Individual Thesis

EXPLAINABLE AI IN THE ASSESMENT OF MORTALITY

Konstantinos Savva

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2026

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

EXPLAINABLE AI IN THE ASSESMENT OF MORTALITY

Konstantinos Savva

Supervisor Professor

Dr. Konstantinos Pattichis

The Individual Thesis was submitted in partial fulfilment of the requirements for obtaining the Computer Science degree of the Department of Computer Science of the University of Cyprus

May 2026

I declare that this dissertation and any accompanying software are my own original work, conducted in full compliance with the University of Cyprus Code of Conduct for Research (https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/), and with full accountability on my part for the accuracy, integrity, and validity of all content.

Acknowledgements

I would like to express my deepest gratitude to Dr. Konstantinos Pattichis, my supervising professor, for his invaluable guidance, continuous support, and insightful feedback throughout the development of this dissertation. His expertise and encouragement have been instrumental in shaping the direction and quality of this work.

I am also sincerely grateful to my family, whose patience, understanding, and unwavering support provided the foundation I needed to pursue my research. Their encouragement kept me motivated during challenging moments.

I would like to thank my friends, who provided both moral support and valuable discussions that enriched my perspective throughout this project, always followed with a beer and a laugh.

A special acknowledgment is due to my work colleagues, whose practical insights, collaboration, and shared experience significantly contributed to the feasibility and relevance of this research. Their support and engagement in discussions related to real-world applications of this work were particularly invaluable.

Abstract

This dissertation focuses on the development and evaluation of Explainable Artificial Intelligence models for mortality prediction in emergency department patients. The work aims to predict mortality risk while also providing interpretable explanations for the generated predictions using explainable AI techniques.

Initially, an extensive review of the relevant literature is presented, highlighting the importance of mortality prediction in emergency medicine, the growing use of machine learning in healthcare, and the need for transparent and interpretable artificial intelligence systems in clinical environments. The study also examines existing approaches related to mortality predictions and explainable AI in healthcare applications.

The methodology adopted in this dissertation is then described in detail. The work is based on the dataset “w6_mortality_multi.csv” and includes the development of three different machine learning models using Python and the scikit-learn framework. The first model focuses on binary mortality prediction within the mortality cohort, distinguishing between death within 24 hours and in-hospital death after 24 hours. The second model addresses the primary binary mortality task, predicting in-hospital death versus no in-hospital death, while the third model extends the problem into a multiclass classification setting with three outcomes: no death, death within 24 hours and in-hospital death after 24 hours.

In addition, the dissertation incorporates the TE2Rules explainability framework to extract interpretable decision rules from the developed models. This enables a better understanding of the reasoning behind the predictions and contributes to the transparency and trustworthiness of the proposed system.

The experimental evaluation demonstrated promising results, with the second model achieving the best overall predictive performance among the developed approaches. The results indicate that machine learning techniques, combined with explainable AI methods, can effectively support mortality risk assessment in emergency department settings.
Finally, the dissemination concludes with a discussion of the findings, the limitations of the current work, and directions for future research.

Contents

Chapter 1

Introduction

1.1 Introduction

Artificial Intelligence (AI) and machine learning technologies have become increasingly important in the healthcare sector during recent years. The ability of machine learning algorithms to analyse large volumes of data and identify complex patterns has led to the development of intelligent systems capable of supporting clinical decision-making processes. These technologies are applied in areas such as disease diagnosis, medical image analysis, patient monitoring and risk prediction. Among these applications, mortality prediction has emerged as an important research field due to its potential to assist healthcare professionals in identifying high-risk patients and improving patient management.

Emergency Departments (EDs) represent one of the most demanding and critical environments within healthcare systems. Medical personnel are often required to make rapid and accurate decisions under time pressure, while dealing with large number of patients with different medical conditions and varying levels of severity. In such situations, the ability to accurately predict mortality risk can provide valuable support to clinicians by helping prioritise patient care, allocate medical resources efficiently and improve overall treatment outcomes.

Although machine learning (ML) models have demonstrated strong predictive performance in healthcare applications, many advanced models operate as “black-box” systems. This means that while they may produce highly accurate predictions, the reasoning behind these predictions is often difficult for healthcare professionals to understand or interpret. As a result, Explainable Artificial Intelligence (XAI) has gained significant attention as an approach that aims to make machine learning systems more understandable, transparent and trustworthy for medical experts.

This dissertation focuses on the development and evaluation of explainable machine learning models for mortality prediction in emergency departments patients. The main objective of this work is not only to predict mortality risk, but also to provide interpretable explanations for the produced predictions.

To achieve this, multiple machine learning models were developed and evaluated using emergency department patient data. In addition, explainable artificial intelligence techniques were applied to improve the transparency and interpretability of the produced predictions. The experimental evaluation demonstrated promising results, highlighting the potential contribution of explainable AI in supporting clinical decision-making within healthcare environments.

The remainder of this dissertation presents the theoretical background, related literature, methodology, experimental evaluation and discussion of the developed explainable mortality prediction models, highlighting the potential contribution of explainable artificial intelligence in healthcare environments.

1.2 Dissertation Outline

This dissertation is organized into seven chapters, each focusing on a different aspect of the research and development process related to explainable artificial intelligence for mortality prediction.

Chapter 1 introduces the research topic, presents the motivation behind the study, and provides an overview of the objectives and structure of the dissertation.

