounded in Riemannian geometry, algebraic topology, and category theory, which seeks to accurately represent local and global structures [235]. UMAP assumes that data are uniformly distributed on local manifolds in high dimensional space, which can be approximated as a fuzzy set that is patched together to form a topological representation. Next, we can construct a low-dimensional topological representation that minimises the difference between the two representations. UMAP creates a neighbourhood around each individual data coordinate and identifies a pre-selected number of neighbours to build high-dimensional manifolds. Finally, the result is a low-dimensional representation that groups similar data representations together on a local scale while preserving long-range topological connections to more distantly related cases[236]. Therefore, ECG segments that are similarly categorised and grouped in the projection by UMAP should contain similar cardiological similarities in the ECGs.

5.5.2.2 PCA

PCA can be considered the most popular dimensionality reduction technique. It tries to learn the orthogonal projections of the original data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximised [234]. PCA converts a group of correlated variables into a set of uncorrelated variables [237], intending to extract the most essential information. Most commonly, PCA is used for exploratory analysis but due to the nature of its examination of the data relationship between features, it can be used for dimensionality reduction.

PCA builds a dxk-dimensional transformation matrix W that maps the original d-dimensional space X to a new k-dimensional space Y $(k \leq d)$ [237]. The linear eigendecomposition method is applied to the covariance matrix ($X.X^T$) to produce the eigenvector (PCs) and eigenvalues. The eigenvectors show the data directions, and the eigenvalues the data magnitude. The eigenvalues are used to order the columns in the matrix W, where each column is an eigenvector. The eigendecomposition method can be defined as follows: $X.X^T \rightarrow B.D.B^T$, where the covariance matrix is decomposed into three other matrices [238]. Here, B is a square matrix (d x d) that contains the eigenvectors; D is the diagonal matrix (d x d) where all the elements except the main diagonal elements are zeros, and the diagonal elements are respective Eigenvalues; B^T is the transposed matrix.

5.5.3 Clustering

We applied cluster analysis on the projections calculated by UMAP and PCA. The data UMAP and PCA utilised were from the 1D-CNN feature matrix representation of the ECG. We next compared results regarding cluster purity and accuracy. Finally, we investigated the DR projection clusters to discover ECG clusters with similar properties or classes. Cluster analysis is an unsupervised learning method widely applied for dividing objects into different groups based on their similarity [239]. K-means is a well-known clustering method to minimise the Euclidean distance between each data point and the cluster's centre to which it belongs. It is essential to distinguish two objectives when clustering with k-means: firstly, the selection of an appropriate value of 'K', and second, that the 'K' value selected results in the selection of a stable, reproducible solution. With these objectives in mind, we implemented the separation and concordance (SeCo) framework proposed by [240], which allowed for the optimal 'K' value and stability to be selected. The SeCo map generated allows for a 2-D map of the local minima found by the k-means algorithm. Not only can values k be identified for analysis, but also a measure of the stability of the clusters are calculated. This study investigated a range of 'K' values from 2 to 100, with 400 initialisations.

5.5.4 Hardware & Software Requirements

To process and train such a deep, sophisticated model with such a large dataset requires a particular computational effort. Therefore, this experimental research is being performed on an AMD Ryzen Threadripper 3970x 32-Core CPU, a RAM of 256 GB, and a hard drive of 2TB SSD, with an additional

14TB mechanical hard drive for storage. In addition, two Nvidia GeForce RTX 3090 24 GB VRAM graphic cards were used during training. The use of the GPUs was needed because of the high computational time due to the volume of the ECG data. The experiments were performed on an Ubuntu operating system using Python (V3.8.8) and R (V4.1.1).

5.6 Results

5.6.1 ECG Segmentation and Selection

From 57 long lead ECG records relating to 214,774 30-second ECG recordings, 144,686 were selected, totalling a recording time of 1205.7 hours, as displayed in Table 15. Collectively 32.6% of all ECG records were omitted. In addition, 7.9% of the ECGs contained missing values, 11.4% contained non-analysable heart rates, 13.5% held consecutive values, 8.6% included unreliable QRS intervals, and 12.4% had calculated ECG metadata which generated null/missing values.

Table 15: Detected number of segments in labelled data.

Records	Total ECG segments	ECG selected	Total AF ECGs	AF ECGs selected	Total time (h)	Selected time (h)
57	214,774	144,686	46,929	31,644	1,789.9	1,205.7

5.6.2 Clinically Reviewed Non-Sepsis ECGs Validation Data

A total of 1809 30-second ECG segments were collected from non-sepsis patients and reviewed. Only 1,147 (63.41%) of the ECGs had a complete agreement among the 5 class labels. Thus, over one-third of the ECGs needed additional review. However, when the 'AF' and 'Possible AF' diagnoses were aggregated into one group, agreement increased to 84.3%. Of the 1809 ECG snippets, 257 (14.21%) were classified as 'AF/Possible AF', 1132 (62.58%) as 'Normal Sinus Rhythm'(NSR), 217 (12%) as 'Uninterpretable', 169 (9.34%) as 'Other Arrythmia' and 34 (1.88%) as 'Unknown' as displayed in Figure 28.

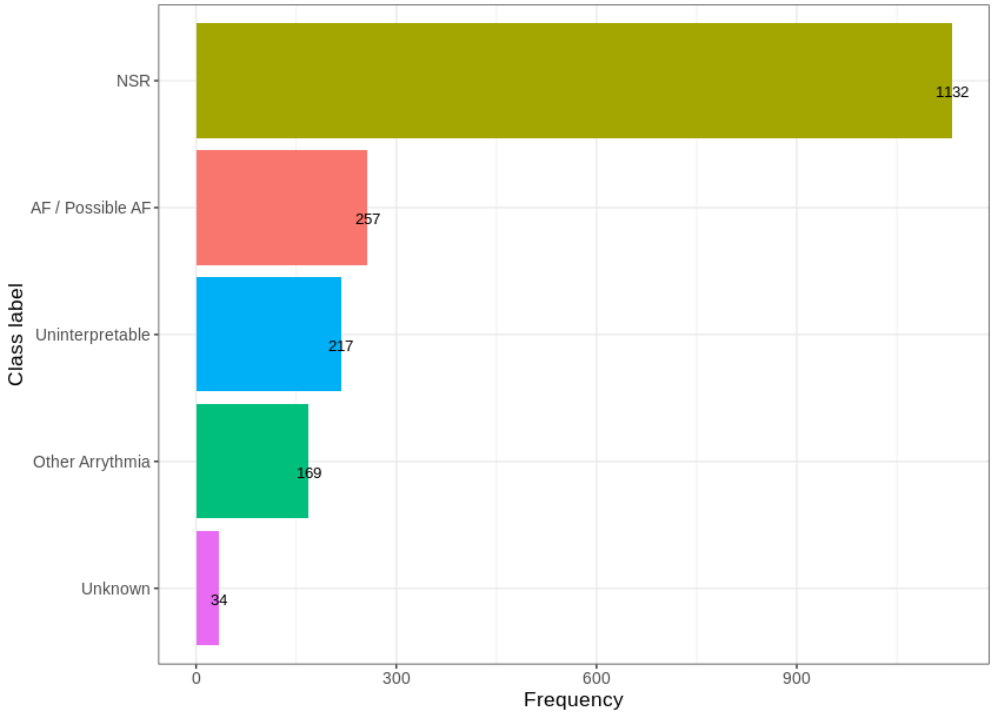


Figure 28: Final class labels for all 1809 ECG segments.

Table 16 displays the confusion matrix of the manually validated ECGs from the three clinicians. Table 17 displays the results from when the fourth final senior cardiologist was needed to decide the final outcome label. The results show a range of agreements and disagreements among the clinicians regarding the five classes. This further emphasises the complexity of the task displaying its subjective nature concerning the outcome labels provided.

Table 16: Summary of confusion matrixes for all clinicians for each validation partition.

		MD3				
	Data Partition 1	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
MD1	AF	28	0	3	2	1
	Possible AF	38	3	37	5	1
	NSR	9	9	293	19	4
	Other Arrhythmia	5	3	56	16	1
	Uninterpretable	24	3	27	4	12
		MD1				
	Data Partition 2	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
MD2	AF	20	18	41	7	13
	Possible AF	1	3	8	2	6
	NSR	16	37	258	27	40
	Other Arrhythmia	6	3	54	15	12
	Uninterpretable	0	2	10	1	2
		MD2				
	Data Partition 3	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
MD3	AF	80	3	3	9	0
	Possible AF	12	1	1	5	1
	NSR	6	9	325	32	13
	Other Arrhythmia	4	1	5	84	1
	Uninterpretable	0	1	0	0	7

Table 17: Summary of confusion matrixes for all clinicians for each validation split for ECG segments without complete agreement 662 ECGs.

		MD1				
	Data Partition 1	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
MD4	AF	0	7	1	1	4
	Possible AF	1	26	7	3	14
	NSR	4	41	21	52	25
	Other Arrhythmia	0	3	9	3	5
	Uninterpretable	1	4	3	6	10
		MD3				
MD4	Data Partition 1	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
	AF	12	1	0	0	0

	Possible AF	46	2	0	3	0
	NSR	8	5	113	12	5
	Other Arrhythmia	1	1	4	14	0
	Uninterpretable	9	6	6	1	2
MD1						
MD4	Data Partition 2	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
	AF	0	3	8	2	5
	Possible AF	0	13	30	5	7
	NSR	19	40	34	28	46
	Other Arrhythmia	4	2	26	2	4
	Uninterpretable	0	2	16	0	9
MD2						
MD4	Data Partition 2	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
	AF	16	0	0	2	0
	Possible AF	46	3	0	5	1
	NSR	11	9	120	25	2
	Other Arrhythmia	0	2	0	35	1
	Uninterpretable	7	3	0	8	9
MD2						
MD4	Data Partition 3	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
	AF	0	0	0	0	0
	Possible AF	11	3	1	10	1
	NSR	8	8	8	21	5
	Other Arrhythmia	1	0	0	9	0
	Uninterpretable	2	3	0	6	9
MD3						
MD4	Data Partition 3	AF	Possible AF	NSR	Other Arrhythmia	Uninterpretable
	AF	0	0	0	0	0
	Possible AF	10	10	5	1	0
	NSR	2	5	35	8	0
	Other Arrhythmia	0	1	9	0	0
	Uninterpretable	3	3	11	2	1

5.6.3 Evaluation of Model Performance

The developed model predicted AF vs non-AF cases with AUC performances of 99.99%, 99.969%, and 99.967% regarding each training validation partition, as displayed in Figure 29. The probabilistic threshold was calibrated using the validation set, obtaining a class threshold value of 0.81. The F1, precision, and recall measures highlight that the model favoured the non-AF class as it displayed slightly higher performance values across all training validation partitions. This may highlight that the non-AF class may be more accessible to capture distinct patterns during model learning.

Metrics	Training	Validation	Test
AUC *	99.99	99.969	99.967
Accuracy	99.506	99.174	99.219
Sensitivity	99.509	99.184	99.284
Specificity	99.493	99.138	98.979
F1 (Non-AF):	99.684	99.471	99.503
F1 (AF):	98.863	98.113	98.183
Precision (Non-AF):	99.509	99.184	99.284
Precision (AF):	99.493	99.138	98.979
Recall (Non-AF):	99.86	99.76	99.722
Recall (AF):	98.24	97.109	97.4

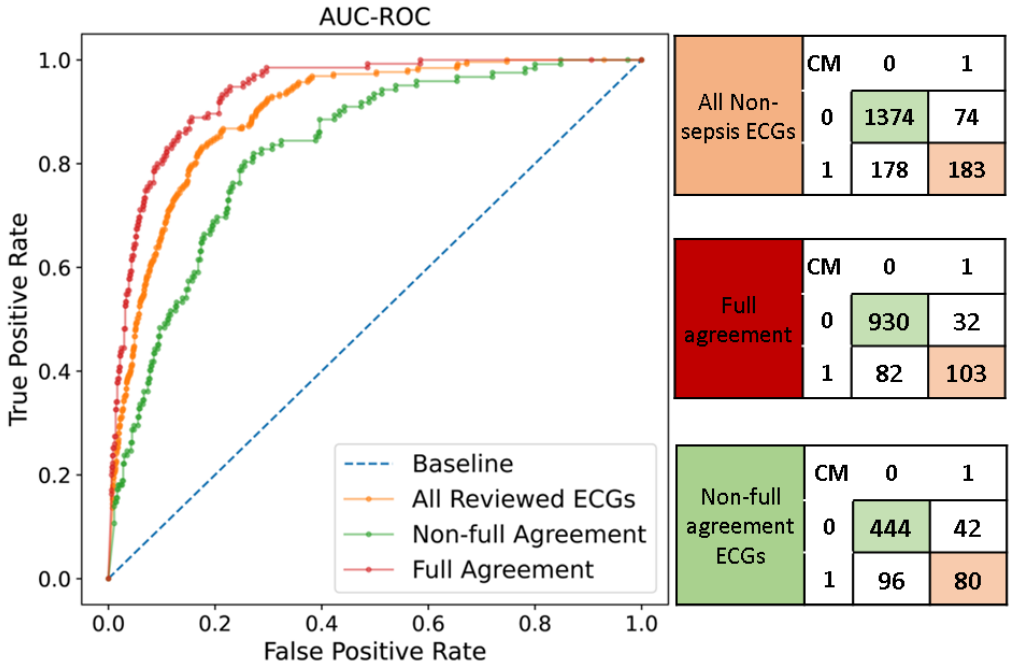
Training	CM	0	1
	0	67739	334
	1	95	18643

Validation	CM	0	1
	0	22485	185
	1	54	6213

Test	CM	0	1
	0	22606	163
	1	63	6106

Figure 29: AUC-ROC scores and confusion matrix for the model's validation splits. Training split (Gray), validation split (yellow) and test (orange).

The non-sepsis ECG data showed decreased performance, obtaining an AUC of 89.754 %. Similarly, the negative class (non-AF) was superior to the positive class (AF) regarding correct classification. The results displayed that the model struggled to classify positive cases correctly from the non-sepsis dataset. This resulted in many false positives being captured. However, the model could still capture non-AF cases efficiently, achieving an F1 and precision values of 91.6% and 94.89%. When we reviewed the ECG cases for which the clinicians had complete agreement, the AUC score of the model increased to 93.59%. Similarly, when focusing on ECG cases where a disagreement occurred that needed the third clinician to tie break, the AUC decreased significantly to 82.61%. This may reflect the complexity of the labelling task, as when the labels are easily agreed upon, the AUC score increases.



Cohorts	AUC	Accuracy	Sensitivity	Specificity	F1 (non-AF)	F1 (AF)	Precision (non-AF)	Precision (AF)	Recall (non-AF)	Recall (AF)
All non-sepsis ECG	89.75	86.07	94.89	50.69	91.60	59.223	94.89	50.69	88.53	71.21
Full Agreement	93.59	90.06	96.67	55.69	94.23	64.38	96.67	55.68	91.90	76.30
Non Agreement	82.61	79.1541	91.358	45.4545	86.55	53.69	91.36	45.45	82.22	65.57

Figure 30: AUC-ROC curve for the non-sepsis data. All none-sepsis ECGs (orange), full agreement ECGs(red) and non-full agreement ECGs (green). The baseline was represented with dases lines (Blue).

5.6.4 Evaluation of Hyperparameter Tuning

The optimal model architecture after hyperparameter tuning is constructed from 2 CNN blocks (4 CNN layers with batch normalisation and pooling) connected to a single dense neuron (output layer), no additional dense layers were selected. The chosen architecture reduces the filters per CNN block, resulting in a one-dimensional feature space of 2,944 for each ECG segment. The optimal batch size and learning rates were 512 and 0.0001, respectively, with an epoch run time of 720 seconds for 25 epochs.

5.6.5 Evaluation of ECG Projections

We ran UMAP and PCA simultaneously using 2944 features calculated from the 1D-CNN's convolutional filters. We trained each unsupervised model on the training data and then predicted the validation cases based on the training representations learnt for each model. PCA and UMAP successfully pooled together clusters based on the true class label of each ECG segment. PCA displayed a single high concentration of the AF class centred at the origin of the first two principal components. In contrast, non-AF cases were more sparsely projected across the rest of the latent space. UMAP displayed a similar result. However, the region in which AF ECG segments were grouped was much more extensive and purer than PCA, representing a larger region of the AF in the latent space formulated.

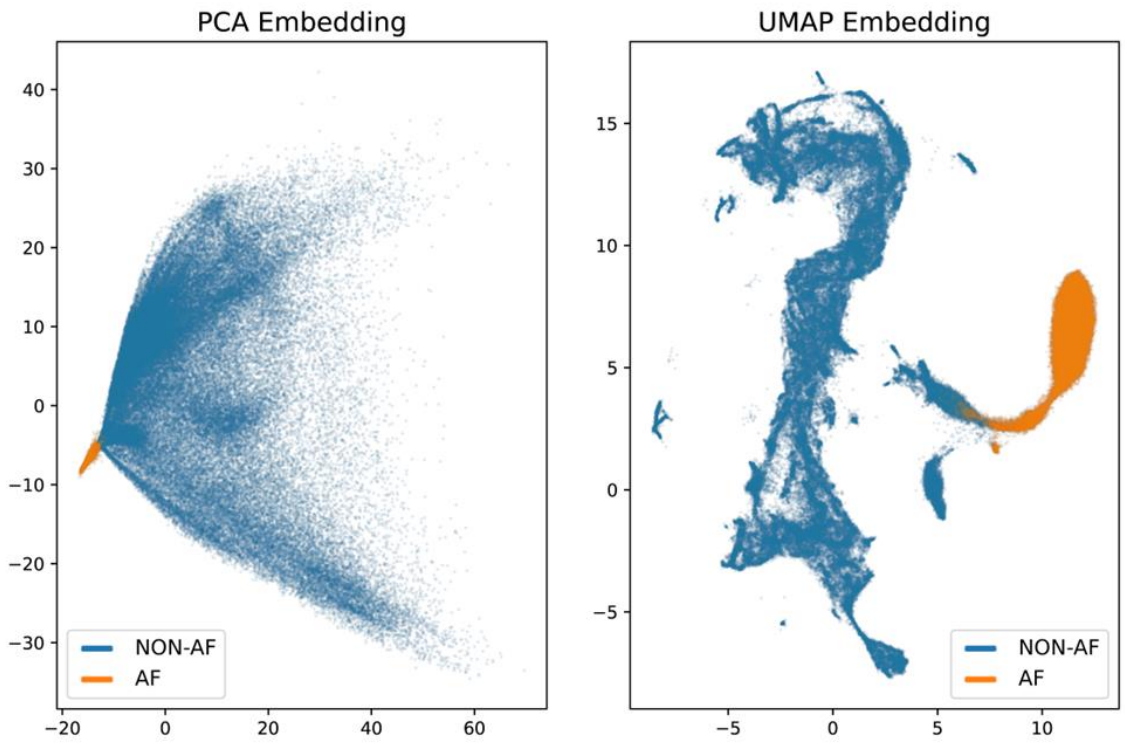


Figure 31: PCA and UMAP projection of all ECG segments displaying true class labels.