Chapter 2 presents the theoretical background required for understanding the concepts used throughout this work. It includes an overview of artificial intelligence and machine learning in healthcare, mortality prediction systems, explainable artificial intelligence and evaluation metrics commonly used in machine learning applications.

Chapter 3 reviews the existing literature and related work associated with mortality prediction and explainable artificial intelligence in healthcare environments. The chapter discusses previous studies, machine learning approaches, explainability techniques and current challenges in the field.

Chapter 4 describes the methodology followed in this dissertation. It presents the dataset used in the experiments, the preprocessing procedures, the developed machine learning models and the explainability framework applied to extract interpretable decision rules from the generated predictions.

Chapter 5 presents the experimental results and evaluates the performance of the developed models. The chapter includes the comparison of the models, performance metrics and the analysis of the explainability results produces by the proposed approach.

Chapter 6 discusses the finding of the dissertation and analyses the strengths and limitations of the developed system. Furthermore, the chapter examines the importance of explainability and transparency in artificial intelligence systems designed for healthcare applications.

Finally, chapter 7 concludes the dissertation by summarizing the main contributions of the work and proposing possible directions for future research and further improvements of the developed models and explainability methods.

Chapter 2

Theoretical Background

2.1 Artificial Intelligence in Healthcare

Artificial Intelligence (AI) has become one of the most rapidly developing technologies in the healthcare sector. The increasing availability of medical data, combined with advances in in computational power and machine learning algorithms, has enabled the development of intelligent systems capable of supporting healthcare professionals in a wide range of clinical tasks. AI technologies are now applied in areas such as disease diagnosis, medical image analysis, patient monitoring, clinical decision support, drug discovery and risk prediction systems.

One of the most important applications of AI in healthcare is the development of Clinical Decision Support Systems (CDSSs). These systems assist healthcare professionals by analysing patient information and generating recommendations that can support medical decision. According to Salimparsa et al. (2025), CDSSs are computer-based systems designed to improve clinical decision-making and healthcare delivery using data-driven models and intelligent algorithms. [1]

Machine learning techniques have significantly contributed to the improvement of healthcare systems by enabling predictive analysis and patterns recognition from complex medical

datasets. AI-based systems have demonstrated promising performance in applications such as cancer detection, cardiovascular disease prediction, medical image interpretation and patient outcome forecasting. Furthermore, AI technologies can support healthcare professionals by improving diagnostic accuracy, reducing medical errors and assisting with early risk assessment. [2]

Despite the advantages offered by AI systems, their integration into healthcare environments also introduces important challenges. Many advanced machine learning models operate as "black-box" systems, where the reasoning behind a prediction or recommendation is not easily understood by clinicians. In healthcare environments, where decisions directly affect patient safety and treatment outcomes, transparency and interpretability are essential. Researchers have therefore emphasised the growing importance of Explainable Artificial Intelligence (XAI) as a means of improving trust, transparency and accountability in AI-driven healthcare applications [3]

Another important concern related to the use of AI in healthcare is the possibility of bias or unfair decision making. Recent studies have shown that AI systems may unintentionally produce biased predictions influenced by socioeconomic or demographic factors present in training data. Such issues highlight the need for careful evaluation, ethical considerations and responsible implementation of AI technologies in medical environments. [4]

Overall, artificial intelligence has the potential to significantly improve healthcare services and clinical decision-making processes. However, the successful integration of AI into real-world medical practice requires not only high predictive performance, but also transparency, interpretability, reliability and ethical responsibility. These requirements have contributed to the increasing research interest in explainable and trustworthy AI systems for healthcare applications.

2.2 Machine Learning and Classification Fundamentals

Machine learning (ML) is a subfield of Artificial Intelligence that focuses on the development of algorithms capable of learning patterns from data and making predictions or decisions without being explicitly programmed for every individual task. In healthcare environments, machine learning techniques are increasingly used to analyse large medical datasets, identify hidden relationships between variables and support clinical decision-making processes.

2.2.1 Supervised Learning

Supervised learning refers to the process where a machine learning model is trained using labelled data. In this approach, the dataset contains both input variables and known output labels, allowing the model to learn the relationship between the provided features and the expected outcome. The objective of the model is to identify patterns that can later be used to make predictions on previously unseen data.

In healthcare applications, supervised learning is frequently used for prediction and classification task such as disease diagnosis, mortality prediction and patient outcome forecasting. The model learns from historical patient data and attempts to predict future outcomes based on previously observed patterns.