The UMAP embedding showed a distinct decision region in which the probability of the ECG becomes confident. Dissimilarly, from visual inspection, PCA failed to highlight such regions where the model becomes less confident in its predictions. Nonetheless, while the ECG segments were not always segregated into completely distinct clusters by UMAP or PCA, the distinction of both classes remains similarly identifiable in both DR methods. However, the UMAP technique surpasses PCA in terms of the identifiable latent space of the AF class. Additionally, it displays a more suitable representation of the ECGs in a lower dimensional space as we can distinctly view the decision boundary.

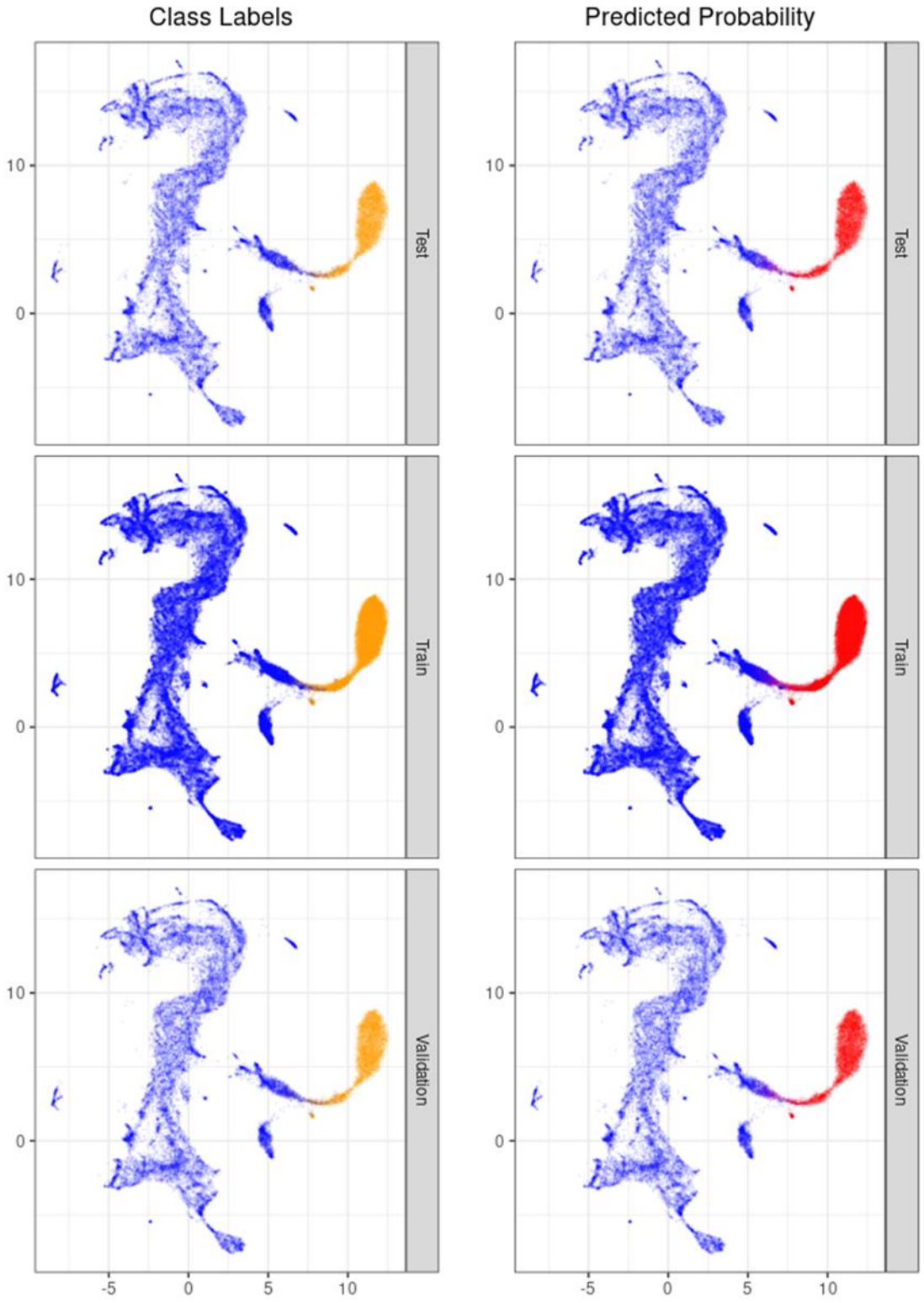


Figure 32: UMAP projections of training, validation, and test split of the data. Blue (True class label | non-AF), orange (true class label), Red (probability of model prediction)

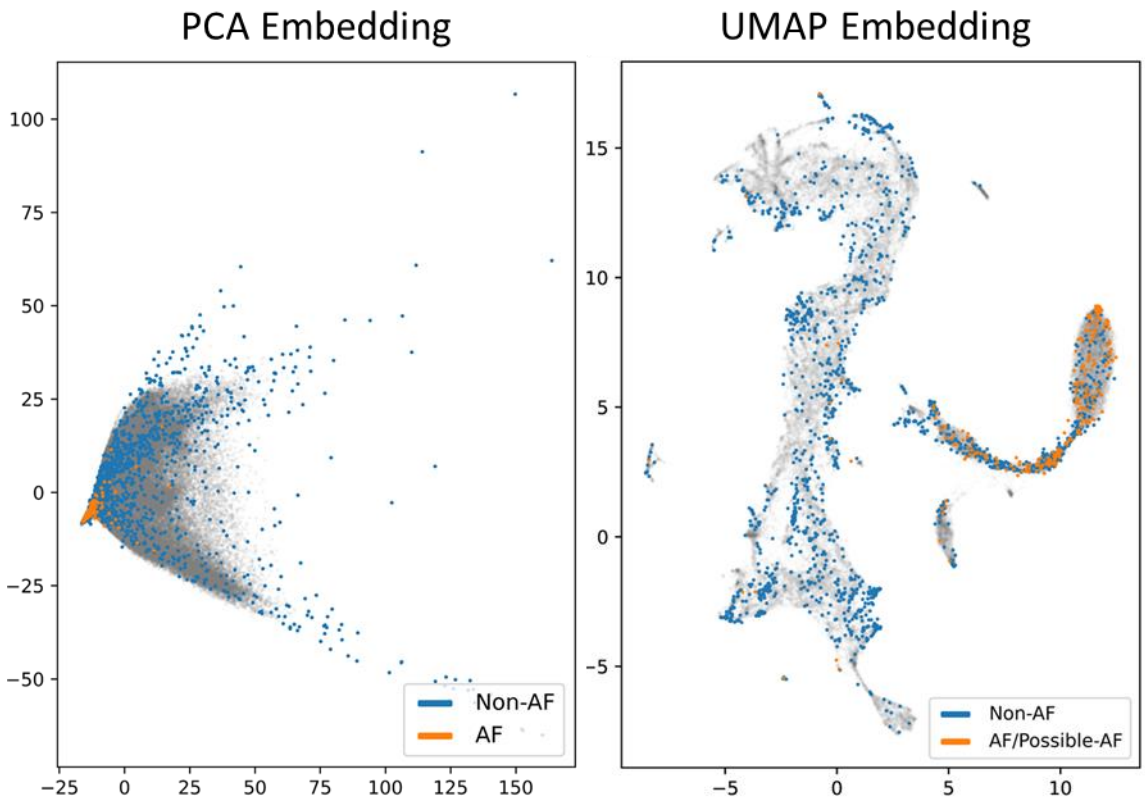


Figure 33:PCA and UMAP embedding of non-sepsis ECG data.

Figure 33 displays the projections of the non-sepsis ECG data with aggregated labels provided and validated by our clinical team. The results of the projections are not as distinguishable compared to the original sepsis ECG projections. Many AF/possible AF classes are more spread across both latent spaces. Additionally, the Non-AF class shows to have more overlapping in regions thought to have strong similarities and relations with the AF class.

5.6.6 Cluster Analysis

We applied K-means to the CNN features in conjunction with the SeCo framework for each DR method. The results demonstrated the optimal 'K' value for K-means, which generates stable, reproducible clusters, where 'K' values 9 and 10 for UMAP and PCA. The SeCo figures can be viewed in supplementary material Figure 35. Furthermore, Table 18 demonstrates the cluster purity regarding the classification of AF ECG cases compared to the non-AF ECG cases. The results showed that UMAP could classify AF ECG segments more accurately than the PCA method. Furthermore, the results displayed by UMAP categorised nearly all training, validation, and test cases for AF ECG segments into three distinguishable clusters (clusters; 3, 7, 9), as displayed in Figure 34, unlike PCA.

Table 18: PCA and UMAP cluster purities for the sepsis and none sepsis ECG data.

Training/Validation/Test ECGs				Non-Sepsis ECGs			
Clusters-PCA	ECGs (n)	AF ECGs (n)	AF ECGs (%)	Clusters-PCA	ECGs (n)	AF ECGs (n)	AF ECGs (%)
1	48206	31644	65.64	1	928	248	26.72
2	8997	0	0	2	36	0	0
3	23981	0	0	3	131	3	2.29
4	7003	0	0	4	103	0	0
5	18328	0	0	5	18	0	0
6	11321	0	0	6	287	5	1.74
7	10023	0	0	7	48	0	0
8	4905	0	0	8	185	1	0.54
9	4405	0	0	9	65	0	0
10	7517	0	0	10	8	0	0
Clusters-UMAP	ECGs (n)	AF ECGs (n)	AF ECGs (%)	Clusters-UMAP	ECGs (n)	AF ECGs (n)	AF ECGs (%)
1	9947	0	0	1	215	2	0.93
2	17864	0	0	2	171	1	0.58
3	25283	25282	99.99	3	193	110	56.99
4	13117	1	0.007	4	186	2	1.08
5	17105	0	0	5	243	4	1.65
6	21503	0	0	6	150	4	2.67
7	14694	312	2.12	7	256	42	16.41
8	18449	0	0	8	113	1	0.88
9	6724	6049	89.96	9	282	91	32.27

The non-sepsis ECGs labelled by the clinicians yielded similar results as the initial training and validation data. However, the cluster purities of the other corresponding groups did not reflect the same performance. For example, the collected total of ECG segments in clusters 3,7,9 still reflected 94.55% of AF ECG segments of the non-sepsis ECG data. However, the collective total of ECG segments represented in the cluster was 731(40.31%) ECGs, of which AF represented 243 of the whole group clusters (33.32%).

5.6.7 Understanding the Visualisation

	Record	ECG Index	Prob	Pred Class	True Class	Cluster	BPM
A	p097547-2125-10-21-23-43	112	0	0	0	8	89
B	p087675-2104-12-05-03-53	376	0	0	0	4	81
C	p058932-2120-10-13-23-15	3130	0	0	0	6	67
D	p099674-2105-06-12-21-00	7	0.22	0	0	7	87
E	p075796-2198-07-25-23-40	488	0.997	1	1	9	87
F	p063039-2157-03-29-13-35	1066	1	1	1	3	72

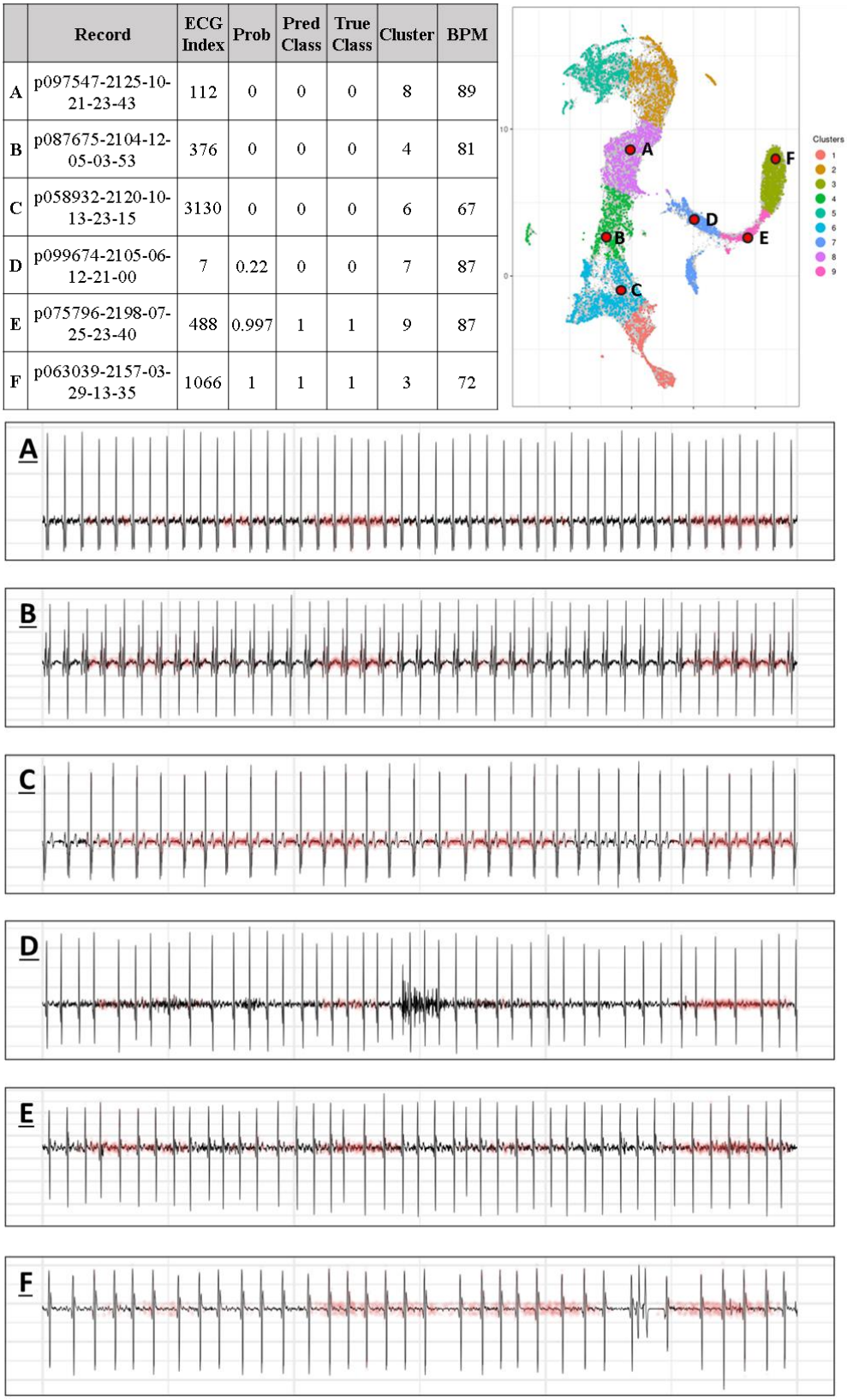


Figure 34: Six randomly selected saliency maps from the test set ECGs cases. The upper right figure displays the kmeans labels projected over the UMAP projects. The table categorises each of the selected ECGs, the class label allocated, and the associated metadata.

Saliency maps were calculated for all ECG segments regarding the outcome class labels AF and non-AF, as displayed in Figure 34. We visualise the saliency maps for six randomly selected ECG segments classified correctly regarding both classes. The saliency maps encode the location of the region with the most impact on the outcome class with respect to the input ECG and thus can be used for object localisation in terms of relative importance. The first four ECG cases are non-AF ECG segments labelled correctly by the CNN model. The model associates high importance with the apparent P-wave regions of the ECG. These patterns and features in the ECG waveforms were commonly revealed as important for classifying non-AF ECG cases. ECGs E and F display a correctly classified AF ECG. The CNN developed displays a high association to regions of extra P-waves or 'recovery beats' after the QRS interval. After further investigation, we discovered that many misclassified cases were due to the ECG belonging to other arrhythmias or contained noise after further clinical review. The saliency maps produced gave insights into how the model is deducing important regions of the ECG, thus allowing a level of interpretability into the rationale of each classified ECG segment.

Furthermore, this could be exploited further in a post hoc review of model implementation to tailor future tuning and ML pipelines. Figure 34 displays six randomly selected ECGs labelled A to F. The ECGs represent different regions of the UMAP projection segmented by the kmeans groups. The upper left table displays the records and ECG information for the representative ECGs, displaying the ECG class label, probability, and cluster assigned. The upper right figure shows the test set UMAP projection with the Kmeans cluster labels. It is clear from previous analysis that the AF ECG segments predominantly reside on the right-hand side of the UMAP projection. ECG B displays a Non-AF ECG, however, after clinical review, it resembles a patient ECG with a pacemaker. In further investigation, more ECGs in this cluster have similar characteristics. Therefore, ECG cluster four may capture ECG records with pacemakers or cardiac pacing devices. ECG C is likened to a normal sinus rhythm. ECG D has a slight increase in probability from 0 to 0.22 compared to ECGs A, B, and C. Although in all these cases, the ECG class was classified correctly as non-AF. However, ECG D contains some noise or artefacts, which may explain the increase in probability and potentially the misclassification error among some of the test ECG and Non-Sepsis ECG cases. Furthermore, this ECG is hard to disentangle visually and would likely be disregarded in real-world applications. ECGs E and F were correctly classified as AF with a probability of 0.98 and 1. The Saliency map displays impactful regions after the QRS, predominantly the recovery beat or additional F wave. The clusters assigned to these groups also had similar resemblances and patterns, again further emphasising the distinct regions in the UMAP projection categories' specific ECG characteristics. Further investigation could lead to clinically meaningful groupings of the clustered ECGs.

5.7 Discussion

We describe an approach to detect AF in ICU patients and a novel way of visualising that data using dimensionality reduction techniques and saliency maps to investigate each ECG segment further. To be useful clinically, AF detection algorithms must be accurate to avoid false positives to reduce alert fatigue. Therefore, one major challenge facing our model was the generability capability of applying our trained model to non-sepsis data to detect AF ECG segments. The ability to detect non-AF ECG cases was still accurate. However, the false positive rate increased considerably. There are many highlights of this research. First, we developed a framework that can classify AF and Non-AF using the near raw waveform data in sepsis patients with high performance. Secondly, we explored ways of visualising the ECG waveforms using dimensionality reduction techniques for a more in-depth understanding of the ECGs in a latent space. Thirdly, we explored clustering of the ECG projections to explore cluster purities with respect to the AF ECG class. Fourth, we explored and deployed saliency maps to gain visual inference of the ECGs, allowing for a level of reasoning behind each ECG classification. Lastly, we tested our model using non-sepsis ECG data from a range of ICU patients to investigate the scalability and generability of the model, which was initially trained on a relatively small cohort size of 47 patients using 57 ICU ECG recordings.

5.7.1 Clinically Validated Non-Sepsis ECGs

The none sepsis clinically validated ECG records displayed curious results. The challenge of Diagnosing AF in short 30-second single lead ECG snippets remained difficult in over 15% of cases and may explain why brief episodes of AF (<30 seconds) often remain undiagnosed. Over 35% of the 1809 ECGs for clinical validation needed the third clinician to determine the final class label, with a further 34 cases being unclassified. This result shows the current practice's complicated and subjective nature of the ECG classification. This complexity is reflected in the modelling objective, where many misclassified cases were due to either the problematic reading or understanding of the ECG or because the ECG was mislabelled. In addition, over 12% of the ECG snippets in our labelling task had inadequate quality for the diagnosis. The quality of the ECG is a critical part of the study, as it played a pivotal role in the classification of each ECG segment. Classical monitoring devices such as defibrillators, novel patch-based devices, or lifestyle gadgets commonly offer single-lead ECG monitoring[241]. However, interference, patient movement and shivering may impact ECG quality and hamper readability. Therefore, diagnostic algorithms based on ML are required and tasked to address these diagnostic uncertainties. Research has highlighted the need for an automated approach to label ECG cases in order to reduce the workload associated with the manual review of ECGs [242]. The approach presented here could be used to reduce the time required to review ECGs in the ICU manually. Reviewing an ECG case can take approximately 20 seconds [243]. However, when coupled with a large number of ECGs requiring review can be an extensive, time-consuming task. Therefore, an approach to reduce the number of manual ECG reviews could significantly reduce the cost and resources of screening.

5.7.2 Modelling and Classifying AF ECGs

ML models have become very accurate in detecting AF, most of them exhibiting accuracies higher than 90%. Several models are designed to detect AF only, but others also identify other arrhythmias. The data usually involves the use of ECGs, either a single or 12 leads. However, some methods use different modalities, such as ballistocardiograms (BCG), photoplethysmogram (PPG), tabular data extracted from EHR or features from the ECG, or combinations of them all[244]. A consideration which was considered during the analysis was whether transforming the data is necessary or valuable before applying ML or if it is possible to use the raw data as inputs to the model. This decision could heavily contribute to the conclusion of what ML algorithm should be implemented. Our study used the CNN model architecture to determine the features and classification of the almost raw ECG data by applying little augmentation to each ECG segment. However, a considerable amount of time was spent determining an ‘analysable’ ECG using an automated, rule-based approach, as detailed in the methods section. This approach aimed to distinguish noise sources and contamination in the long-term continuous ECG recording obtained from the mimic-III database. However, we found various sources of artefacts which led to poor signal quality and portions that did not even contain ECG waveforms. Most likely due to bad contact with electrodes, movement artefacts, the sudden disappearance of the ECG tracing, high-frequency noise, or poor device calibration. Due to these various types of noise artefacts, we had to develop an approach to discriminate between segments with and without ECG waveforms.

We found that 67.4 % of the data was analysable ECG recordings from our initial labelled data. The primary goal of the noise detection was to reduce the false positive rate of AF, which is why we were mainly interested in retaining only the analysable actual ECG records. This led to the development of an analysis of the QRS and RR intervals to negate any corrupted ECG signals. Thus, we analysed all ECG segments for features listed in Supplementary Table 21. If none of these expected features, such as heart rate or QRS intervals, could not be calculated in the ECG, then the ECG was determined as noise and disregarded, as these ECGs did not resemble any characteristics of expected ECG waveforms, thus negated from the analysis. A similar strategy has worked well with the MIT-BIH [245][246]and MIMC-III [208], so it is hoped that similar success with noise detection will occur in the future using similar strategies. Our rule-based strategy showed promising results in identifying various types of noise artefacts, albeit this study was performed on a limited number of subjects.

Furthermore, this preprocessing is primarily driven by the fact that we were mainly interested in the automatic identification of AF ECG segments. Hence the development of our noise detection strategy was to aid this objective by identifying and negating artefacts in the data to reduce the AF false positive rate. Further validation of the proposed preprocessing strategy will need to be undertaken using more subjects in the future. Still, the results from our current implementation showed promise as our approach accounted for most of the artefacts in the data encountered in the ICU setting from our subset.

5.7.3 Dimensionality Reduction Methods

The results of applying PCA and UMAP to the features extracted from the ECG signals are shown in Figure 31. The UMAP projection offered better distinctions between AF and Non-AF, further explored with the cluster analysis showing similar results. For example, PCA predominantly grouped all AF cases towards the centre of the origin, compared to UMAP, which displayed more spread concerning the AF class. The UMAP approach also displayed the decision boundary clearly, as the ECG cases tended to resign more towards the righthand side of the embedding, with the likelihood of the ECG class being AF increased where it was not as apparent with PCA. This was further seen with the clustering analysis. However, the UMAP embedding features space still shows some overlap between AF and Non-AF, considerably less than PCA. The overlap between the two groups could also stem from miss-labelled ECGs or ECGs containing motions and noise artefacts. Overall, the projections and embedding of the non-sepsis ECG cases resulted in similar results. However, it displayed more significant overlap and unexpected projections resulting in lesser purity in regions associated with AF.

5.7.4 ECG Visualisations

The results of the visualisation of the CNN model are highlighted in Figure 34. The visualisations of the critical regions of the ECG are shown by the network represented as red highlighted regions of the ECG. The region highlighted denoted the importance of a given time point in predicting the target class. The following observations can be made by studying the plots for each class. First, t visualisations were observed to align with the literature in the ECG interpretation [247]. The peaks are observed to have a crucial role in this classification task, similar to how ECG technicians infer the significance of rhythmics to diagnose arrhythmia. It can also be noted that for non-AF, the CNN visualisation provides importance to the P and T waves. By observing the visualisations of the CNN network for the AF class, it can be seen that the network gives importance to the R peak followed by fibrillation or extra recovery waves. Finally, t visualisation was observed to line up with the clinical literature in ECG interpretation[248]. Our future work will be focused on expanding the visualisations to other classes and validation in the clinical setting. Collectively, we can see that using saliency maps can aid in determining the rationale of the CNN model concerning each outcome class and has been well documented in the literature to help aid in understanding ECG records[249], [250]. Therefore, with further investigation, we aim to dissect and investigate the patterns presented in each class to align further clinical literature with the presented trends and patterns shown in the CNN model.

5.7.5 Cluster Analysis

The behaviour of the clustering algorithm implemented depends essentially on the features constructed from the CNN model. The kmeans algorithm and SeCo map were calculated on the embedding feature space created by the UMAP algorithm. The aim was to partition the latent space into meaningful clusters. The results have shown that the groups resembled ECGs of different overall characteristics. This was prevalent with cluster four, which displayed many pacemaker cases after a post hoc review of the groups. A similar result was shown by clusters three and nine, which primarily captured AF cases. In future research, we aim to determine meaningful clusters from the embedding space created by UMAP.

5.7.6 Study Limitations

Although we have used a large number of ECG segments to classify AF, this is still a single-centre study, and data may differ at other centres. Furthermore, the data comes from very few patients and

therefore does not integrate a diverse set of patient characteristics like the datasets used by [251] [252], which are shown to be very generalisable due to the large datasets used to train the models. Therefore, further validation of the proposed automated AF detection algorithm in different settings is undoubtedly warranted. Given the large number of subjects and long duration of recordings available in the complete MIMIC-III database, we have been limited to developing our algorithm from only a tiny fraction of the data due to the nature of the available labelled AF ECG records. In addition, the limitation of ECG record labels and the manually extensive task of clinically validating ECGs resulted in the relatively small non-sepsis ECG data set, as manually annotating thousands of hours of raw ECG data is physically unfeasible. Hence, due to these limitations, it is possible that we may not have accounted for other types of noise sources or artefacts in the data or, additionally, AF or non-AF unique global characteristics. We hope this work can be expanded and further developed to provide a solution to automatically label all ECG records for the entire MIMIC-III database, allowing for further analysis of AF ECG waveforms in critically ill patients on a much larger scale. Indeed, this strategy has worked well with the original sepsis ECG data. However, further development is required to extend this to all ECG cases, i.e., non-sepsis AF ECG records. Lastly, when partitioning the data into training validation and test partitions where unique patients were in each representative dataset, there may be correlations across ECGs from the same patient, which we have not accounted for in our study. This may have an impact on generability and predictive performance.

5.8 Conclusion

In conclusion, we have proposed several tools for addressing the interpretability issue of ML ECG classification. First, we propose a method to detect and negate noise and motion artefacts in the ECG data. Next, we built a 1D CNN model based on state-of-the-art architecture and explored dimensionality reduction techniques to explore the CNN model features in a latent space for clusters and natural groupings. Furthermore, we utilised Saliency maps to visualise what areas of each beat of the ECG record are important to the final classification decision. Additionally, we have shown how the raw ECG record can be visualised in a latent space to allow for easy consumption and identification of the ECG likely class label based on its geographical location in the 2D latent space embedding created by the UMAP algorithm. Lastly, we segmented each saliency map into clusters to quantitatively compare the areas the model looks at for each cluster and investigate any differences amongst the ECGs. Model interpretability is necessary when applying these models to clinical applications. Therefore, the ability to understand the rationale behind the model's prediction is paramount.

Moreover, we tested our model with non-sepsis data to check generability and achieved adequate results. However, further development is needed to utilise this model and framework for non-sepsis cases. Finally, our method provides a technique for comparing which section of an ECG is important for each class assigned. Comparing ECG saliency maps allows us to determine the patterns picked up by the model for each class as a whole, allowing us to compare this to clinical understanding and investigate any outliers. Since new-onset AF during sepsis can significantly increase complications and risk of poor outcomes, this study can help expand research into AF during sepsis by allowing rapid scaling up of AD detection in banked ECG waveform data and improving clinical decision of AF among the critically ill [216]. Overall, we feel that this would lead to an improvement in clinical care for patients with AF. In addition, further studies can explore whether enhanced AF detection may improve patient outcomes by informing more rapid and better-informed treatment. Another avenue of research could explore whether different clustering techniques may enhance cluster purity and improve AF classification, improving patient outcomes by informing more rapid and better-informed treatment.

6 CHAPTER 6: Discussion

6.1 Conclusion

This research has shown various methods and techniques to understand and infer insight when modelling ICU patients in an ML environment. This research has covered several ICU data sources and their advantages and limitations. In addition, we described the process of data cleaning, feature engineering, imputation, and missing value methods, amongst other traditional data pipelining operations. The ICU environment has been shown to be diverse and data-rich with various data types and structures, allowing various modelling tasks to be investigated. We covered several classification and regression methods with popular ML models like linear and logistic regression, RF and GBM. We explored different inference methods to obtain insight from the model, such as odds ratios, sankey diagrams, feature importance, partial dependencies, and SHAP analysis. We further investigated model optimisation frameworks to increase predictive performance using the piling MTL method. Lastly, we explored waveform analysis of ECGs to detect AF in 30-second ECG segments. This research was then extended using dimensionality reduction techniques and clustering in conjunction with saliency maps to provide a rationale behind the ECG classification. Collectively throughout the research undertaken, we have provided methods of gaining a level of interpretability from the developed models to provide insight clinically. Furthermore, we have demonstrated ML's ability to outperform current methods of severity quantification in the ICU and its use case applicability.

The overall aim of this thesis was to leverage model predictive performance and interpretability when modelling various outcomes in the ICU. Additionally, we aim to display the use case practicality of ML in the ICU, highlighting how ML can improve and optimise current medical understanding of outcomes of interest and current practices. The need for interpretable models in the clinical setting is paramount in understanding the rationale behind a particular prediction. Within this thesis, we build tools for clinical implementation for the prediction of mortality and LOS for general hospital and ICU settings. Furthermore, we investigated a sepsis cohort and compared clinical biomarkers in relation to sepsis subtypes and showed the use case application of MTL in the ICU setting for the increased model predictive performance. Lastly, we explored ECG classification in addition to dimensionality reduction, clustering and class activation maps for model interpretability using waveform data. These developed models could inform intervention and additional cues and alerts in medical practice to aid clinical decision-making.

Chapter 3 describes the data sources used throughout the thesis. We analysed and compared three critical care databases, the MIMIC-III, eICU and AUMC, from the big data environment, which covers a range of clinical features and data sources in addition to unique characteristics. We explored the ML pipeline and the considerations that must be addressed to manipulate and construct an actionable dataset for analysis. We compared three database sources in a univariate and multivariate analysis classifying and predicting ICU LOS and mortality. The databases showed distinct differences in performances, and data availability, however, some unique similarities amongst biomarkers were displayed in the respected modelling outcomes. We demonstrated that biomarkers are heterogeneous depending on the data source and, similarly, regarding predictive performance from the data sources. This research is unique in that many studies have compared the MIMIC-III and eICU databases such as [253], however, few studies compare these databases to the AUMC for predictive performance and feature availability[254]. Additionally, we compare a univariate comparison between the databases, which showed statistical differences, again enforcing heterogeneity within the data sources and the need to further external validation models before implementation into clinical practice. Furthermore, geographical factors impact the predictive performance when modelling in the ICU. Therefore, developing models for international use will need data representing this demographic to negate any bias relating to the training data source.

Chapter 4 utilized research from Chapter 3 as the basic schematic to construct and model sepsis in the ICU. This chapter contains three sub-studies. Firstly, we use traditional statistical methods to investigate

in-hospital mortality in a sepsis cohort and compare predisposing factors and biomarkers. We compared our model to current critical care deterioration scores and showed that our tailored approach using ML yielded better results than traditional methods used in current practice. Additionally, we displayed that the origin of infection in relation to sepsis affects the associated feature of in-hospital mortality. Therefore, one model does not fit all sepsis subtypes. Finally, we displayed the odd ratios using sankey diagrams in a novel visualisation method. Although studies have been undertaken to analyse predisposing factors of sepsis, little take into account the origin of infection. Due to our unique study, this research resulted in a journal publication with validated medical insight. Next, we extended this study and investigated a range of outcomes, both hospital and ICU LOS and mortality, deploying a range of commonly used ML algorithms to compare inference and predictive performance. Our results show that the prediction performance favoured non-linear methods such as GBM and RF compared to traditional statistical methods such as linear and logistic regression in most modelling scenarios. Lastly, we extended the in-hospital mortality prediction methods with ML models to adapt an MTL framework to increase predictive performance. This framework has been shown to be successful in increasing the model's predictive performance in this instance. Many MTL methods have been used in the ICU with mortality, such as [255], [256]. However, none have implemented this with sepsis and utilised the subtypes to increase sepsis predictive performances, thus showing that the related tasks in the MTL framework were closely connected and, therefore, able to add inference to the other modelling tasks. However, further analysis would be required to determine if there were any clinical relevance to the inference of the features selected from the MTL models.

In Chapter 5, we aimed to investigate AF waveforms from a sepsis cohort and classify 30-second ECG segments to detect regions in long lead ECG records which contained AF characteristics. The training, validation and test set yielded excellent results, however, when testing the model with a random sample of non-sepsis ECG cases, the performance decreased. This may have been due to the relatively small sample size of the model's training cohort. Although we trained our model using over 160,000 samples of 30-second ECG segments (over 1200+ hours of ECG recording), this data only reflects a small cohort of 45 patients relating to 57 ECG records collectively. Ideally, a cohort size in the hundreds or thousands of ECG record annotations would have yielded great potential to overcome any generability or data artefact issues in the current research. However, the practicality of labelling thousands of hours of ECG records is a labour-intensive task and unrealistic. Our study used the 1809 30-second none sepsis ECG samples to test our model's performance. This external validation set collectively represents only 15 hours of ECG recording time, which, compared to the training data and overall records available in the MIMIC-III, is a tiny fraction of the potential data available.

Furthermore, the1809 non-sepsis clinically labelled ECGs displayed to be a difficult task, with many factors affecting the decision of the label given, such as the quality of the ECG, the presence of artefacts in the waveform, and the subjective nature of the interpretation of the ECG amongst the clinicians, amongst others. This was further investigated in relation to performance, we tested the ECGs with complete agreement between the clinical team, and we found that our developed model had improved predictive performance, displaying AUC scores of 93%. In contrast, compared to cases where a tiebreaker was needed to finalise the decision on the label, the AUC score reduced significantly to 82.61%. Moreover, in Chapter 5, we proposed several methods to address interpretability regarding the ECG classification through saliency maps and dimensionality reduction techniques. This allowed for further insight into the rationale of the ECG classification label and the probability given to each case. The research undertaken here could have far-reaching impacts in the medical setting as firstly, we have elaborated on the steps of detecting noise artefacts in ICU waveform data. Secondly, we have developed a model to classify AF vs Non-AF ECG cases. Furthermore, we can deduce a level of reasoning behind the 1D CNN's actions regarding each classified ECG. With further refinement and testing, clinicians and decision making could easily use and interpret the results from our model and methods.

In summary, this thesis aims to develop models that can be used to help aid clinical intervention in the ICU setting. We developed models for a range of outcomes, including in-hospital/in-ICU mortality and in-hospital/in-ICU LOS, regarding a range of cohorts, using various data sources, parameters, and

modelling techniques. The results and models generated can easily be used and interpreted by clinicians who could consequently inform intervention and best practices concerning patient care in the ICU environment.

6.2 Strength & Limitations

In this section, we aim to reflect on the experiments conducted. Although the research has accomplished clinical insight, shown ML applicability in the ICU, and progressed developed throughout the chapters. However, we also recognize that there are areas where some improvement and further considerations could be applied. This section will discuss the strengths and weaknesses of the research, reflecting on specific examples and discussing how these factors have impacted the research.

The research in Chapter 3 highlights a broad set of considerations that must be addressed as ML establishes its place in the healthcare domain. A key strength of this chapter is that first, we demonstrated that factors of importance, predictive performance and data availability were homogeneous depending on the data source. Secondly, We have displayed the foundations for future application of ICU analysis based on the ML frameworks, which were applied throughout the thesis to model a range of outcomes successfully. Lastly, we also explored a range of data attributes available in the three large open-source ICU databases, highlighting each data source's potential strengths and limitations.