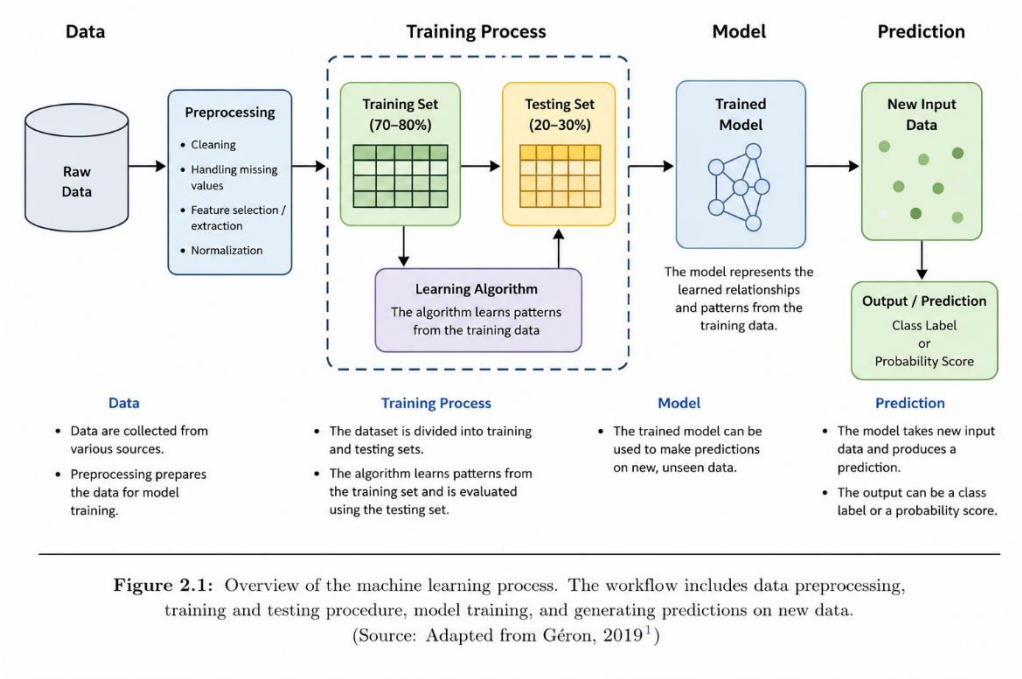


Figure 2.1: Overview of the machine learning process. The workflow includes data preprocessing, training and testing procedure, model training, and generating predictions on new data. (Source: Adapted from Géron, 2019[1])

Typically, a supervised learning process begins with the collection and preprocessing of data. The dataset is the divided into training and testing subsets. The training dataset is used to train the machine learning model, while the testing dataset is used to evaluate its performance and generalization ability.

2.2.2 Classification Problems

Classification is one of the most important supervised learning tasks and is widely applied in healthcare environments. In classification problems, the objective of the model is to assign input data into predefined categories or classes.

The two main types of classification problems:

- Binary Classification
- Multiclass Classification

Binary Classification:

Binary classification involves problems where the model predicts one of two possible outcomes. In healthcare applications, binary classification is commonly used for tasks such as predicting whether a patient has a disease or not, or whether a patient will survive or die.

In this dissertation, binary classification was applied to mortality prediction tasks, including the prediction of in-hospital death versus no in-hospital death.

Multiclass Classification:

Multiclass classification extends binary classification by allowing the prediction of more than two classes. In such problems, the model must determine which category among multiple possible outcomes best matches the input data.

In this dissertation, multiclass classification was used to distinguish between three different outcomes: no death, death within 24 hours and in-hospital death after 24 hours.

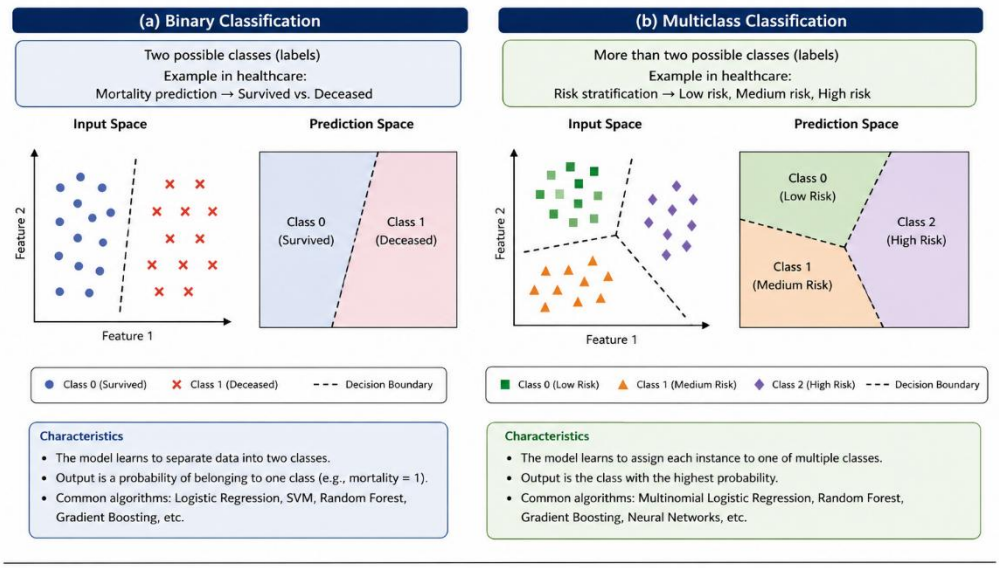


Figure 2.2: Comparison between binary and multiclass classification.
Binary classification involves two possible classes, while multiclass classification involves more than two classes.
(Source: Adapted from Hastie, Tibshirani and Friedman, 2009[1]; Géron, 2019[2])

2.2.3 Training and Testing Process

The performance of machine learning models depends heavily on the quality of the training process and the ability of the model to generalize to unseen data. During training, the model learns patterns and relationships from the training dataset. After the training phase is completed, the model is evaluated using a separate testing dataset to estimate its predictive performance.

One important challenge in machine learning is overfitting. Overfitting occurs when a model learns the training data too closely, including noise and irrelevant patterns, resulting in poor performance when applied to new unseen data. To reduce overfitting, techniques such as cross-validation, feature selection and regularization are commonly used.