Our research in Chapter 4 has displayed that the ML approach outperformed traditional methods of measuring mortality in the ICU. In addition, we have displayed a range of inference methods, such as the use of odd ratios, sankey diagrams, variable importance and SHAP analysis, to gain a more granular understanding of a particular outcome. As such, we have shown that ML algorithms can identify complex patterns and relationships within the data, which traditional clinical and statistical methods may miss. Therefore, this enables the algorithms to identify risk factors and predict outcomes more accurately[257]. This has been further highlighted throughout Chapter 4, where a more tailored and advanced approach to model sepsis and its subtypes achieved superior predictive performance to traditional methods. Ultimately, the research generated in Chapter 4 resulted in validated clinical insight, highlighting unique differences among the sepsis groups. Research conducted in the chapter was further published in clinical literature [258].

In our research in Chapter 5, we define an approach to detect AF in ICU patients. We also applied an innovative way of visualising data using dimensionality reduction techniques, clustering and saliency maps to investigate each ECG segment further for patterns and similarities. First, we developed a deep learning model that can classify AF and Non-AF ECGs using the near raw waveform data in sepsis patients with high performance. Lastly, we investigated the clustering of the ECG projections to explore cluster purities concerning the AF ECG class and whether the clusters would have similar ECG properties. The study's strengths included the in-depth analysis of the ECGs to gain an understanding or rationale of the logic of the 1D-CNN model. We displayed that we could generate inferences from the 1D-CNN model, where we utilised saliency maps to visualise what areas of each beat of the ECG record are essential to the final classification decision. Model interpretability is vital when applying these models to clinical applications. Therefore, understanding the rationale behind the model's prediction is paramount. We further presented how the raw ECG record can be visualised in a latent space to allow for easy end-user consumption. Additionally, this would enable the end-user to potentially identify the likelihood of the ECG class label based on its geographical location in the 2D latent space embedding created by the UMAP algorithm.

Collectively, the thesis displays that ML methods can be adapted and personalized to the individual patient's data, providing more accurate predictions than traditional methods. For instance, ML algorithms can use patient-specific features such as demographics, previous medical history, laboratory results, vital signs, and other relevant clinical factors to personalize predictions[259]. Chapter 4 highlighted this in the unique differences found in the sepsis subgroups and that unique models for each sepsis type allowed for optimal predictive performance and outperformed commonly used scoring

systems in the ICU. Ultimately there are many strengths of applying ML to healthcare applications, as displayed throughout Chapters 3-5. For example, we have displayed that predictive models, which were developed using ML methods, can achieve superior predictive performance. Therefore, it could aid healthcare providers in allocating better resources, including staff, beds, and equipment, or can be more efficient at managing patients and reducing mortality rates [260]. Ultimately ML can optimise a range of ICU areas and broader healthcare environments by adopting automated ML solutions for patient alerts and management. This thesis displays our ability to develop ML solutions for various ICU outcomes.

Holistically, this thesis has demonstrated a range of strengths, although the generalisability of the results is subject to certain limitations. One example is shown in the analysis, which was limited regarding the ability to externally validate the models developed on other ICU data sources. Although we had access to three large ICU databases, as discussed in Chapter 3, the available data is unique to each data source. Therefore, this would limit the ability to test models developed on eICU data to then test and validate on the MIMIC-III data. For instance, the coding ontologies used to recreate the analysis for experiments conducted in Chapter 4 could not be recreated using the MIMIC-III data because each data source coded the sepsis groups differently. Although we accounted for bias using the nested k-fold for the model validation method, in order to assess performance, generalisability, overfitting and confidence in our results, it is essential to further test on other clinical data sources for consistency. Similarly, in Chapter 5, this remains the same problem regarding externally cross-validating our results using other clinical data sources, such as local hospitals. However, the principal limitation of this chapter was the small number of patients and ECG records available with AF labels. Consequently, this resulted in acquiring a large amount of ECG records for a small number of patients, reflecting a small fraction of the total ECG data available in the MIMIC-III. Therefore, our training data may be an inadequate data representation, which may have introduced learning bias. In addition, it may not fully represent the diversity and variability of the wider AF and sepsis population used to train our model.

The study limitations are discussed to make an overall conclusion about the use and applicability of the developed models for local use in the UK. Although our research conducted throughout chapters 3-5 achieved excellent predictive performance, the necessity for validation for local use in the UK would require further investigation. In addition, a study published in the Journal of General Internal Medicine in 2018[261]examined the performance of a clinical prediction model for identifying patients at risk of hospital readmission. The model was developed using US data, and the study evaluated its performance on UK data. Moreover, the study discovered that the model performed poorly on the UK data, with poor discrimination and calibration. The researchers concluded that population differences were a likely contributing factor to the model's poor performance on the UK data. Similar research by [262] examined the transferability of clinical prediction models across different populations. The review found that many models developed in one cohort do not perform as well when applied to another population and that the population differences in patient characteristics, healthcare practices, and healthcare systems can all impact model performance. These findings were further emphasised in Chapter 3, as our results from the eICU and AUMC were significantly different regarding data structure, availability and predictive performance. As shown in the literature and our experiments in Chapter 3, population differences may be a potential obstacle to overcome and must be considered.

Overall, while it is possible for clinical models to work well across different populations, it is crucial to carefully evaluate their performance on new populations and consider potential population differences that may impact their performance. Therefore, future research would require further steps to mitigate any risk associated with population variations. The first step is obtaining a representative dataset from the new population we want to evaluate the model on, in our case, data locally sourced from UK healthcare practices. This dataset should include the same variables as the original datasets used throughout the thesis experiments and must be large enough to evaluate the model's performance appropriately. Consider that the model does not perform well on the new dataset. In this case, we should investigate any differences between the new population and the original US or Netherland population we trained our models on that may contribute to the poor performance. These differences could include

variations in patient characteristics, the prevalence of certain conditions, or how medical care is delivered. Depending on the nature of the differences between the new population and the original population, we may need to adapt the model to improve its performance on the new data. This could require us to retrain the model on the new data, adjust the model parameters or features to account for population differences, or use multi-task/transfer learning techniques to adapt the model to the new population. Once we have adapted the model, we should validate its performance on a separate validation dataset from the new population. Thus, this will allow us to assess whether the adapted model can accurately predict outcomes in the new population. Following such approaches will allow for more optimal model development and validation, ultimately allowing us to test our current research built from diverse data sources to UK patients and potentially other healthcare practices worldwide.

6.3 Future Work

While the results of this research are promising, with additional time and resources, the utilities of the model prediction may be improved and further developed into an interactive dashboard and software tools. Chapter 3 could be extended by considering a broader range of clinical parameters and combinations of variables to be used when modelling. There is also the potential to incorporate a multi-outcome solution to the problem to collectively determine the probability of mortality and LOS in one model rather than having separate models for each task. This could further lead to better generability and model predictive performance as the loss function of both tasks have to be considered simultaneously when modelling. Similarly, it could also be further developed in a multi-task or transfer learning setting due to the interconnected nature of these tasks.

Work in Chapter 4 could be extended by seeking data to validate the LOS and mortality prediction models externally. Although good predictive performance was yielded from the prediction models implemented, the need for external validation is paramount for the models to be deemed as reliable and thus not data specific. This result was captured in Chapter 3 as the data source significantly impacted the model's predictive performance. Similar to Chapter 3, in further work regarding Chapter 4, we would like to develop a multi-outcome model for the four outcomes investigated and extend this approach further in an MTL framework. An example would be developing a feed-forward neural network with task-specific outcome nodes for in-ICU/hospital mortality and LOS. These tasks are highly relatable, as a more severely ill patient has an increased chance of dying and thus an increased likelihood of the time spent in care. As the tasks are highly correlated, a multi-outcome model utilising an MTL approach would likely yield greater predictive performance.

Furthermore, methods such as SHAP and partial dependencies can be used to gain inference from the neural network, allowing rationales to be still determined regarding the outcome. Additionally, from a practical view, one model able to measure four outcomes is more likely manageable in practice than four models measuring a specific outcome. Lastly, research in Chapter 5 could further be extended, exploring the interpretability methods proposed in the multiclass classification domain using data such as in [263], [264] where more labels are present. In addition, it would be interesting to investigate if similar results can be generated in a multiclass classification setting and thus further explore the meaning of some of the regions of the ECGs embedded by UMAP. Additionally, we would like to validate the current implementation in a K-fold environment to account for the bias-variance trade-off not used in our current validation method. This limitation was due to the computational time to process thousands of hours of ECG recording and hyperparameter tuning and training of the models implemented.

Collectively, validating the models on data from the UK would help cement the usability of these models in Britain and other regions globally. However, the demographics of the populations vary between locations, and in our current research, we have experimented with data collected from only the US and the Netherlands. Whereas if we collected data gathered from the UK, where the models developed could potentially be used. We could use optimisation methods such as MTL to incorporate

the data from the US and Amsterdam currently available (MIMC, eICU and AUMC databases) to create internationally validated models for ICU care.

To conclude, the creation of models in the ICU environment has the potential to save money and enhance medical care and, most importantly, inform intervention and medical practice in the ICU. In addition, these models developed could help clinicians and medical staff make critical decisions in a fast-paced data-rich environment. Holistically, this thesis has demonstrated ML applicability in the ICU and other healthcare environments to optimise current practices, which in turn could potentially save patient lives and resources.

7 Supplement Material

Table 19: The Sequential Organ Failure Assessment (SOFA) score criteria [265]

The Sequential Organ Failure Assessment (SOFA) score	**SOFA Score**				
Organ system	0	1	2	3	4
Respiratory, PO2/FiO2, mmHg (kPa)	≥400 (53.3)	<400 (53.3)	<300 (40)	<200 (26.7) with respiratory support	<100 (13.3) with respiratory
Coagulation, Platelets, ×103/mm3	≥150	<150	<100	<50	<20
Liver, Bilirubin, mg/dL	<1.2	1.2–1.9	2.0–5.9	6.0–11.9	>12.0
Cardiovascular	MAP ≥70 mmHg	MAP <70 mmHg	Dopamine <5 or dobutamine (any dose)	Dopamine 5.1–15 or epinephrine ≤0.1 or norepinephrine ≤0.1b	Dopamine >15 or epinephrine >0.1 or norepinephrine >0.1b
Central nervous system, Glasgow Coma Scale	15	13–14	10–12	6–9	<6
Renal, Creatinine, mg/dL.	<1.2	1.2–1.9	2.0–3.4	3.5–4.9	>5.0
Urine output, mL/d				<500	<200

Table 20: Quick Sequential Organ Failure Assessment (SOFA) score criteria [265]

qSOFA (Quick SOFA) Criteria	points
Respiratory rate ≥22/min	1
Change in mental status, Glasgow coma scale ≤ 14	1
Systolic blood pressure ≤100 mmHg	1

Table 21: SIRS: Systemic Inflammatory Response score criteria [266].

SIRS Criteria	points
36 > Temperature > 38	1
Respiratory rate ≥22/min	1
Heart Rate > 90 bpm	1
4000 > White cell count > 12000	1

Table 22: Charlson Comorbidity index definition [267]. MI (myocardial infraction) CHF(congestive heart failure).

Charlson Comorbidity Index	Assigned weights for each condition
MI	1
CHF	1
peripheral vascular disease	1
cerebrovascular disease	1
dementia	1
chronic pulmonary disease	1
connective tissue disease	1
Ucler disease	1
diabetes	1
Hemiplegia	2
Moderate or severe renal disease	2
Diabetes with end organ damage	2
any tumor without metastasis	2
leukema	2
Lymphoma	2
Moderate or severe liver disease	3
Metastatic solid tumer	6
AIDS	6

Table 23: Acute Physiology and Chronic Health Evaluation (APACHE) IV score criteria [268]

Acute Physiology and Chronic Health Evaluation (APACHE) IV Score	
Clinical Features	Definition
Age	Continuous Measure Plus Five Spline Terms
APS variables	Weight determined by most abnormal value within first APACHE day; sum of weights equals the APS, which ranges from 0 to 252. Five spline terms added. Variables include pulse rate, mean blood pressure, temperature, respiratory rate, PaO2/FIO2 ratio (or P(A-a)O2 for intubated patients with FIO2 0.5), hematocrit, white blood cell count, creatinine, urine output, blood urea nitrogen, sodium, albumin, bilirubin, glucose, acid base abnormalities, and neurological abnormalities based on Glasgow Coma Score
Chronic health variables	AIDS, cirrhosis, hepatic failure, immunosupression, lymphoma, leukemia or myeloma, metastatic tumor. Not used for elective surgery patients
ICU admission diagnosis	116 categories
ICU admission source	Floor, emergency room, operating/recovery room, stepdown unit, direct admission, other ICU, other hospital, other admission source
Length of stay before ICU admission	Square root plus four spline terms
Emergency surgery	Y/N
Unable to assess Glasgow	Y/N
Thrombolytic therapy	For patients with acute myocardial infarction (Y/N)
Glasgow Coma Scale score rescaled	15 minus measured Glasgow Coma Scale score
PaO2/FIO2 ratio	
Mechanical ventilation	Y/N

Table 24: All variables used in each sepsis group for the final developed model.

Sepsis Groups	All Features Selected for Logistic Regression
Abdominal	Age, Albumin Var, Asthma, Atrial Fibrillation, Avg Albumin, Avg Creatinine, Avg FiO2, Avg GCS Total, Avg Glucose, Avg Heart Rate, Avg Hematocrit, Avg Lymphs, Avg MAP, Avg PaCO2, Avg PaO2, Avg PH, Avg Platelets, Avg Resp Rate, Avg Sodium, Avg Temp °C, Avg Total Bilirubin, Avg Urine, Avg WBC, BUN Var, CABG, Cancer, Creatinine Var, CTD, Dementia, Dobutamine, Dopamine, Endocrine, FiO2 Var, GCS Total Var, Gender (Male), Glucose Var, Heart Rate Var, Hematocrit Var, Hemiplegia, Hypertension, Hypothyroidism, Infectious Diseases, Intubated, Lymphs Var, MAP Var, Mild Liver Disease, Neurologic, Norepinephrine, Oncology, PaCO2 Var, PaO2 Var, Phenylephrine, Platelets Var, Pulmonary, Renal Disease, Resp Rate Var, Respiratory Failure, SaO2 Var, Seizures, Severe Liver Disease, Sodium Var, Temp °C Var, Uncomplicated DM, Unit Stay Type (Admit), Unit Stay Type (Other/Stepdown/Transfer), Unit Stay Type (Readmit), Unit Type (SICU), Urine Var, Vasopressin, WBC Var.
Pulmonary	Age, Atrial Fibrillation, Avg Albumin, Avg BUN, Avg FiO2, Avg GCS Total, Avg Heart Rate, Avg MAP, Avg PaO2. Avg SaO2. Avg Temp °C, Avg Total Bilirubin, Avg Urine, Cancer, Cardiovascular, CHF, Dementia, GCS Total Var, Gender (Male), Heart Rate Var, Hypothyroidism, Intubated, Norepinephrine, Oncology, PaCO2 Var, Phenylephrine, Platelets Var, Renal Disease, Resp Rate Var, Respiratory Failure, SaO2 Var, Total Bilirubin Var, Unit Stay Type (Admit), Unit Stay Type (Readmit), Unit Type (Med-Surg ICU), Unit Type (MICU), Vasopressin.
Renal/UTI	Age, Albumin Var, Asthma, Atrial Fibrillation, Avg Albumin, Avg BUN, Avg Creatinine, Avg FiO2, Avg GCS Total, Avg Glucose, Avg Heart Rate, Avg Hematocrit, Avg Lymphs, Avg MAP, Avg PaCO2, Avg PaO2, Avg PH, Avg Platelets, Avg Resp Rate, Avg SaO2, Avg Sodium, Avg Temp °C, Avg Total Bilirubin, Avg Urine, Avg WBC, BUN Var, CABG, Cancer, Cardiovascular, CHF, COPD, Creatinine Var, CTD, Dementia, Dobutamine, Dopamine, Endocrine, Epinephrine, FiO2 Var, Gastrointestinal, GCS Total Var, Gender (Male), Glucose Var, Heart Rate Var, Hematocrit Var, Hemiplegia, Hypertension, Hypothyroidism, Infectious Diseases, Intubated, Lymphs Var, MAP Var, Mild Liver Disease, Myocardial Infarction, Neurologic, Norepinephrine, Oncology, PaCO2 Var, PaO2 Var, Peptic Ulcer Disease, PH Var, Phenylephrine, Platelets Var, Pulmonary, PVD, Renal, Renal Disease, Resp Rate Var, Respiratory Failure, SaO2 Var, Seizures, Severe Liver Disease, Sodium Var, Temp °C Var, Total Bilirubin Var, Uncomplicated DM, Unit Stay Type (Admit), Unit Stay Type Other/Stepdown/Transfer), Unit Stay Type (Readmit),Unit Type (Med-Surg ICU), Unit Type (MICU), Unit Type (SICU), Urine Var, Vasopressin, WBC Var.

Table 25: Table of eICU variable names and clinical groupings.

Original name	Clean name	Group	Clinical group	Index
Pulmonary	Pulmonary	Admission diagnosis	A	1
Cardiovascular	Cardiovascular	Admission diagnosis	A	2
Infectious_Diseases	Infectious Diseases	Admission diagnosis	A	3
Renal	Renal	Admission diagnosis	A	4
Gastrointestinal	Gastrointestinal	Admission diagnosis	A	5
Oncology	Oncology	Admission diagnosis	A	6
Neurologic	Neurologic	Admission diagnosis	A	7
Endocrine	Endocrine	Admission diagnosis	A	8
Age	Age	Demographics	B	9
Gender	Gender (Male)	Demographics	B	10
Coronary_Artery_Bypass	CABG	Comorbidities	C	11
Myocardial_Infarction	Myocardial Infarction	Comorbidities	C	12
Congestive_Heart_Failure	CHF	Comorbidities	C	13
Peripheral_Vascular_Disease	PVD	Comorbidities	C	14
Hypertension	Hypertension	Comorbidities	C	15
AF	Atrial Fibrillation	Comorbidities	C	16
Hemiplegia	Hemiplegia	Comorbidities	C	17
Heart_Rate	Avg Heart Rate	Vitals	C	18
Heart_Rate_SD	Heart Rate Var	Vitals	C	19
Non_Invasive_BP_Systolic	Avg Non-Inv SBP	Vitals	C	20
Non_Invasive_BP_Diastolic	Avg Non-Inv DBP	Vitals	C	21
Non_Invasive_BP_Mean	Avg Non-Inv MBP	Vitals	C	22
Invasive_BP_Systolic	Avg Invasive SBP	Vitals	C	23
Invasive_BP_Diastolic	Avg Invasive DBP	Vitals	C	24
Invasive_BP_Mean	Avg Invasive MBP	Vitals	C	25
Non_Invasive_BP_Systolic_SD	Non-Inv SBP Var	Vitals	C	26
Non_Invasive_BP_Diastolic_SD	Non-Inv DBP Var	Vitals	C	27
Non_Invasive_BP_Mean_SD	Non-Inv MBP Var	Vitals	C	28
Invasive_BP_Systolic_SD	Invasive SBP Var	Vitals	C	29
Invasive_BP_Diastolic_SD	Invasive DBP Var	Vitals	C	30
Invasive_BP_Mean_SD	Invasive MBP Var	Vitals	C	31
Asthma	Asthma	Comorbidities	D	32
Respiratory_Failure	Respiratory Failure	Comorbidities	D	33
COPD	COPD	Comorbidities	D	34
PaO2	Avg PaO2	Laboratory	D	35
PaCO2	Avg PaCO2	Laboratory	D	36
FiO2	Avg FiO2	Laboratory	D	37
PH	Avg PH	Laboratory	D	38
PaO2_SD	PaO2 Var	Laboratory	D	39
PaCO2_SD	PaCO2 Var	Laboratory	D	40
FiO2_SD	FiO2 Var	Laboratory	D	41
PH_SD	PH Var	Laboratory	D	42

Ventilation	Ventilation	Respiration	D	43
Spontaneous	Spontaneous	Respiration	D	44
Non_Invasive_Ventilation	NIV	Respiration	D	45
Intubated	Intubated	Respiration	D	46
O2_Saturation	Avg SaO2	Vitals	D	47
Respiratory_Rate	Avg Resp Rate	Vitals	D	48
O2_Saturation_SD	SaO2 Var	Vitals	D	49
Respiratory_Rate_SD	Resp Rate Var	Vitals	D	50
Renal_Disease	Renal Disease	Comorbidities	E	51
Creatinine	Avg Creatinine	Laboratory	E	52
BUN	Avg BUN	Laboratory	E	53
Albumin	Avg Albumin	Laboratory	E	54
Sodium	Avg Sodium	Laboratory	E	55
Creatinine_SD	Creatinine Var	Laboratory	E	56
BUN_SD	BUN Var	Laboratory	E	57
Albumin_SD	Albumin Var	Laboratory	E	58
Sodium_SD	Sodium Var	Laboratory	E	59
WBC	Avg WBC	Laboratory	F	60
Lymphs	Avg Lymphs	Laboratory	F	61
Platelets	Avg Platelets	Laboratory	F	62
Hct	Avg Hematocrit	Laboratory	F	63
WBC_SD	WBC Var	Laboratory	F	64
Lymphs_SD	Lymphs Var	Laboratory	F	65
Platelets_SD	Platelets Var	Laboratory	F	66
Hct_SD	Hematocrit Var	Laboratory	F	67
Temperature_C	Avg Temp °C	Vitals	F	68
Temperature_C_SD	Temp °C Var	Vitals	F	69
Mild_Liver_Disease	Mild Liver Disease	Comorbidities	G	70
Sever_Liver_Disease	Severe Liver Disease	Comorbidities	G	71
Total_Bilirubin	Avg Total Bilirubin	Laboratory	G	72
Total_Bilirubin_SD	Total Bilirubin Var	Laboratory	G	73
Norepinephrine	Norepinephrine	Drugs	H	74
Vasopressin	Vasopressin	Drugs	H	75
Phenylephrine	Phenylephrine	Drugs	H	76
Dopamine	Dopamine	Drugs	H	77
Epinephrine	Epinephrine	Drugs	H	78
Dobutamine	Dobutamine	Drugs	H	79
Connective_Tissue_Disease	CTD	Comorbidities	I	80
Peptic_Ulcer_Disease	Peptic Ulcer Disease	Comorbidities	I	81
Uncomplicated_Diabetes	Uncomplicated DM	Comorbidities	I	82
Hypothyroidism	Hypothyroidism	Comorbidities	I	83
Seizures	Seizures	Comorbidities	I	84
Dementia	Dementia	Comorbidities	I	85
Cancer	Cancer	Comorbidities	I	86
Glucose	Avg Glucose	Laboratory	I	87

Glucose_SD	Glucose Var	Laboratory	I	88
Urine	Avg Urine	Laboratory	I	89
Urine_SD	Urine Var	Laboratory	I	90
GCS_Total	Avg GCS Total	Vitals	I	91
GCS_Total_SD	GCS Total Var	Vitals	I	92
tsk_col	Task	Task	J	93
Hospital_Discharge_Status	In-hospital Mortality	Outcome	Outcome	94
Unit_Stay_Type_admit	Unit Stay Type (Admit)	Unit Stay Type	K	95
Unit_Stay_Type_Other_Stepdown_Transfer	Unit Stay Type (Other/Stepdown/Transfer)	Unit Stay Type	K	96
Unit_Stay_Type_readmit	Unit Stay Type (Readmit)	Unit Stay Type	K	97
Unit_Type_Med_Surg_ICU	Unit Type (Med-Surg ICU)	Unit Type	L	98
Unit_Type_MICU	Unit Type (MICU)	Unit Type	L	99
Unit_Type_SICU	Unit Type (SICU)	Unit Type	L	100
MAP	Avg MAP	Vitals	C	101
MAP_SD	MAP Var	Vitals	C	102
Unit_Type	Unit Type	Unit Type	L	103
Unit_Stay_Type	Unit Stay Type	Unit Stay Type	L	104
Unit_Discharge_Status	In-ICU Mortality	Outcome2	Outcome	105
Time_ICU	ICU LOS	Admission Duration	Outcome	106
Time_Admission	Hospital LOS	Admission Duration	Outcome	107
SOFA	SOFA	Scores	M	108
SIRS	SIRS	Scores	M	109
qSOFA	qSOFA	Scores	M	110
Patient_Unit_Stay_ID	Patient Unit Stay ID	ID	Z	111
Patient_Health_System_Stay_ID	Patient Health System Stay ID	ID	Z	112
Charlson_CI	Charlson CI	Scores	M	113
APACHE_IV	APACHE IV	Scores	M	114
Apache_Admissiondx	Apache Admission	ICU Admission Diagnosis	N	115

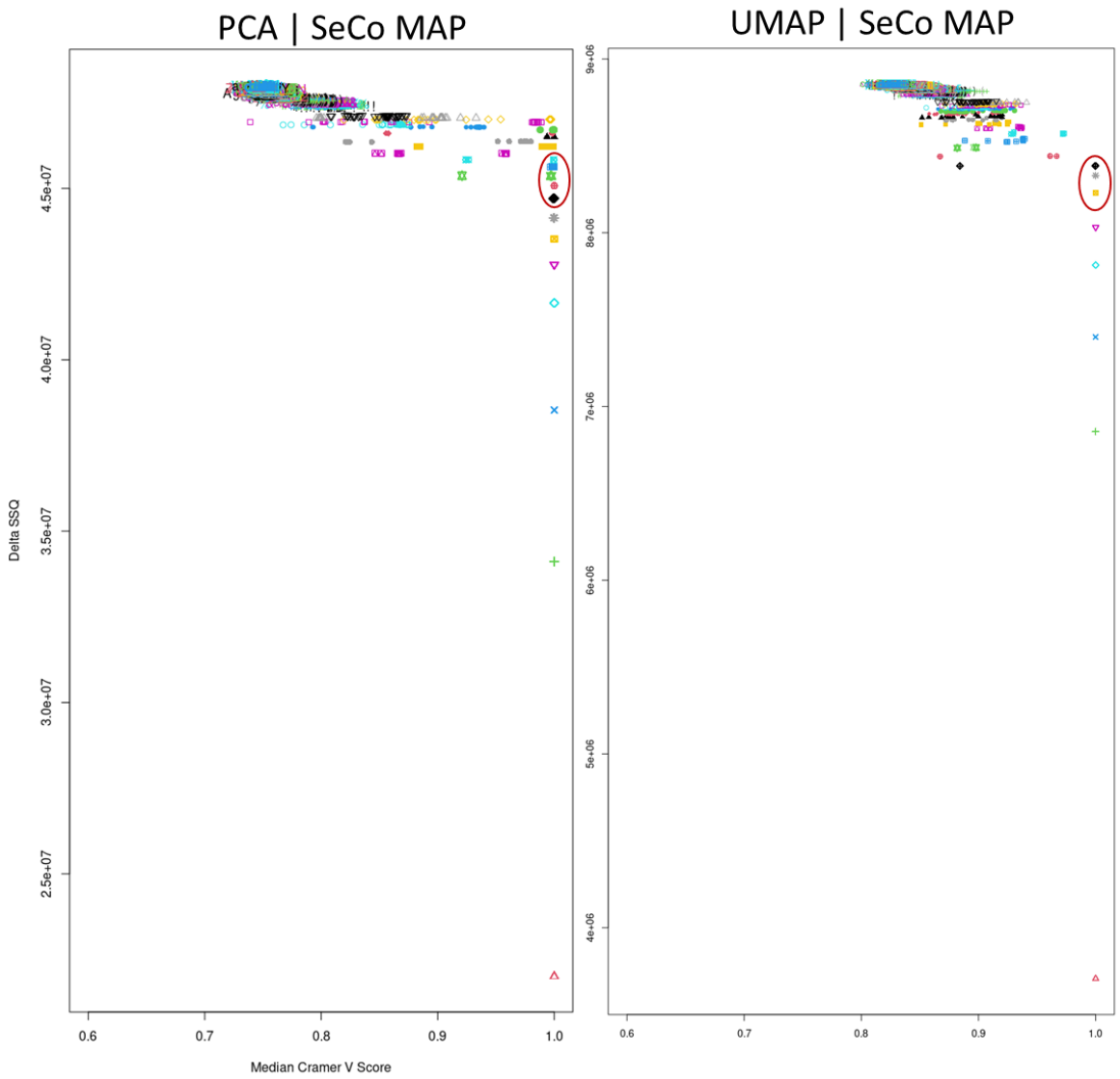


Figure 35: The SeCo maps for the UMAP and PCA projections to find the optimal kmeans value.

Table 26: ECG Metadata collected for each ECG record and feature definition.

Feature name	Description
HR_Dectected	Heart detected in ECG segment
BPM	Beats per minute
IBI	Interbeat interval
SDNN	The standard deviation of RR intervals
SDSD	The standard deviation of successive differences
RMSSD	Root mean square of successive differences
PNN20	The proportion of successive differences above 20ms
PNN50	The proportion of successive differences above 50ms
hr_mad	Median absolute deviation of RR intervals
SD1	Poincare parameters: standard deviation perpendicular to the identity line
SD2	Poincare parameters: standard deviation along the identity line
S	Poincare parameters: Area of ellipse described by SD1 and SD2
SD1/SD2	Poincare parameters: SD1/SD2 ratio
BreathingRate	Breathing rate
VLF	Time-series measurements calculated by welch: very low-frequency, frequency spectrum between 0.05.15Hz
LF	Time-series measurements calculated by welch: low-frequency, frequency spectrum between 0.05-0.15Hz
HF	Time-series measurements calculated by welch: high-frequency, frequency spectrum between 0.05-0.15Hz
LF/HF	Time-series measurements calculated by welch: the ratio of high frequency / low frequency
p_total	Time-series measurements calculated by welch: the ratio of high frequency / low frequency
vlf_perc	Time-series measurements calculated by welch: very low-frequency, frequency spectrum between 0.05.15Hz - %
lf_perc	Time-series measurements calculated by welch: low-frequency, frequency spectrum between 0.05-0.15Hz - %
hf_perc	Time-series measurements calculated by welch: high-frequency, frequency spectrum between 0.05-0.15Hz - %
lf_nu	Time-series measurements calculated by welch: low-frequency, frequency spectrum between 0.05-0.15Hz - normalised
hf_nu	Time-series measurements calculated by welch: high-frequency, frequency spectrum between 0.05-0.15Hz - normalised
n_rejected_peaks	The total number of rejected peaks
n_rr	The total number of RR peaks detected
n_qrs	The total number of QRS peaks detected
mean	Mean value of normalised ECG segment
var	The variance of ECG values
skew	The skew of ECG values
kurtosis	The Kurtosis of ECG values
min	The min value of the ECG segment
max	The max value of the ECG segment

8 References

[1] P. J. Pronovost, D. C. Angus, T. Dorman, K. A. Robinson, T. T. Dremsizov, and T. L. Young, "Physician staffing patterns and clinical outcomes in critically III patients: A systematic review," *J Am Med Assoc*, vol. 288, no. 17, pp. 2151–2162, Nov. 2002, doi: 10.1001/jama.288.17.2151.

[2] N. A. Halpern and S. M. Pastores, "Critical care medicine in the United States 2000–2005: An analysis of bed numbers, occupancy rates, payer mix, and costs*," *Crit Care Med*, vol. 38, no. 1, pp. 65–71, Jan. 2010, doi: 10.1097/CCM.0b013e3181b090d0.

[3] J. L. Vincent *et al.*, "Assessment of the worldwide burden of critical illness: The Intensive Care Over Nations (ICON) audit," *Lancet Respir Med*, vol. 2, no. 5, pp. 380–386, May 2014, doi: 10.1016/S2213-2600(14)70061-X.

[4] A. Davoudi *et al.*, "Intelligent ICU for Autonomous patient Monitoring Using pervasive sensing and Deep Learning", doi: 10.1038/s41598-019-44004-w.

[5] L. J. Hirsch, "Continuous EEG monitoring in the intensive care unit: An overview," *Journal of Clinical Neurophysiology*, vol. 21, no. 5. pp. 332–340, 2004. doi: 10.1097/01.WNP.0000147129.80917.0E.

[6] J.-L. Vincent and J. Creteur, "Paradigm shifts in critical care medicine: the progress we have made," 2015. doi: 10.1186/cc14728.

[7] C. W. Hug and P. Szolovits, "ICU acuity: real-time models versus daily models.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2009, pp. 260–264, 2009.

[8] W. A. Knaus *et al.*, "The APACHE III prognostic system: Risk prediction of hospital mortality for critically III hospitalized adults," *Chest*, vol. 100, no. 6, pp. 1619–1636, Dec. 1991, doi: 10.1378/chest.100.6.1619.

[9] J. R. Gall, S. Lemeshow, and F. Saulnier, "A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study," *JAMA: The Journal of the American Medical Association*, vol. 270, no. 24, pp. 2957–2963, Dec. 1993, doi: 10.1001/jama.1993.03510240069035.

[10] U. Janssens *et al.*, "Evaluation of the sofa score: A single-center experience of a medical intensive care unit 303 consecutive patients with predominantly cardiovascular disorders," *Intensive Care Med*, vol. 26, no. 8, pp. 1037–1045, 2000, doi: 10.1007/s001340051316.

[11] J.-R. Le Gall, "The use of severity scores in the intensive care unit," *Intensive Care Med*, vol. 31, no. 12, pp. 1618–1623, Dec. 2005, doi: 10.1007/s00134-005-2825-8.

[12] E. Barbini, G. Cevenini, S. Scolletta, B. Biagioli, P. Giomarelli, and P. Barbini, "A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery - Part I: Model planning," *BMC Medical Informatics and Decision Making*, vol. 7. BioMed Central, p. 35, 2007. doi: 10.1186/1472-6947-7-35.

[13] S. et Al, "Predictive models in critical care: a scoping review.," *Crit Care*, vol. 23, no. 1, pp. 1–9, 2019, doi: 10.1186/s13054-019-2437-4.

[14] W. et Al, "A systematic review of the use of administrative data for longitudinal outcomes research in emergency medicine.," *Academic Emergency Medicine*, vol. 27, no. 8, pp. 649–652, Aug. 2020, doi: 10.1111/acem.13745.

[15] A. Rajkomar *et al.*, “Scalable and accurate deep learning with electronic health records,” *npj Digital Medicine 2018 1:1*, vol. 1, no. 1, pp. 1–10, May 2018, doi: 10.1038/s41746-018-0029-1.

[16] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, “A Review of Challenges and Opportunities in Machine Learning for Health,” *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020, Accessed: Apr. 27, 2023. [Online]. Available: /pmc/articles/PMC7233077/

[17] D. C. Bouch and J. P. Thompson, “Severity scoring systems in the critically ill,” *Continuing Education in Anaesthesia Critical Care & Pain*, vol. 8, no. 5, pp. 181–185, Sep. 2008, doi: 10.1093/bjaceaccp/mkn033.

[18] L. A. Winters-Miner, *Practical predictive analytics and decisioning systems for medicine : informatics accuracy and cost-effectiveness for healthcare administration and delivery including medical research*. Academic Press, 2015.

[19] T. Sinuff *et al.*, “Mortality predictions in the intensive care unit: Comparing physicians with scoring systems*,” *Crit Care Med*, vol. 34, no. 3, pp. 878–885, Mar. 2006, doi: 10.1097/01.CCM.0000201881.58644.41.

[20] E. S. Ford *et al.*, “Explaining the Decrease in U.S. Deaths from Coronary Disease, 1980–2000,” *New England Journal of Medicine*, vol. 356, no. 23, pp. 2388–2398, Jun. 2007, doi: 10.1056/NEJMsa053935.

[21] E. G. Nabel and E. Braunwald, “A Tale of Coronary Artery Disease and Myocardial Infarction,” *New England Journal of Medicine*, vol. 366, no. 1, pp. 54–63, Jan. 2012, doi: 10.1056/NEJMra1112570.

[22] W. S. Weintraub, A. C. Fahed, and J. S. Rumsfeld, “Translational Medicine in the Era of Big Data and Machine Learning.,” *Circ Res*, vol. 123, no. 11, pp. 1202–1204, Nov. 2018, doi: 10.1161/CIRCRESAHA.118.313944.

[23] Z. Yan *et al.*, “Multi-Instance Deep Learning: Discover Discriminative Local Anatomies for Bodypart Recognition,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1332–1343, May 2016, doi: 10.1109/TMI.2016.2524985.

[24] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, “Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network,” *IEEE Trans Med Imaging*, vol. 35, no. 5, pp. 1207–1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[25] J. Mehta and A. Majumdar, “RODEO: Robust DE-aliasing autoencOder for real-time medical image reconstruction,” *Pattern Recognit*, vol. 63, pp. 499–510, Mar. 2017, doi: 10.1016/j.patcog.2016.09.022.

[26] M. Havaei *et al.*, “Brain tumor segmentation with Deep Neural Networks,” *Med Image Anal*, vol. 35, pp. 18–31, Jan. 2017, doi: 10.1016/j.media.2016.05.004.

[27] B. E. Bejnordi *et al.*, “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer,” *JAMA - Journal of the American Medical Association*, vol. 318, no. 22, pp. 2199–2210, Dec. 2017, doi: 10.1001/jama.2017.14585.