Another important concept is generalization, which refers to the ability of a machine learning model to perform accurately on previously unseen data. A model with good generalization capability is more reliable and suitable for real-world applications.

2.2.4 Machine Learning Models Used in Healthcare

Several machine learning algorithms are commonly used in healthcare applications due to their predictive capabilities and ability to handle complex datasets. Among the most frequently used models are Logistic Regression, Decision Trees, Random Forests and Gradient Boosting algorithms.

Logistic Regression is a statistical classification algorithm widely used in medical prediction systems because of its simplicity and interpretability. Random Forest and Gradient Boosting models are ensemble learning methods capable of achieving high predictive performance by combining multiple decision trees.

Breiman (2001) introduced Random Forests as an ensemble learning technique that improves predictive accuracy and reduces overfitting by combining the outputs of multiple decision trees. Similarly, Friedman (2001) proposed Gradient Boosting as a method that sequentially builds decision trees to minimize prediction errors and improve model performance. [5] [6]

These machine learning techniques have demonstrated promising results in healthcare applications, particularly in prediction tasks involving large and complex medical datasets.

2.3 Mortality Prediction in Emergency Medicine

Mortality prediction in emergency medicine is a critical research area because accurate early identification of high-risk patients can significantly improve clinical outcomes and optimise resource allocation. Emergency Departments (EDs) are fast-paced environments where clinicians must make rapid decisions, often with incomplete patient information. Predictive systems that estimate the likelihood of in-hospital mortality help healthcare professionals prioritize treatment, allocate resources efficiently and potentially reduce preventable deaths. [7]

Traditional mortality prediction methods in EDs rely on scoring systems such as the Acute Physiology and Chronic Health Evaluation (APACHE), the Simplified Acute Physiology Score (SAPS) and the Emergency Severity Index (ESI). While these scoring systems are well-established, they often lack the ability to incorporate complex interactions among variables and may not capture all the relevant patterns present in large clinical datasets. Machine learning models, by contrast, can analyse multidimensional data, identify non-linear relationships and produce more accurate risk predictions. [8]

Recent studies have demonstrated that effectiveness or machine learning approaches for mortality prediction in emergency medicine. For example, Naemi et al. (2021) systematically reviewed multiple machine learning models, including Logistic Regression, Random Forests and Gradient Boosting algorithms, applied to ED datasets. Their analysis showed three-based ensemble methods and Gradient Boosting models often outperform traditional scoring systems in predictive accuracy, highlighting the potential of data-driven approaches in clinical decision-making. [9]

In addition to predictive accuracy, interpretability and explainability are essential for the adoption of mortality prediction models in clinical practice. Clinicians are more likely to trust and act upon model predictions when the rationale behind the prediction can be understood. Explainable AI frameworks, such as TE2Rules, SHAP and LIME, have been applied to mortality prediction to extract interpretable rules, identify key risk factors and provide insights into the decision-making process of complex machine learning models. [10]

Overall, mortality predictions in emergency medicine illustrates the practical value of combining machine learning with explainable AI. These approaches enable not only accurate

risk assessment but also transparent reasoning, thereby supporting clinicians in making timely, informed decisions that can improve patient outcomes.

2.4 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to a set of methods and techniques that make the predictions and decision-making process of machine learning models understandable to humans. In healthcare, the use of “black-box” models such as ensemble methods or deep learning can produce highly accurate predictions, but the lack of transparency often reduces clinicians trust in AI systems. For decisions that directly affect patient outcomes, interpretability and transparency are essential. [11]

XAI aims to bridge the gap by providing insight into how models generate predictions, highlighting which features or patient characteristics contribute most to a particular outcome. For example, in mortality prediction for emergency department patients, explainable models can reveal the key risk factors influencing the likelihood of in-hospital death, enabling clinician to a better understand and act on model recommendations.

Several techniques have been developed to make machine learning models more interpretable. Local interpretable methods, such as LIME (Ribeiro et al., 2016), provide explanations for individual predictions by approximating complex models with simpler, interpretable models locally around the input instance. SHAP (Shapley Additive exPlanations) values (Lundberg & Lee, 2017) quantify the contribution of each feature to a specific prediction, offering both global and local interpretability. Rule-based extraction methods, such as TE2Rules, transform model decisions into a set of human-readable rules, making them especially suitable for clinical environments where interpretability is critical. [12] [13]

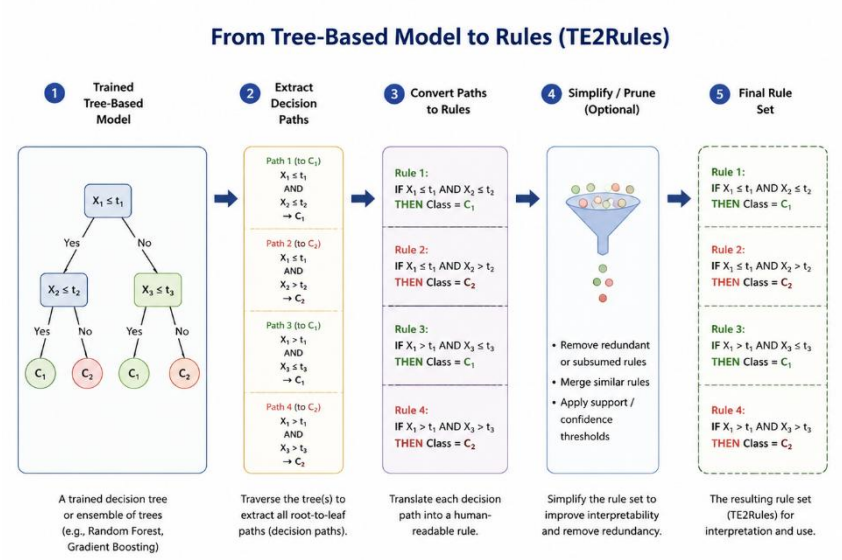


Figure 2.4: Rule generation process using TE2Rules.
A trained tree-based model is traversed to extract decision paths, which are converted into human-readable IF–THEN rules, optionally simplified, and returned as the final rule set.

The benefits of XAI in healthcare extend beyond trust. By understanding model behaviour, clinicians can identify potential biases in predictions, validate model decisions against established medical knowledge and ensure compliance with ethical standards. Moreover, interpretable AI systems can facilitate clinical adoption, as healthcare professionals are more likely to use models whose reasoning they can comprehend and justify to patients.

In the context of this dissertation, XAI serves as a fundamental component that complements mortality prediction models. By applying the TE2Rules framework to the developed models, interpretable rules are generated that explain why certain patients are predicted to be at high risk of mortality. This not only improves transparency but also provides actionable insights that can support clinical decision-making in emergency departments.

2.5 Evaluation Metrics

In machine learning, evaluation metrics are essential to quantify the performance of predictive models and to compare different algorithms. In the context of mortality prediction in emergency department patients, selecting appropriate metrics ensures that the models are not only accurate but also reliable for clinical decision-making.

2.5.1 Accuracy

Accuracy measures the proportion of correct predictions among all predictions:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Where TP is the true positives, TN is the true negatives, FP is the false positives and FN is the false negatives. While accuracy is easy to interpret, it may be misleading in imbalanced datasets where one class dominates (e.g. mortality is often less frequent than survival)

2.5.2 Precision and Recall

Precision evaluates the proportion of correct positive predictions:

$$Precision = \frac{(TP)}{(TP + FP)}$$

Recall (or sensitivity) evaluates the proportion of actual positives correctly identified:

$$Recall = \frac{TP}{(TP + FN)}$$

These metrics are particularly important in healthcare applications where false negatives (e.g. failing to predict a high-risk patient) may have severe consequences.

2.5.3 F1-Score

The F1-Score is the harmonic mean of precision and recall:

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

It provides a single measure that balances precision and recall, making it useful for evaluating models on imbalanced datasets.

2.5.4 Receiver Operating Characteristics (ROC) Curve and AUC

The ROC curve plots the true positive rate against the false positive rate at various classification thresholds. The Area Under the Curve (ROC-AUC) quantifies the model's ability to distinguish between classes:

- AUC = 1: Perfect classifier
- AUC = 0.5: Random guessing

ROC-AUC is widely used in medical applications because it summarizes performance across all thresholds.

2.5.5 Precision-Recall (PR) Curve and AUC

The PR curve plots precision against recall at different thresholds. PR-AUC is especially informative for imbalanced datasets, where ROC-AUC can sometimes overestimate model performance.

2.5.6 Specificity

Specificity (true negative rate) measures the proportion of actual negatives correctly identified:

$$Specificity = \frac{TN}{(TN + FP)}$$

In combination with sensitivity, it gives a complete picture of model performance.

Chapter 3

Literature Review

3.1 AI and Machine Learning for Mortality Prediction

Artificial Intelligence (AI) has become increasingly important in emergency medicine, particularly for the prediction of patient outcomes. Machine Learning models can analyse complex patient data to identify patterns that may indicate high risk of mortality, allowing clinicians to make more informed decisions in time-sensitive environments such as Emergency Departments (EDs).

Traditional approaches, such as scoring systems, are limited in their ability to model complex interactions between variables. Recent studies have shown that machine learning methods provide a significant improvement in predictive performance. Naemi et al. (2021) conducted a systematic review of machine learning techniques applied to mortality prediction in EDs, highlighting that tree-based models, gradient boosting and ensemble methods often outperform conventional scoring systems. [7]

Specific models such as Random Forests and Gradient Boosting have been applied successfully in clinical datasets, capturing non-linear relationships and interactions between patient features. Fransvea et al. (2022) demonstrated that explainable machine learning models could effectively predict post-surgical in-hospital mortality in elderly patients, combining predictive accuracy with clinically interpretable outputs. Similarly, Findik et al. (2026) showed that ensemble models could predict in-hospital mortality for patients with suspected pulmonary embolism, achieving high accuracy while maintaining interpretability through explainable AI techniques. [8] [10]

3.2 Explainable AI for Mortality Prediction

While machine learning models can achieve high accuracy, their “black-box” nature poses challenges in clinical adoption. Explainable AI (XAI) techniques provide insight into the reasoning of models, increasing trust and transparency. Okada et al. (2023) emphasized the importance of explainable AI in emergency medicine, noting that interpretable models enable clinicians to understand and verify predictions before acting on them. [11]

Common XAI methods include SHAP, LIME and rule-based extraction frameworks such as TE2Rules. These approaches help clinicians identify which patient features contribute most to predicted outcomes, allowing for informed decision-making. For example, Findik et al. (2026) applied TE2Rules to convert complex ensemble model outputs into a set of human-readable rules, highlighting the key factors influencing mortality risk. [10]

3.3 Existing Explainable Mortality Prediction Systems

Recent research has focused on combining predictive performance with interpretability to make mortality prediction models usable in clinical settings. The main goal of these systems is not only to accurately predict patient outcomes but also to provide explanations that clinicians can understand and trust.