[28] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” Nov. 2017.

[29] V. Gulshan *et al.*, “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,” *JAMA - Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016, doi: 10.1001/jama.2016.17216.

[30] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[31] X. Zeng and G. Luo, “Progressive Sampling-Based Bayesian Optimization for Efficient and Automatic Machine Learning Model Selection,” *Health Inf Sci Syst*, vol. 5, no. 1, Dec. 2018, doi: 10.1007/s13755-017-0023-z.

[32] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997, doi: 10.1109/4235.585893.

[33] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, and N. K. Jha, “Systematic poisoning attacks on and defenses for machine learning in healthcare,” *IEEE J Biomed Health Inform*, vol. 19, no. 6, pp. 1893–1905, Nov. 2015, doi: 10.1109/JBHI.2014.2344095.

[34] M. A. Ahmad, A. Teredesai, and C. Eckert, “Interpretable machine learning in healthcare,” in *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, p. 447. doi: 10.1109/ICHI.2018.00095.

[35] F. Doshi-Velez *et al.*, “Accountability of AI Under the Law: The Role of Explanation,” *SSRN Electronic Journal*, Nov. 2017.

[36] A. Vellido, “The importance of interpretability and visualization in machine learning for applications in medicine and health care,” *Neural Comput Appl*, pp. 1–15, Feb. 2019, doi: 10.1007/s00521-019-04051-w.

[37] J. Labarère, R. Bertrand, and M. J. Fine, “How to derive and validate clinical prediction models for use in intensive care medicine,” *Intensive Care Med*, vol. 40, no. 4, pp. 513–527, Feb. 2014, doi: 10.1007/s00134-014-3227-6.

[38] K. G. M. Moons *et al.*, “Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker,” *Heart*, vol. 98, no. 9. BMJ Publishing Group Ltd and British Cardiovascular Society, pp. 683–690, May 01, 2012. doi: 10.1136/heartjnl-2011-301246.

[39] O. Yildirim, U. B. Baloglu, R. S. Tan, E. J. Ciaccio, and U. R. Acharya, “A new approach for arrhythmia classification using deep coded features and LSTM networks,” *Comput Methods Programs Biomed*, 2019, doi: 10.1016/j.cmpb.2019.05.004.

[40] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Interpretable machine learning: definitions, methods, and applications,” *Proc Natl Acad Sci U S A*, vol. 116, no. 44, pp. 22071–22080, Jan. 2019, doi: 10.1073/pnas.1900654116.

[41] F. Fan, J. Xiong, and G. Wang, “On Interpretability of Artificial Neural Networks,” Jan. 2020.

[42] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”, doi: 10.1038/s42256-019-0048-x.

[43] P. J. G. Lisboa, S. Ortega-Martorell, S. Cashman, and I. Olier, “The Partial Response Network: a neural network nomogram,” Aug. 2019.

[44] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining Explanations: An Overview of Interpretability of Machine Learning,” 2019.

[45] “Broad Agency Announcement Explainable Artificial Intelligence (XAI),” 2016.

[46] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. Van Calster, “A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 110. Elsevier USA, pp. 12–22, Jun. 01, 2019. doi: 10.1016/j.jclinepi.2019.02.004.

[47] D. D. Gutierrez, *Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R by Daniel D. Gutierrez*, 1st ed. Technics Publications , 2015.

[48] Shai. Shalev-Shwartz and Shai. Ben-David, “From Theory to Algorithms,” in *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

[49] M. Z. Alom *et al.*, “The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches,” Mar. 2018.

[50] D. H. Maulud and A. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/JASTT1457.

[51] M. E. Shipe, S. A. Deppen, F. Farjah, and E. L. Grogan, “Developing prediction models for clinical use using logistic regression: An overview,” *Journal of Thoracic Disease*, vol. 11, no. Suppl 4. AME Publishing Company, pp. S574–S584, 2019. doi: 10.21037/jtd.2019.01.25.

[52] C. Y. J. Peng, T. S. H. So, F. K. Stage, and E. P. St. John, “The use and interpretation of logistic regression in higher education journals: 1988-1999,” *Research in Higher Education*. 2002. doi: 10.1023/A:1014858517172.

[53] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statiscal Learning with Applications in R*. 2013. doi: 10.1016/j.peva.2007.06.006.

[54] and C. J. S. L. Breiman, J. H. Friedman, R. A. Olshen, “Classification and Regression Trees, no, Ed. Belmont, CA: Wadsworth International Group, 1984.,” *Mach Learn*, 1993, doi: 10.1109/ICETET.2008.143.

[55] T. Yiu, “Understanding Random Forest - Towards Data Science,” 2019. https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (accessed Jan. 06, 2020).

[56] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms,” in *ACM International Conference Proceeding Series*, 2006, pp. 161–168. doi: 10.1145/1143844.1143865.

[57] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statiscal Learning with Applications in R*. 2013. doi: 10.1016/j.peva.2007.06.006.

[58] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurorobot*, vol. 7, no. DEC, p. 21, Dec. 2013, doi: 10.3389/fnbot.2013.00021.

[59] Z. Yang *et al.*, “Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets,” *Comput Mater Sci*, vol. 151, pp. 278–287, Aug. 2018, doi: 10.1016/j.commatsci.2018.05.014.

[60] L. Bermudez, “Overview of Neural Networks - machinevision - Medium,” 2017. https://medium.com/machinevision/overview-of-neural-networks-b86ce02ea3d1#targetText=The human brain consists of,that neuron will also fire.&targetText=This neural network has one,layers%2C inputs%2C or outputs. (accessed Aug. 24, 2019).

[61] A. C. Ian Goodfellow, Yoshua Bengio, *Deep Learning*. 2016.

[62] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42. Elsevier B.V., pp. 146–157, Jul. 01, 2018. doi: 10.1016/j.inffus.2017.10.006.

[63] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao, "Deep Learning and Its Applications to Machine Health Monitoring: A Survey."

[64] S. Pouyanfar *et al.*, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys*. 2018. doi: 10.1145/3234150.

[65] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[66] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3. pp. 1157–1182, Mar. 2003. doi: 10.1162/153244303322753616.

[67] E. A. Mohammed, C. Naugler, and B. H. Far, "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics," in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools*, Elsevier Inc., 2015, pp. 577–602. doi: 10.1016/B978-0-12-802508-6.00032-6.

[68] A. Sarica, A. Cerasa, and A. Quattrone, "Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review," *Front Aging Neurosci*, vol. 9, no. OCT, p. 329, Oct. 2017, doi: 10.3389/fnagi.2017.00329.

[69] Z. Xu, G. Huang, K. Q. Weinberger, and A. X. Zheng, "Gradient boosted feature selection," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, New York, USA: Association for Computing Machinery, 2014, pp. 522–531. doi: 10.1145/2623330.2623635.

[70] M. F. Bellolio, L. A. Serrano, and L. G. Stead, "Understanding statistical tests in the medical literature: which test should I use?," *Int J Emerg Med*, vol. 1, no. 3, pp. 197–199, Sep. 2008, doi: 10.1007/s12245-008-0061-z.

[71] M. L. McHugh, "The Chi-square test of independence," *Biochem Med (Zagreb)*, vol. 23, no. 2, pp. 143–149, Jun. 2012, doi: 10.11613/BM.2013.018.

[72] J. Kloke and J. W. McKean, *Nonparametric Statistical Methods Using R*. 2014. doi: 10.1201/b17501.

[73] P. E. McKight and J. Najab, "Kruskal-Wallis Test," in *The Corsini Encyclopedia of Psychology*, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2010, pp. 1–1. doi: 10.1002/9780470479216.corpsy0491.

[74] J. D. Spurrier, "Nonparametric Statistics On the null distribution of the Kruskal-Wallis statistic ON THE NULL DISTRIBUTION OF THE KRUSKAL-WALLIS STATISTIC," *Nonparametric Statistics*, vol. 15, no. 6, pp. 685–691, 2003, doi: 10.1080/10485250310001634719.

[75] T. Fawcett, "doi:10.1016/j.patrec.2005.10.010," 2005, doi: 10.1016/j.patrec.2005.10.010.

[76] B. Publishing Ltd, J. M. Lobo, A. Jiménez-Valverde, and R. Real, "ECOLOGICAL SOUNDING AUC: a misleading measure of the performance of predictive distribution models," 2007, doi: 10.1111/j.1466-8238.2007.00358.x.

[77] N. Sarang, “Understanding AUC - ROC Curve - Towards Data Science,” 2018. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 (accessed Aug. 21, 2019).

[78] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)?,” *Geosci. Model Dev. Discuss*, vol. 7, pp. 1525–1534, 2014, doi: 10.5194/gmdd-7-1525-2014.

[79] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” 1995.

[80] A. S. U. Refailzadeh Payam, Lei Tang, Huan Liu, “Cross-Validaton,” 2009. doi: 10.1159/000323569.

[81] J. Wainer and G. Cawley, “Nested cross-validation when selecting classifiers is overzealous for most practical applications,” *Expert Syst Appl*, vol. 182, p. 115222, Nov. 2021, doi: 10.1016/J.ESWA.2021.115222.

[82] C. O. S. Sorzano, J. Vargas, and A. P. Montano, “A survey of dimensionality reduction techniques,” Mar. 2014, doi: 10.48550/arxiv.1403.2877.

[83] S. Jialin Pan, J. T. Kwok, and Q. Yang, “Transfer Learning via Dimensionality Reduction”, Accessed: Nov. 06, 2022. [Online]. Available: www.aaai.org

[84] Y. Lin, X. Zhu, Z. Zheng, Z. Dou, and R. Zhou, “The individual identification method of wireless device based on dimensionality reduction and machine learning,” *The Journal of Supercomputing 2017 75:6*, vol. 75, no. 6, pp. 3010–3027, Dec. 2017, doi: 10.1007/S11227-017-2216-2.

[85] R. R. Zebari, A. Mohsin Abdulazeez, D. Q. Zeebaree, D. A. Zebari, and J. N. Saeed, “A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction,” vol. 01, no. 02, pp. 56–70, 2020, doi: 10.38094/jastt1224.

[86] M. Verleysen and D. François, “The curse of dimensionality in data mining and time series prediction,” *Lecture Notes in Computer Science*, vol. 3512, pp. 758–770, 2005, doi: 10.1007/11494669_93/COVER.

[87] H. Hotelling, “Analysis of a complex of statistical variables into principal components,” *J Educ Psychol*, vol. 24, no. 6, pp. 417–441, Sep. 1933, doi: 10.1037/H0071325.

[88] M. Lovrić *et al.*, “Should We Embed in Chemistry? A Comparison of Unsupervised Transfer Learning with PCA, UMAP, and VAE on Molecular Fingerprints,” *Pharmaceuticals 2021, Vol. 14, Page 758*, vol. 14, no. 8, p. 758, Aug. 2021, doi: 10.3390/PH14080758.

[89] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb. 2018, doi: 10.48550/arxiv.1802.03426.

[90] T. S. Madhulatha, “An Overview on Clustering Methods,” *IOSR Journal of Engineering*, vol. 02, no. 04, pp. 719–725, May 2012, doi: 10.48550/arxiv.1205.1117.

[91] P. Makwana, T. M. Kodinariya, and P. R. Makwana, “Review on Determining of Cluster in K-means Clustering Review on determining number of Cluster in K-Means Clustering,” *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 6, 2013, Accessed: Nov. 06, 2022. [Online]. Available: https://www.researchgate.net/publication/313554124

[92] A. Meyer *et al.*, “Machine learning for real-time prediction of complications in critical care: a retrospective study,” *Lancet Respir Med*, vol. 6, no. 12, pp. 905–914, Dec. 2018, doi: 10.1016/S2213-2600(18)30300-X.

[93] J. L. Vincent *et al.*, “Assessment of the worldwide burden of critical illness: the Intensive Care Over Nations (ICON) audit,” *Lancet Respir Med*, vol. 2, no. 5, pp. 380–386, May 2014, doi: 10.1016/S2213-2600(14)70061-X.

[94] I. Yoo *et al.*, “Data Mining in Healthcare and Biomedicine: A Survey of the Literature,” *Journal of Medical Systems 2011 36:4*, vol. 36, no. 4, pp. 2431–2448, May 2011, doi: 10.1007/S10916-011-9710-5.

[95] P. B. Jensen, L. J. Jensen, and S. Brunak, “Mining electronic health records: towards better research applications and clinical care,” *Nature Reviews Genetics 2012 13:6*, vol. 13, no. 6, pp. 395–405, May 2012, doi: 10.1038/nrg3208.

[96] I. W. M. Verburg *et al.*, “Which Models Can I Use to Predict Adult ICU Length of Stay? A Systematic Review*,” *Crit Care Med*, vol. 45, no. 2, pp. e222–e231, Feb. 2017, doi: 10.1097/CCM.0000000000002054.

[97] M. Kılıç, N. Yüzkat, C. Soyalp, and N. Gülhaş, “Cost Analysis on Intensive Care Unit Costs Based on the Length of Stay,” *Turk J Anaesthesiol Reanim*, vol. 47, no. 2, p. 142, Apr. 2019, doi: 10.5152/TJAR.2019.80445.

[98] C. Herman, W. Karolak, A. M. Yip, K. J. Buth, A. Hassan, and J. F. Légarè, “Predicting prolonged intensive care unit length of stay in patients undergoing coronary artery bypass surgery – development of an entirely preoperative scorecard,” *Interact Cardiovasc Thorac Surg*, vol. 9, no. 4, pp. 654–658, Oct. 2009, doi: 10.1510/ICVTS.2008.199521.

[99] I. W. M. Verburg, N. F. De Keizer, E. De Jonge, and N. Peek, “Comparison of Regression Methods for Modeling Intensive Care Length of Stay,” *PLoS One*, vol. 9, no. 10, p. e109684, Oct. 2014, doi: 10.1371/JOURNAL.PONE.0109684.

[100] K. Feldman, L. Faust, X. Wu, C. Huang, and N. V. Chawla, “Beyond volume: The impact of complex healthcare data on the machine learning pipeline,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10344 LNAI, pp. 150–169, 2017, doi: 10.1007/978-3-319-69775-8_9/COVER/.

[101] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, “The eICU collaborative research database, a freely available multi-center database for critical care research,” *Sci Data*, vol. 5, Sep. 2018, doi: 10.1038/sdata.2018.178.

[102] L. L. McGhie, “Health Insurance Portability and Accountability Act (HIPAA),” in *Encyclopedia of Information Assurance*, CRC Press, 2010, pp. 1299–1309. doi: 10.1081/e-eia-120046838.

[103] B. Moody and G. Moody, “MIMIC-III Waveform Database v1.0.” https://physionet.org/content/mimic3wdb/1.0/ (accessed Aug. 25, 2020).

[104] B. Moody and G. Moody, “MIMIC-III Waveform Database Matched Subset v1.0.” https://physionet.org/content/mimic3wdb-matched/1.0/ (accessed Aug. 25, 2020).

[105] P. J. Thoral *et al.*, “Explainable Machine Learning on AmsterdamUMCdb for ICU Discharge Decision Support: Uniting Intensivists and Data Scientists,” *Crit Care Explor*, vol. 3, no. 9, p. e0529, Sep. 2021, doi: 10.1097/CCE.0000000000000529.

[106] “ICD - ICD-10-CM - International Classification of Diseases,(ICD-10-CM/PCS Transition.” https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm (accessed Nov. 20, 2022).

[107] H. Meye, “Coding Complexity: US Health Care Gets Ready For The Coming Of ICD-10,” *https://doi.org/10.1377/hlthaff.2011.0319*, vol. 30, no. 5, pp. 968–974, Aug. 2017, doi: 10.1377/HLTHAFF.2011.0319.

[108] E. S. Fisher *et al.*, “The accuracy of Medicare’s hospital claims data: progress has been made, but problems remain.,” *https://doi.org/10.2105/AJPH.82.2.243*, vol. 82, no. 2, pp. 243–248, Aug. 2011, doi: 10.2105/AJPH.82.2.243.

[109] C. Raina MacIntyre, M. J. Ackland, E. J. Chandraraj, and J. E. Pilla, “Accuracy of ICD–9–CM codes in hospital morbidity data, Victoria: implications for public health research,” *Aust N Z J Public Health*, vol. 21, no. 5, pp. 477–482, Aug. 1997, doi: 10.1111/J.1467-842X.1997.TB01738.X.

[110] S. Nuthakki, S. Neela, J. W. Gichoya, and S. Purkayastha, “Natural language processing of MIMIC-III clinical notes for identifying diagnosis and procedures with neural networks”.

[111] K. Xu *et al.*, “Multimodal Machine Learning for Automated ICD Coding,” *Proceedings of Machine Learning Research*, vol. 106. PMLR, pp. 197–215, Oct. 28, 2019. Accessed: Nov. 20, 2022. [Online]. Available: https://proceedings.mlr.press/v106/xu19a.html

[112] “Data Mining: Concepts and Techniques - Jiawei Han, Jian Pei, Micheline Kamber - Google Books.” https://books.google.co.uk/books?hl=en&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&dq=Han,+J.,+Pei,+J.,+Kamber,+M.:+Data+mining:+concepts+and+techniques.+Elsevier+(2011)&ots=tAIx2_oz2Z&sig=5OkqVYpd043biGBL7xz9e_gewmE#v=onepage&q=Han%2C J.%2C Pei%2C J.%2C Kamber%2C M.%3A Data mining%3A concepts and techniques. Elsevier (2011)&f=false (accessed Jul. 14, 2022).

[113] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, “Data preprocessing in predictive data mining,” *Knowl Eng Rev*, vol. 34, 2019, doi: 10.1017/S026988891800036X.

[114] C. B. Gokulnath and S. P. Shantharajah, “An optimized feature selection based on genetic approach and support vector machine for heart disease,” *Cluster Computing 2018 22:6*, vol. 22, no. 6, pp. 14777–14787, Mar. 2018, doi: 10.1007/S10586-018-2416-4.

[115] S. Gupta and R. R. Sedamkar, “Machine Learning for Healthcare: Introduction,” pp. 1–25, 2020, doi: 10.1007/978-3-030-40850-3_1.

[116] J. Zhao, P. Papapetrou, L. Asker, and H. Boström, “Learning from heterogeneous temporal data in electronic health records,” *J Biomed Inform*, vol. 65, pp. 105–119, Jan. 2017, doi: 10.1016/J.JBI.2016.11.006.