Fransvea et al. (2022) developed an explainable machine learning model aimed at predicting post-surgical in-hospital mortality in elderly patients. Their approach utilized tree-based ensemble models, which were chosen for their strong predictive performance and inherent structure suitable for rule extraction. To enhance interpretability, the TE2Rules framework was applied to the trained models, generating a set of human-readable rules that describe how patient features contribute to the predicted risk. For example, the system could identify combinations of physiological measures, comorbidities and laboratory values that significantly increased mortality risk, providing actionable insights for clinicians. This study demonstrated that explainable models can simultaneously maintain high accuracy and offer clinically relevant explanations. [8]

Findik et al. (2026) applied a similar methodology to predict in-hospital mortality for emergency department patients with suspected pulmonary embolism. Using ensemble machine learning models combined with TE2Rules, their system achieved high level of predictive accuracy while converting complex decision logic of the models into interpretable IF-THEN

rules. These rules highlighted key patient factors, such as oxygen saturation, heart rate and prior medical history, which contributed most to the predicted mortality risk. The study emphasized that providing interpretable outputs allows clinicians to validate and understand predictions, which is crucial for adoption in fast-paced emergency settings. [10]

Naemi et al. (2021), in a systematic review of machine learning techniques for mortality prediction in emergency departments, highlighted that tree-based models and gradient boosting algorithms are among the most effective approaches for predictive performance. While the review focuses mainly on accuracy, it also noted that combining these models with explainability techniques significantly improves clinical usability. Rule-based extraction and feature importance methods were recommended as effective strategies to increase the trust in predictions. This reinforces the importance of interpretability paired with high accuracy performance for clinical applications. [7]

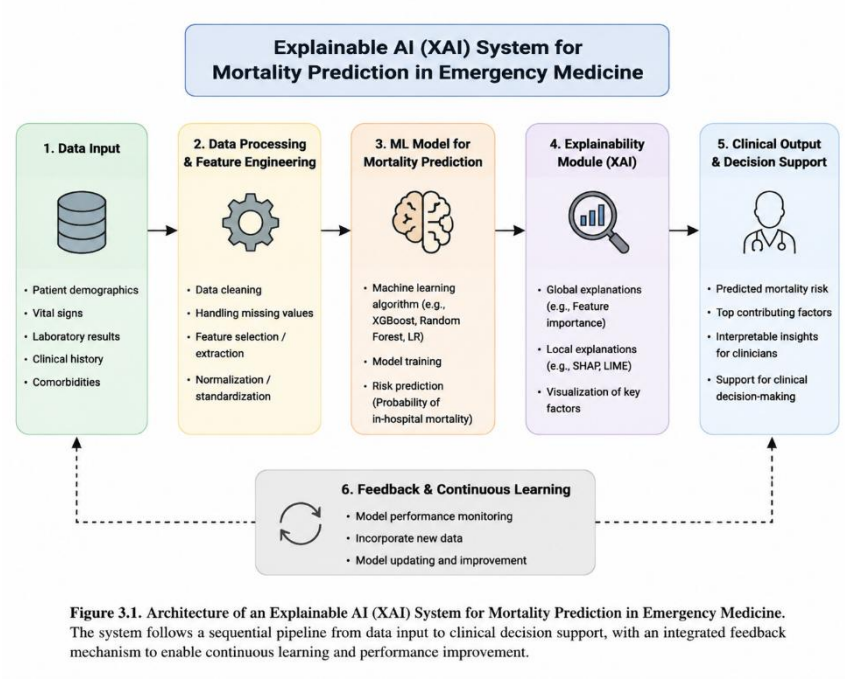


Figure 3.1. Architecture of an Explainable AI (XAI) System for Mortality Prediction in Emergency Medicine. The system follows a sequential pipeline from data input to clinical decision support, with an integrated feedback mechanism to enable continuous learning and performance improvement.

Collectively, these studies illustrate several key trends in the development of explainable mortality prediction systems:

1. Tree-based and ensemble models dominate predictive performance due to their ability to handle complex, non-linear relationships in patient data.
2. Rule extraction methods like TE2Rules are highly effective for translating model logic into human-readable explanations.
3. Explainability enhances clinical trust and adoption, as clinicians can understand, validate and act on model predictions.

4. Feature importance and interpretable outputs allow identification of key risk factors, supporting patient prioritization and intervention in emergency contexts.

These findings demonstrate that modern mortality prediction system in emergency medicine can achieve a balance between high predictive accuracy and interpretability, a critical factor for their practical adoption in clinical decision-making. By generating actionable and understandable explanations, these systems provide a pathway to integrating AI responsibly into patient care.

3.4 Research Gaps and Motivation for the Proposed Work

Despite the success of recent approaches, several limitations remain. Many models are evaluated on small or disease-specific datasets, limiting generalizability. Furthermore, the trade-off between predictive accuracy and interpretability is still an open challenge, particularly for high-stakes applications in emergency medicine.

This dissertation addresses these gaps by developing multiple machine learning models for ED mortality prediction, evaluating them and applying TE2Rules to generate interpretable rules. By doing so, it combines predictive performance with transparency, supporting clinical trust and adoption and providing actionable insights to guide patient care.

Chapter 4

Methodology and Experimental Setup

4.1 Overview

The mortality prediction pipeline in this dissertation was designed as a structured end-to-end workflow that integrates predictive modelling with post explainability using the TE2Rules framework. The pipeline operated on the tabular emergency department (ED) dataset, w6_mortality_multi.csv, and addresses three related prediction tasks:

1. Model 1: Binary mortality timing within patients who died in hospital (death within 24 hours vs later).
2. Model 2: Binary full-cohort-in-hospital mortality prediction (death vs no death).
3. Model 3: Multiclass mortality timing prediction (no death, death withing 24 hours, in-hospital death after 24 hours).

The goal of this pipeline is to combine accurate predictive modelling with interpretable, rule-based explanations to support clinical decision-making in emergency settings.

The workflow proceeds through several key stages:

1. Raw Data Input and Cohort Definition:
   The pipeline begins with ED encounter data containing demographic information, vital signs, laboratory results and outcome labels. Task-specific cohorts are created for each prediction objective.
2. Data Partitioning:
   The dataset is split at the patient level into training, validation and locked test sets to prevent information leakage, preserving class prevalence across splits.
3. Target Construction:
   Outcome variables are defined according to each model's task (binary or multiclass), producing feature matrixes and response vectors.
4. Feature Selection and Extraction Rules:
   Identifiers and outcome labels are excluded from the predictors to prevent leakage. Retained features include demographics, vital signs, laboratory results and missingness indicators.
5. Preprocessing:
   Separate pipelines are applied to numeric and categorical features using scikit-learn, including median imputation, standardization, most-frequent imputation and one-hot encoding.
6. Class Imbalance Handling:
   Training variants are created with different positive-to-negative sampling ratios to address rare-event mortality.

7. Predictive Model Development:
   Three machine learning algorithms—Logistic Regression, Random Forest, and Gradient Boosting—are trained and evaluated using cross-validation and standard metrics (ROC-AUC, PR-AUC, F1-score, etc.).
8. Model Selection and Threshold Optimization:
   The best performing model and threshold are selected based on validation performance to maximize predictive accuracy under class imbalance conditions.
9. Locked Test Evaluation:
   The final model is evaluated on the untouched test set, producing final performance metrics, confusion matrixes and bootstrap confidence intervals.
10. Explainable AI with TE2Rules:
    TE2Rules is applied to post hoc to tree-ensemble models to extract human-readable IF-THEN rules that approximate the model's positive prediction logic. For the multiclass model, one-vs-rest binary surrogate models are used for explainability.
11. Final Outputs:
    The pipeline produces processed datasets, trained predictive models, threshold-optimized classification outputs and TE2Rules explanation artifacts, supporting both reproducibility and interpretability.

In summary, this integrated workflow enables both accurate mortality prediction and transparent, interpretable explanations of model decisions, providing actionable insights for clinicians while maintaining methodological rigor.

4.2 Dataset Description

The dataset used in this study, w6_mortality_multi.csv, contains structured emergency department (ED) encounter data for adult patients. It comprises 424,952 ED encounters from 205,427 unique patients, indicating that some patients had multiple encounters. Each row corresponds to a single ED stay, with unique identifier stay_id assigned to each eanconter. The mean number of encounters per patient is approximately 2.07 and 69,928 patients have more than one recorded encounter.

The dataset includes both administrative identifiers and clinical variables. Identifier-related fields are stay_id, subject_id, hadm_id, ed_intime and ed_outtime. After removing identifiers and mortality outcome labels, the datasets contain 49 candidate input features, including both categorical and numerical variables:

- Categorical variables (4): gender, arrival_transport, race and chiefcomplaint.

- Numerical variables (45): demographic and encounter-level measures (e.g. age_at_ed, ed_los_hours), vital signs collected during the first six hours, initial laboratory measurements, derived variability measures and binary missingness indicators (e.g. missing_lactate_first_6h, missing_troponin_first_6h).

The numerical features can be grouped into clinically meaningful categories:

1. **Demographic and encounter-level measures:** age, ED length of stay, number of vital sign measurements within the first six hours.
2. **Physiological measurements:** systolic and diastolic blood pressure, heart rate, respiratory rate, oxygen saturation, and body temperature.
3. **Laboratory variables:** lactate, troponin, creatinine, potassium, sodium, bicarbonate, white blood cell count, hemoglobin, and platelet count.
4. **Missingness indicators:** preserve information about whether measurements were unavailable, which may convey clinical relevance.

Categorical variables show meaningful distributions. Gender is relatively balanced, with **229,838 female** and **195,114 male** encounters. Arrival mode is dominated by WALK IN (251,799) and AMBULANCE (155,719). Race is represented across multiple categories, with the largest being WHITE (228,077) and BLACK/AFRICAN AMERICAN (76,777). Chief complaints are high-cardinality text fields reflecting a variety of ED presentations.