[117] S. Ortega-Martorell, M. Pieroni, B. W. Johnston, I. Olier, and I. D. Welters, “Development of a Risk Prediction Model for New Episodes of Atrial Fibrillation in Medical-Surgical Critically Ill Patients Using the AmsterdamUMCdb,” *Front Cardiovasc Med*, vol. 0, p. 1259, May 2022, doi: 10.3389/FCVM.2022.897709.

[118] M. B. Mohammed, H. S. Zulkafli, M. B. Adam, N. Ali, and I. A. Baba, “Comparison of five imputation methods in handling missing data in a continuous frequency table,” *AIP Conf Proc*, vol. 2355, no. 1, p. 40006, May 2021, doi: 10.1063/5.0053286/684231.

[119] J. Huang, C. Osorio, and L. W. Sy, “An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes,” *Comput Methods Programs Biomed*, vol. 177, pp. 141–153, Aug. 2019, doi: 10.1016/J.CMPB.2019.05.024.

[120] A. P. Kurniati, E. Rojas, D. Hogg, G. Hall, and O. A. Johnson, “The assessment of data quality issues for process mining in healthcare using Medical Information Mart for Intensive Care III, a freely available e-health record database,” *Health Informatics J*, vol. 25, no. 4, pp. 1878–

1893, Dec. 2019, doi: 10.1177/1460458218810760/ASSET/IMAGES/LARGE/10.1177_1460458218810760-FIG2.JPEG.

[121] N. Nawalkar, V. Z. Attar, and S. P. Kalamkar, "Automated ICD-9 Medical Code Assignment from Given Free Text Using Deep Learning Approach," *Lecture Notes in Networks and Systems*, vol. 318, pp. 317–327, 2022, doi: 10.1007/978-981-16-5689-7_28/COVER.

[122] T. Gangavarapu, A. Jayasimha, G. S. Krishnan, and S. Sowmya Kamath, "Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes," *Knowl Based Syst*, vol. 190, p. 105321, Feb. 2020, doi: 10.1016/J.KNOSYS.2019.105321.

[123] C. C. Diamond, F. Mostashari, and C. Shirky, "Collecting And Sharing Data For Population Health: A New Paradigm," *https://doi.org/10.1377/hlthaff.28.2.454*, vol. 28, no. 2, pp. 454–466, Aug. 2017, doi: 10.1377/HLTHAFF.28.2.454.

[124] R. Hillestad *et al.*, "Identity Crisis: An Examination of the Costs and Benefits of a Unique Patient Identifier for the U.S. Health Care System," 2008.

[125] Z. Harel *et al.*, "Rehospitalizations and Emergency Department Visits after Hospital Discharge in Patients Receiving Maintenance Hemodialysis," *J Am Soc Nephrol*, vol. 26, no. 12, pp. 3141–3150, Dec. 2015, doi: 10.1681/ASN.2014060614.

[126] D. Hosmer, *Applied logistic regression*. Hoboken, New Jersey: Wiley, 2013.

[127] L. Breiman, "Random Forests," *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[128] J. H. Friedman, "Stochastic gradient boosting," *Comput Stat Data Anal*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[129] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif Intell*, vol. 97, no. 1–2, pp. 273–324, Dec. 1997, doi: 10.1016/S0004-3702(97)00043-X.

[130] D. Benrimoh *et al.*, "Editorial: ML and AI Safety, Effectiveness and Explainability in Healthcare," *Front Big Data*, vol. 4, Jul. 2021, doi: 10.3389/FDATA.2021.727856.

[131] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," *Entropy 2021, Vol. 23, Page 18*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/E23010018.

[132] V. Danilatou *et al.*, "Outcome Prediction in Critically-Ill Patients with Venous Thromboembolism and/or Cancer Using Machine Learning Algorithms: External Validation and Comparison with Scoring Systems," *International Journal of Molecular Sciences 2022, Vol. 23, Page 7132*, vol. 23, no. 13, p. 7132, Jun. 2022, doi: 10.3390/IJMS23137132.

[133] Z. Sun, S. Peng, Y. Yang, X. Wang, and F. Li, "A General Fine-tuned Transfer Learning Model for Predicting Clinical Task Acrossing Diverse EHRs Datasets," *Proceedings - 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2019*, pp. 490–495, Nov. 2019, doi: 10.1109/BIBM47256.2019.8983098.

[134] W. Liu *et al.*, "A Simple Weaning Model Based on Interpretable Machine Learning Algorithm for Patients With Sepsis: A Research of MIMIC-IV and eICU Databases," *Front Med (Lausanne)*, vol. 8, p. 814566, Jan. 2022, doi: 10.3389/FMED.2021.814566/FULL.

[135] M. Pieroni, I. Olier, S. Ortega-Martorell, B. W. Johnston, and I. D. Welters, "In-Hospital Mortality of Sepsis Differs Depending on the Origin of Infection: An Investigation of

Predisposing Factors," *Front Med (Lausanne)*, vol. 9, p. 2041, Jul. 2022, doi: 10.3389/FMED.2022.915224/BIBTEX.

[136] M. Singer *et al.*, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, p. 801, Feb. 2016, doi: 10.1001/jama.2016.0287.

[137] E. Volakli, C. Spies, A. Michalopoulos, A. J. Groeneveld, Y. Sakr, and J.-L. Vincent, "Infections of respiratory or abdominal origin in ICU patients: what are the differences?," *Crit Care*, vol. 14, no. 2, p. R32, Mar. 2010, doi: 10.1186/cc8909.

[138] J. A. Stortz *et al.*, "Phenotypic heterogeneity by site of infection in surgical sepsis: a prospective longitudinal study," *Crit Care*, vol. 24, no. 1, p. 203, Dec. 2020, doi: 10.1186/s13054-020-02917-3.

[139] C. W. Seymour *et al.*, "Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis," *JAMA*, vol. 321, no. 20, p. 2003, May 2019, doi: 10.1001/jama.2019.5791.

[140] M. Bauer, H. Gerlach, T. Vogelmann, F. Preissing, J. Stiefel, and D. Adam, "Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019—results from a systematic review and meta-analysis," *Crit Care*, vol. 24, no. 1, p. 239, Dec. 2020, doi: 10.1186/s13054-020-02950-2.

[141] C. W. Seymour *et al.*, "Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis," *JAMA*, vol. 321, no. 20, p. 2003, May 2019, doi: 10.1001/jama.2019.5791.

[142] J. C. Marshall, "Why have clinical trials in sepsis failed?," *Trends Mol Med*, vol. 20, no. 4, pp. 195–203, 2014, doi: 10.1016/J.MOLMED.2014.01.007.

[143] C. Nedeva, J. Menassa, and H. Puthalakath, "Sepsis: Inflammation is a necessary evil," *Front Cell Dev Biol*, vol. 7, no. JUN, p. 108, 2019, doi: 10.3389/FCELL.2019.00108/BIBTEX.

[144] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU collaborative research database, a freely available multi-center database for critical care research," *Sci Data*, vol. 5, Sep. 2018, doi: 10.1038/sdata.2018.178.

[145] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Sci Data*, vol. 5, no. 1, p. 180178, Dec. 2018, doi: 10.1038/sdata.2018.178.

[146] G. Kong, K. Lin, and Y. Hu, "Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU," *BMC Med Inform Decis Mak*, vol. 20, no. 1, pp. 1–10, Oct. 2020, doi: 10.1186/S12911-020-01271-2/FIGURES/1.

[147] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, "Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients," *Crit Care Med*, vol. 34, no. 5, pp. 1297–1310, 2006, doi: 10.1097/01.CCM.0000215112.84523.F0.

[148] D. J. Hand, "Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis by Geoff Cumming," *International Statistical Review*, vol. 80, no. 2, pp. 344–345, Aug. 2012, doi: 10.1111/J.1751-5823.2012.00187_26.X.

[149] L. G. Halsey, D. Curran-Everett, S. L. Vowler, and G. B. Drummond, "The fickle P value generates irreproducible results," *Nature Methods 2015 12:3*, vol. 12, no. 3, pp. 179–185, Feb. 2015, doi: 10.1038/nmeth.3288.

[150] S. N. Goodman, “Toward evidence-based medical statistics. 2: The Bayes factor,” *Ann Intern Med*, vol. 130, no. 12, pp. 1005–1013, Jun. 1999, doi: 10.7326/0003-4819-130-12-199906150-00019.

[151] T. M. Khoshgoftaar, K. Gao, A. Napolitano, and R. Wald, “A comparative study of iterative and non-iterative feature selection techniques for software defect prediction,” *Information Systems Frontiers*, vol. 16, no. 5, pp. 801–822, Oct. 2014, doi: 10.1007/S10796-013-9430-0/METRICS.

[152] J. L. Vincent *et al.*, “The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine,” *Intensive Care Med*, vol. 22, no. 7, pp. 707–710, 1996, doi: 10.1007/BF01709751.

[153] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” 2009, doi: 10.1007/978-0-387-84858-7.

[154] J. L. Vincent and R. Moreno, “Clinical review: Scoring systems in the critically ill,” *Critical Care*, vol. 14, no. 2. BioMed Central, p. 207, Mar. 2010. doi: 10.1186/cc8204.

[155] W. P. T. M. van Doorn *et al.*, “A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis,” *PLoS One*, vol. 16, no. 1, p. e0245157, Jan. 2021, doi: 10.1371/journal.pone.0245157.

[156] F. Daviaud *et al.*, “Timing and causes of death in septic shock,” *Ann Intensive Care*, vol. 5, no. 1, pp. 1–9, Dec. 2015, doi: 10.1186/S13613-015-0058-8/FIGURES/4.

[157] R. S. Hotchkiss, G. Monneret, and D. Payen, “Sepsis-induced immunosuppression: From cellular dysfunctions to immunotherapy,” *Nat Rev Immunol*, vol. 13, no. 12, pp. 862–874, Dec. 2013, doi: 10.1038/NRI3552.

[158] H. C. Prescott, C. S. Calfee, B. Taylor Thompson, D. C. Angus, and V. X. Liu, “Toward Smarter Lumping and Smarter Splitting: Rethinking Strategies for Sepsis and Acute Respiratory Distress Syndrome Clinical Trial Design,” *Am J Respir Crit Care Med*, vol. 194, no. 2, pp. 147–155, Jul. 2016, doi: 10.1164/RCCM.201512-2544CP.

[159] K. M. Demerle *et al.*, “Sepsis Subclasses: A Framework for Development and Interpretation,” *Crit Care Med*, vol. 49, no. 5, pp. 748–759, 2021, doi: 10.1097/CCM.0000000000004842.

[160] E. E. Davenport *et al.*, “Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study,” *Lancet Respir Med*, vol. 4, no. 4, pp. 259–271, Apr. 2016, doi: 10.1016/S2213-2600(16)00046-1.

[161] T. E. Sweeney *et al.*, “Unsupervised Analysis of Transcriptomics in Bacterial Sepsis Across Multiple Datasets Reveals Three Robust Clusters,” *Crit Care Med*, vol. 46, no. 6, pp. 915–925, 2018, doi: 10.1097/CCM.0000000000003084.

[162] B. P. Scicluna *et al.*, “Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study,” *Lancet Respir Med*, vol. 5, no. 10, pp. 816–826, Oct. 2017, doi: 10.1016/S2213-2600(17)30294-1.

[163] N. L. Stanski and H. R. Wong, “Prognostic and predictive enrichment in sepsis,” *Nat Rev Nephrol*, vol. 16, no. 1, pp. 20–31, Jan. 2020, doi: 10.1038/S41581-019-0199-3.

[164] Z. Zhang, G. Zhang, H. Goyal, L. Mo, and Y. Hong, “Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: A latent profile analysis 11 Medical and Health Sciences 1103 Clinical Sciences,” *Crit Care*, vol. 22, no. 1, pp. 1–11, Dec. 2018, doi: 10.1186/S13054-018-2279-3/FIGURES/2.

[165] B. Gårdlund, N. O. Dmitrieva, C. F. Pieper, S. Finfer, J. C. Marshall, and B. Taylor Thompson, "Six subphenotypes in septic shock: Latent class analysis of the PROWESS Shock study," *J Crit Care*, vol. 47, pp. 70–79, Oct. 2018, doi: 10.1016/J.JCRC.2018.06.012.

[166] S. v. Bhavani, K. A. Carey, E. R. Gilbert, M. Afshar, P. A. Verhoef, and M. M. Churpek, "Identifying Novel Sepsis Subphenotypes Using Temperature Trajectories," *Am J Respir Crit Care Med*, vol. 200, no. 3, pp. 327–335, Aug. 2019, doi: 10.1164/RCCM.201806-1197OC.

[167] D. Wang *et al.*, "A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients," *Front Public Health*, vol. 9, p. 1534, Oct. 2021, doi: 10.3389/FPUBH.2021.754348/BIBTEX.

[168] C. A. Motzkus and R. Luckmann, "Does Infection Site Matter? A Systematic Review of Infection Site Mortality in Sepsis," *J Intensive Care Med*, vol. 32, no. 8, pp. 473–479, Sep. 2017, doi: 10.1177/0885066615627778.

[169] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Adv Neural Inf Process Syst*, vol. 2017-Decem, pp. 4766–4775, May 2017.

[170] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and Information Systems 2013 41:3*, vol. 41, no. 3, pp. 647–665, Aug. 2013, doi: 10.1007/S10115-013-0679-X.

[171] Z. Jones and F. Linder, "Exploratory Data Analysis using Random Forests *," in *Annual MPSA conference*, Apr. 2015, pp. 16–19.

[172] W. P. T. M. van Doorn *et al.*, "A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis," *PLoS One*, vol. 16, no. 1, p. e0245157, Jan. 2021, doi: 10.1371/journal.pone.0245157.

[173] N. Peiffer-Smadja *et al.*, "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications," *Clinical Microbiology and Infection*, vol. 26, no. 5. Elsevier B.V., pp. 584–595, May 2020. doi: 10.1016/j.cmi.2019.09.009.

[174] N. R. Panda, J. K. Pati, J. N. Mohanty, and R. Bhuyan, "A Review on Logistic Regression in Medical Research," *National Journal of Community Medicine*, vol. 13, no. 04, pp. 265–270, Apr. 2022, doi: 10.55489/NJCM.134202222.

[175] E. Christodoulou, J. Ma, G. S. Collins, E. W. Steyerberg, J. Y. Verbakel, and B. van Calster, "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," *J Clin Epidemiol*, vol. 110, pp. 12–22, Jun. 2019, doi: 10.1016/J.JCLINEPI.2019.02.004.

[176] S. Bruckert, B. Finzel, and U. Schmid, "The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions," *Front Artif Intell*, vol. 3, p. 75, Sep. 2020, doi: 10.3389/FRAI.2020.507973/BIBTEX.

[177] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, "Interpretability of machine learning-based prediction models in healthcare," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 10, no. 5, p. e1379, Sep. 2020, doi: 10.1002/WIDM.1379.

[178] D. v. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine Learning Interpretability: A Survey on Methods and Metrics," *Electronics 2019, Vol. 8, Page 832*, vol. 8, no. 8, p. 832, Jul. 2019, doi: 10.3390/ELECTRONICS8080832.

[179] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[180] M. A. Ahmad, C. Eckert, and A. Teredesai, “Interpretable Machine Learning in Healthcare,” 2018, doi: 10.1145/3233547.3233667.

[181] R. Elshawi, M. H. Al-Mallah, and S. Sakr, “On the interpretability of machine learning-based model for predicting hypertension,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, pp. 1–32, Jul. 2019, doi: 10.1186/S12911-019-0874-0/FIGURES/48.

[182] N. Wallace and D. Castro, “The Impact of the EU’s New Data Protection Regulation on AI,” 2018.

[183] Y. Zhang and Q. Yang, “An overview of multi-task learning,” *Natl Sci Rev*, vol. 5, no. 1, pp. 30–43, Jan. 2018, doi: 10.1093/nsr/nwx105.

[184] R. Caruana, L. Pratt, and S. Thrun, “Multitask Learning *,” Kluwer Academic Publishers, 1997.

[185] C. M. Bishop, *Pattern recognition and machine learning*, 3rd editio. New York, NY: Springer, 2006.

[186] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi, “The myth of generalisability in clinical research and machine learning in health care,” *Lancet Digit Health*, vol. 2, no. 9, pp. e489–e492, 2020, doi: 10.1016/S2589-7500(20)30186-2.

[187] K.-H. Thung and C.-Y. Wee, “A brief review on multi-task learning,” *Multimed Tools Appl*, vol. 77, no. 22, pp. 29705–29725, Nov. 2018, doi: 10.1007/s11042-018-6463-x.

[188] H. Yuan, I. Paskov, H. Paskov, A. J. González, and C. S. Leslie, “Multitask learning improves prediction of cancer drug sensitivity,” *Sci Rep*, vol. 6, no. August, pp. 1–11, 2016, doi: 10.1038/srep31619.

[189] N. Sadawi *et al.*, “Multi-task learning with a natural metric for quantitative structure activity relationship learning,” *J Cheminform*, vol. 11, no. 1, p. 68, Dec. 2019, doi: 10.1186/s13321-019-0392-1.

[190] W. Zhang *et al.*, “Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis,” *IEEE Trans Big Data*, vol. 6, no. 2, pp. 322–333, Jun. 2020, doi: 10.1109/TBDATA.2016.2573280.

[191] X. Wang *et al.*, “Cross-type biomedical named entity recognition with deep multi-task learning,” *Bioinformatics*, 2019, doi: 10.1093/bioinformatics/bty869.

[192] H. Suresh, J. J. Gong, and J. V Guttag, “Learning Tasks for Multitask Learningg: Heterogenous Patient Populations in the ICU,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2018, pp. 802–810. doi: 10.1145/3219819.3219930.

[193] J. Wiens, J. Guttag, E. Horvitz, B. M. Marlin, D. Page, and S. Saria, “Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach,” 2016.

[194] M. B. A. McDermott *et al.*, “A Comprehensive Evaluation of Multi-task Learning and Multi-task Pre-training on EHR Time-series Data,” *ArXiv*, Jul. 2020.

[195] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, “Multitask learning and benchmarking with clinical time series data,” *Sci Data*, vol. 6, no. 1, p. 96, Dec. 2019, doi: 10.1038/s41597-019-0103-9.

[196] G. Shmueli, “To Explain or to Predict?,” *Statistical Science*, vol. 25, no. 3, pp. 289–310, Aug. 2010, doi: 10.1214/10-STS330.

[197] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

[198] O. Day and T. M. Khoshgoftaar, “A survey on heterogeneous transfer learning,” *J Big Data*, vol. 4, no. 1, pp. 1–42, Dec. 2017, doi: 10.1186/s40537-017-0089-0.

[199] H. Midi, S. K. Sarkar, and S. Rana, “Collinearity diagnostics of binary logistic regression model,” *Journal of Interdisciplinary Mathematics*, vol. 13, no. 3, pp. 253–267, Jun. 2010, doi: 10.1080/09720502.2010.10700699.

[200] P. Liu, X. Qiu, and X. Huang, “Adversarial Multi-task Learning for Text Classification,” *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 1–10, Apr. 2017, doi: 10.48550/arxiv.1704.05742.

[201] Y. Zhang and Q. Yang, “An overview of multi-task learning,” *Natl Sci Rev*, vol. 5, no. 1, pp. 30–43, Jan. 2018, doi: 10.1093/nsr/nwx105.

[202] K.-H. Thung and C.-Y. Wee, “A brief review on multi-task learning,” *Multimed Tools Appl*, vol. 77, no. 22, pp. 29705–29725, Nov. 2018, doi: 10.1007/s11042-018-6463-x.

[203] Y. Xu, J. Ma, A. Liaw, R. P. Sheridan, and V. Svetnik, “Demystifying Multi-Task Deep Neural Networks for Quantitative Structure-Activity Relationships,” *J Chem Inf Model*, p. acs.jcim.7b00087, 2017, doi: 10.1021/acs.jcim.7b00087.

[204] P. G. McCabe, S. Ortega-Martorell, and I. Olier, “Benchmarking multi-task learning in predictive models for drug discovery,” in *Proceedings of the International Joint Conference on Neural Networks*, 2019. doi: 10.1109/IJCNN.2019.8852074.

[205] N. Sadawi *et al.*, “Multi-task learning with a natural metric for quantitative structure activity relationship learning,” *J Cheminform*, vol. 11, no. 1, p. 68, Dec. 2019, doi: 10.1186/s13321-019-0392-1.

[206] M. Singer *et al.*, “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3),” *JAMA*, vol. 315, no. 8, p. 801, Feb. 2016, doi: 10.1001/jama.2016.0287.

[207] S. Kanji, D. R. Williamson, B. M. Yaghchi, M. Albert, and L. McIntyre, “Epidemiology and management of atrial fibrillation in medical and noncardiac surgical adult intensive care unit patients,” *J Crit Care*, vol. 27, no. 3, pp. 326.e1-326.e8, Jun. 2012, doi: 10.1016/J.JCRC.2011.10.011.

[208] E. Y. Ding *et al.*, “Novel Method of Atrial Fibrillation Case Identification and Burden Estimation Using the MIMIC-III Electronic Health Data Set,” *J Intensive Care Med*, vol. 34, no. 10, pp. 851–857, Oct. 2019, doi: 10.1177/0885066619866172.

[209] A. J. Walkey *et al.*, “Atrial fibrillation among Medicare beneficiaries hospitalized with sepsis: incidence and risk factors,” *Am Heart J*, vol. 165, no. 6, 2013, doi: 10.1016/J.AHJ.2013.03.020.

[210] N. A. Bosch, J. Cimini, and A. J. Walkey, “Atrial Fibrillation in the ICU.,” *Chest*, vol. 154, no. 6, pp. 1424–1434, Dec. 2018, doi: 10.1016/j.chest.2018.03.040.

[211] S. Sibley and J. Muscedere, “New-onset atrial fibrillation in critically ill patients,” *Canadian Respiratory Journal : Journal of the Canadian Thoracic Society*, vol. 22, no. 3, p. 179, May 2015, doi: 10.1155/2015/394961.

[212] S. A. Christian, C. Schorr, L. Ferchau, M. E. Jarbrink, J. E. Parrillo, and D. R. Gerber, “Clinical characteristics and outcomes of septic patients with new-onset atrial fibrillation,” *J Crit Care*, vol. 23, no. 4, pp. 532–536, Dec. 2008, doi: 10.1016/J.JCRC.2007.09.005.

[213] A. J. Walkey, R. S. Wiener, J. M. Ghobrial, L. H. Curtis, and E. J. Benjamin, "Incident Stroke and Mortality Associated with New-onset Atrial Fibrillation in Patients Hospitalized with Severe Sepsis," *JAMA : the journal of the American Medical Association*, vol. 306, no. 20, p. 2248, Nov. 2011, doi: 10.1001/JAMA.2011.1615.

[214] C. Guenancia *et al.*, "Incidence and Predictors of New-Onset Atrial Fibrillation in Septic Shock Patients in a Medical ICU: Data from 7-Day Holter ECG Monitoring," *PLoS One*, vol. 10, no. 5, May 2015, doi: 10.1371/JOURNAL.PONE.0127168.

[215] "The Impact of the MIT-BIH Arrhythmia Database History, Lessons Learned, and Its Influence on Current and Future Databases".

[216] S. K. Bashar, M. B. Hossain, E. Ding, A. J. Walkey, D. D. McManus, and K. H. Chon, "Atrial Fibrillation Detection during Sepsis: Study on MIMIC III ICU Data," *IEEE J Biomed Health Inform*, vol. 24, no. 11, pp. 3124–3135, Nov. 2020, doi: 10.1109/JBHI.2020.2995139.

[217] S. Khairul Bashar *et al.*, "Noise Detection in Electrocardiogram Signals for Intensive Care Unit Patients", doi: 10.1109/access.2019.2926199.

[218] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG signal quality during arrhythmia and its application to false alarm reduction," *IEEE Trans Biomed Eng*, vol. 60, no. 6, pp. 1660–1666, 2013, doi: 10.1109/TBME.2013.2240452.

[219] G. D. Clifford, F. Azuaje, P. E. Mcsharry, and A. House Boston|london, "Advanced Methods and Tools for ECG Data Analysis".

[220] G. Lu *et al.*, "Removing ECG noise from surface EMG signals using adaptive filtering," *Neurosci Lett*, vol. 462, no. 1, pp. 14–19, Sep. 2009, doi: 10.1016/J.NEULET.2009.06.063.

[221] J. S. Paul, M. Ramasubba Reddy, and V. J. Kumar, "A transform domain SVD filter for suppression of muscle noise artefacts in exercise ECG's," *IEEE Trans Biomed Eng*, vol. 47, no. 5, pp. 654–663, 2000, doi: 10.1109/10.841337.

[222] F. Enseleit and F. Duru, "Long-term continuous external electrocardiographic recording: a review," *EP Europace*, vol. 8, no. 4, pp. 255–266, Apr. 2006, doi: 10.1093/EUROPACE/EUJ054.

[223] S. Vijayarangan, B. Murugesan, R. Vignesh, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 300–303, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176396.

[224] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization." pp. 2921–2929, 2016.

[225] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization." pp. 618–626, 2017.

[226] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, 2000, doi: 10.1161/01.CIR.101.23.E215.

[227] B. Tutuko *et al.*, "AFibNet: an implementation of atrial fibrillation detection with convolutional neural network," *BMC Med Inform Decis Mak*, vol. 21, no. 1, pp. 1–17, Dec. 2021, doi: 10.1186/S12911-021-01571-1/TABLES/11.

[228] J. Bergstra, J. B. Ca, and Y. B. Ca, “Random Search for Hyper-Parameter Optimization Yoshua Bengio,” *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012, doi: 10.5555/2188385.

[229] D. Kolar, D. Lisjak, M. Pajak?, and M. Gudlin, “Intelligent Fault Diagnosis of Rotary Machinery by Convolutional Neural Network with Automatic Hyper-Parameters Tuning Using Bayesian Optimization,” *Sensors 2021, Vol. 21, Page 2411*, vol. 21, no. 7, p. 2411, Mar. 2021, doi: 10.3390/S21072411.

[230] M. G. Ragab, S. J. Abdulkadir, N. Aziz, H. Alhussian, A. Bala, and A. Alqushaibi, “An Ensemble One Dimensional Convolutional Neural Network with Bayesian Optimization for Environmental Sound Classification,” *Applied Sciences 2021, Vol. 11, Page 4660*, vol. 11, no. 10, p. 4660, May 2021, doi: 10.3390/APP11104660.

[231] J. Wu, X. Y. Chen, H. Zhang, L. D. Xiong, H. Lei, and S. H. Deng, “Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization,” *Journal of Electronic Science and Technology*, vol. 17, no. 1, pp. 26–40, Mar. 2019, doi: 10.11989/JEST.1674-862X.80904120.

[232] K. Simonyan, K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” *In Workshop at International Conference on Learning Representations*, 2014.

[233] P. S. Parvatharaju, R. Doddaiah, T. Hartvigsen, and E. A. Rundensteiner, “Learning Saliency Maps to Explain Deep Time Series Classifiers,” *International Conference on Information and Knowledge Management, Proceedings*, pp. 1406–1415, Oct. 2021, doi: 10.1145/3459637.3482446.

[234] G. Chao, Y. Luo, and W. Ding, “Recent Advances in Supervised Dimension Reduction: A Survey,” *Machine Learning and Knowledge Extraction 2019, Vol. 1, Pages 341-358*, vol. 1, no. 1, pp. 341–358, Jan. 2019, doi: 10.3390/MAKE1010020.

[235] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” Feb. 2018, doi: 10.48550/arxiv.1802.03426.

[236] E. Becht *et al.*, “Dimensionality reduction for visualizing single-cell data using UMAP,” *Nature Biotechnology 2018 37:1*, vol. 37, no. 1, pp. 38–44, Dec. 2018, doi: 10.1038/nbt.4314.

[237] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip Rev Comput Stat*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/WICS.101.

[238] H. Abdi, “The Eigen-Decomposition: Eigenvalues and Eigenvectors,” *Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage*, 2007.

[239] S. Monti *et al.*, “Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data,” *Machine Learning 2003 52:1*, vol. 52, no. 1, pp. 91–118, Jul. 2003, doi: 10.1023/A:1023949509487.

[240] P. J. G. Lisboa, T. A. Etchells, I. H. Jarman, and S. J. Chambers, “Finding reproducible cluster partitions for the k-means algorithm,” *BMC Bioinformatics*, vol. 14, no. SUPPL.1, pp. 1–19, Jan. 2013, doi: 10.1186/1471-2105-14-S1-S8/TABLES/3.

[241] W. Choi *et al.*, “Comparison of Continuous ECG Monitoring by Wearable Patch Device and Conventional Telemonitoring Device,” *J Korean Med Sci*, vol. 35, no. 44, Nov. 2020, doi: 10.3346/JKMS.2020.35.E363.

[242] M. Adeniji *et al.*, “Prioritising electrocardiograms for manual review to improve the efficiency of atrial fibrillation screening,” *Proceedings of the Annual International Conference of the*

IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2022-July, pp. 3239–3242, 2022, doi: 10.1109/EMBC48229.2022.9871092.

[243] E. Svennberg *et al.*, "Safe automatic one-lead electrocardiogram analysis in screening for atrial fibrillation," *EP Europace*, vol. 19, no. 9, pp. 1449–1453, Sep. 2017, doi: 10.1093/EUROPACE/EUW286.

[244] I. Olier, S. Ortega-Martorell, M. Pieroni, and G. Y. H. Lip, "How machine learning is impacting research in atrial fibrillation: implications for risk prediction and future management," *Cardiovasc Res*, vol. 117, no. 7, pp. 1700–1717, Jun. 2021, doi: 10.1093/CVR/CVAB169.

[245] M. G., "A new method for detecting atrial fibrillation using R-R intervals," *Comput Cardiol*, pp. 227–230, 1983.

[246] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45–50, 2001, doi: 10.1109/51.932724.

[247] S. Vijayarangan, B. Murugesan, R. Vignesh, S. P. Preejith, J. Joseph, and M. Sivaprakasam, "Interpreting Deep Neural Networks for Single-Lead ECG Arrhythmia Classification," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 300–303, Jul. 2020, doi: 10.1109/EMBC44109.2020.9176396.

[248] E. H. Ph. D. John, "Guyton and Hall Textbook of Medical Physiology 12th Edition," 2011.

[249] M. Y. Jones, F. Deligianni, and J. Dalton, "Improving ECG Classification Interpretability using Saliency Maps," *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, pp. 675–682, Oct. 2020, doi: 10.1109/BIBE50027.2020.00114.

[250] P. S. Parvatharaju, R. Doddaiah, T. Hartvigsen, and E. A. Rundensteiner, "Learning Saliency Maps to Explain Deep Time Series Classifiers," *International Conference on Information and Knowledge Management, Proceedings*, pp. 1406–1415, Oct. 2021, doi: 10.1145/3459637.3482446.

[251] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine 2019 25:1*, vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

[252] A. H. Ribeiro *et al.*, "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–9, Apr. 2020, doi: 10.1038/s41467-020-15432-4.

[253] T. N. Pattalung and S. Chaichulee, "Comparison of machine learning algorithms for mortality prediction in intensive care patients on multi-center critical care databases," *IOP Conf Ser Mater Sci Eng*, vol. 1163, no. 1, p. 012027, Aug. 2021, doi: 10.1088/1757-899X/1163/1/012027.

[254] C. M. Sauer *et al.*, "Systematic Review and Comparison of Publicly Available ICU Data Sets-A Decision Guide for Clinicians and Data Scientists," *Crit Care Med*, vol. 50, no. 6, pp. e581–e588, Jun. 2022, doi: 10.1097/CCM.0000000000005517.

[255] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data 2019 6:1*, vol. 6, no. 1, pp. 1–18, Mar. 2017, doi: 10.1038/s41597-019-0103-9.

[256] L. Liu *et al.*, “Multi-task Learning via Adaptation to Similar Tasks for Mortality Prediction of Diverse Rare Diseases,” *AMIA Annual Symposium Proceedings*, vol. 2020, p. 763, 2020, Accessed: Dec. 11, 2022. [Online]. Available: /pmc/articles/PMC8075548/

[257] A. M. Alaa, “Machine Learning for Mortality Prediction in Critically Ill Patients: A Systematic Review and Meta-Analysis,” *Intensive Care Med*, vol. 45, no. 10, pp. 1320–1334, 2019.

[258] M. Pieroni, I. Olier, S. Ortega-Martorell, B. W. Johnston, and I. D. Welters, “In-Hospital Mortality of Sepsis Differs Depending on the Origin of Infection: An Investigation of Predisposing Factors,” *Front Med (Lausanne)*, vol. 0, p. 2041, Jul. 2022, doi: 10.3389/FMED.2022.915224.

[259] M. Komorowski and et al, “Personalized Prognostic Modeling for Critical Care Patients Using Sequentially Learned Deep Neural Networks,” *Sci Rep*, vol. 8, no. 1, pp. 1–13, 2018.

[260] A. J. McGregor and et al, “Machine Learning-Based Mortality Prediction in the ICU,” *PLoS One*, vol. 15, no. 5, 2020.

[261] A. B. Jena, D. Khullar, and I. Ho, “A predictive model for hospital readmission risk: poor performance on a UK population,” *J Gen Intern Med*, vol. 33, no. 4, pp. 489–491, 2018.

[262] L. Wynants *et al.*, “Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal,” *BMJ*, vol. 369, p. 29, Apr. 2020, doi: 10.1136/BMJ.M1328.

[263] A. S. Eltrass, M. B. Tayel, and A. I. Ammar, “Automated ECG multi-class classification system based on combining deep learning features with HRV and ECG measures,” *Neural Comput Appl*, vol. 34, no. 11, pp. 8755–8775, Jun. 2022, doi: 10.1007/S00521-022-06889-Z/TABLES/7.

[264] H. Lassoued and R. Ketata, “ECG multi-class classification using neural network as machine learning model,” *2018 International Conference on Advanced Systems and Electric Technologies, IC_ASET 2018*, pp. 473–478, Jun. 2018, doi: 10.1109/ASET.2018.8379901.

[265] P. E. Marik and A. M. Taeb, “SIRS, qSOFA and new sepsis definition,” *J Thorac Dis*, vol. 9, no. 4, p. 943, Apr. 2017, doi: 10.21037/JTD.2017.03.125.

[266] M. G. Davies and P. O. Hagen, “Systemic inflammatory response syndrome,” *British Journal of Surgery*, vol. 84, no. 7, pp. 920–935, Jul. 1997, doi: 10.1002/BJS.1800840707.

[267] M. E. Charlson, D. Carrozzino, J. Guidi, and C. Patierno, “Charlson Comorbidity Index: A Critical Review of Clinimetric Properties,” *Psychother Psychosom*, vol. 91, no. 1, pp. 8–35, Jan. 2022, doi: 10.1159/000521288.

[268] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila, “Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today’s critically ill patients,” *Crit Care Med*, vol. 34, no. 5, pp. 1297–1310, 2006, doi: 10.1097/01.CCM.0000215112.84523.F0.

9 Glossary

Abbreviation	Definition
AE	Autoencoder
AF	Atrial fibrillation
AI	Artificial intelligence
AIC	Akaike information criterion
ANN	Artificial neural network
APACHE	Acute physiology and chronic health evaluation
AUC	Area under the curve
AUC-ROC	Area under the receiver operating characteristic
AUMC	AmsterdamUMC or Amsterdam university medical centre
BIC	Bayesian information criterion
BO	Bayesian optimisation
BPM	Beats per minute
BSR	Backwards stepwise regression
CAE	Convolutional autoencoders
CART	Classification and regression trees
CCU	Critical Care Unit
CI	Confidence intervals
CM	Confusion matrix
CNN	Convolutional neural network
CV	Cardiovascular
DB	Database
DGPR	General data protection regulations
DL	Deep learning
DR	Dimensionality reduction
ECG	Electrocardiogram
EHR	Electronic health records
eICU	eICU collaborative research database
FNN	Feed-forward neural network
FSR	Forward stepwise regression
GAN	Generative adversarial networks
GBM	Gradient boosting machine
ICD	International classification of disease
ICU	Intensive care unit
LOS	Length of stay
LR	Logistic/linear regression (depending on context)
LSTM	Long-short term memory
MAE	Mean absolute error
MD	Medical doctor
MIMIC	Medical information mart for intensive care
ML	Machine learning
MLP	Multilayer perceptron
MSE	Mean squared error

MTL	Multi-task learn
NN	Neural networks
NSR	Normal sinus rhythm
OR	Odds ratios
PCA	Principal component analysis
PD	Partial dependencies
PRN	Partial response network
Prob	Probability
qSOFA	Quick sequential organ failure assessment
ReLu	Rectified linear unit
RF	Random forest
RMSE	Root mean square error
RNN	Recurrent neural networks
ROC	Area under the
SAPS	Simplified acute physiology score
SD	Standard deviation
SeCo	Separation and concordance
SHAP	Shapley additive explanations
SIRS	Systemic inflammatory response syndrome
SOFA	Sequential organ failure assessment
STL	Single-task learning
SVM	Support vector machines
TL	Transfer learning
UMAP	Uniform manifold approximation and projection