Feature Type	Count	Examples
Administrative identifiers	5	stay_id, subject_id, hadm_id, ed_intime, ed_outtime
Categorical variables	4	gender, arrival_transport, race, chiefcomplaint
Numerical variables	45	Age, vital signs, labs, variability measures
Missingness indicators	Multiple	missing_lactate_first_6h, missing_troponin_first_6h

Table 4.1: Summary of Dataset Structure

The dataset contains **three mortality-related outcome variables**:

Outcome	Positive Cases	Negative Cases	Prevalence
y_death_hosp	4,587	420,365	1.08%
y_death_24h	595	424,357	0.14%
y_death_48h	1,119	423,833	0.26%

Table 4.2: Outcome Variable Distribution

Mortality is a rare event, particularly for very early death, creating **severe class imbalance**. This characteristic necessitates careful consideration during model training and evaluation.

The dataset also exhibits **substantial missingness**, especially in laboratory measurements: troponin_first_6h is missing in 98.89% of encounters, lactate_first_6h in 94.74%, and several other lab variables have roughly 90% missing values. Some vital sign variability measures, such as `sbp_std_6h` and `hr_std_6h`, are missing in ~32% of cases, while temperature variables are missing in ~10% of encounters. This pattern indicates selective testing practices in the ED, and the inclusion of missingness indicators allows models to leverage this information.

Clinical distributions are skewed and heterogeneous. For example, age_at_ed ranges from 18 to 103 years (mean 52.87, median 53), and ED length of stay ranges from 0 to 493 hours (mean 7.16, median 5.47). These extreme values highlight the presence of outliers and support the use of robust preprocessing strategies, such as imputation and scaling.

Variable	Mean	Median	Range	Missing %
age_at_ed	52.87	53	18–103	0
ed_los_hours	7.16	5.47	0–493	0
n_vitalsign_measurements_6h	—	2	1–105	0
sbp_std_6h	—	—	—	32.55
hr_std_6h	—	—	—	32.20
troponin_first_6h	—	—	—	98.89
lactate_first_6h	—	—	—	94.74

Table 4.3: Example Summary Statistics for Key Numerical Variables

4.3 Data Preprocessing

The preprocessing stage was designed to transform the raw ED dataset into a format suitable for machine learning, while preserving methodological rigor and reducing the risk of bias or information leakage. Given the characteristics of w6_mortality_multi.csv, including mixed data types, substantial missingness, severe class imbalance and skewed clinical measurements, preprocessing was treated as a critical component of the modelling pipeline rather than a purely technical preliminary step.

4.3.1 Feature Exclusion and Selection

Feature preparation began with the exclusion of variables that could either identify patients directly, encode temporal metadata unsuitable for generalizable prediction, or leak outcome information into the predictors. The variables stay_id, subject_id, hadm_id, ed_intime, and ed_outtime were removed from the feature space before modeling. These variables served administrative or timestamp functions and were not appropriate predictive inputs for the intended clinical risk modelling task. Their inclusion would have risked memorization, reduced transportability, or indirect leakage of encounter chronology.

All target variables were also excluded from the predictors. The global target set consisted of y_death_hosp, y_death_24h, and y_death_48h. When one of these served as the target for a given model, the others were explicitly removed from the feature matrix as well. This was especially important because the mortality labels are logically related to one another. For example, allowing y_death_hosp to remain in the feature set while predicting y_death_24h would trivialize the task, and allowing y_death_24h to remain while predicting the multiclass mortality timing outcome would directly contaminate the learning process.

Beyond exclusion, the preprocessing framework also formalized feature construction through clinically informed engineering. Rather than relying only on the raw recorded variables, the pipeline created several derived variables intended to represent physiological patterns more directly related to mortality risk. The first engineered variable, shock_index, was defined as heart rate divided by systolic blood pressure and was included because it is a clinically recognized marker of hemodynamic instability. Patients with tachycardia relative to blood pressure may exhibit compensated or overt shock even when individual raw measurements appear only moderately abnormal.

The second engineered feature, map_6h, represented mean arterial pressure derived from systolic and diastolic pressure. Mean arterial pressure is more physiologically informative than isolated blood pressure components in the context of perfusion, especially when mortality is related to circulatory failure. The third feature, n_abnormal_vitals, counted how many key vital signs crossed clinically meaningful abnormality thresholds. This aggregation was intended to summarize overall physiological derangement in a compact form. Rather than forcing the model to discover these threshold combinations independently, the feature provided an explicit signal of multi-system instability.

The fourth feature, vitals_per_hour, quantified the number of recorded vital sign measurements normalized by emergency department length of stay. This variable was motivated by the idea that monitoring intensity itself may carry prognostic information. Patients who are clinically unstable are often reassessed more frequently, and such operational intensity may reflect concern that is not fully captured by the absolute physiological values. The fifth feature, pulse_pressure, was defined as the difference between mean systolic pressure and minimum diastolic pressure. Pulse pressure can reflect vascular tone, cardiac output dynamics, or hemodynamic stress, all of which may be relevant to severe deterioration.

The sixth engineered feature, age_over_75, represented age as a binary risk indicator in addition to the continuous age variable. The rationale was that mortality risk may increase non-linearly beyond an elderly threshold, and a dedicated indicator can help the model capture a clinically meaningful transition in vulnerability. The seventh feature, any_labs_ordered, summarized whether at least one major laboratory test had been obtained. This variable was designed to encode the fact that laboratory ordering is itself a marker of clinical concern and care intensity. In emergency department practice, the decision to order blood tests often reflects triage judgment, perceived instability, or suspicion of serious illness.

Engineered Feature	**Description**	**Clinical Motivation**
shock_index	HR / SBP	Hemodynamic instability
map_6h	Mean arterial pressure	Perfusion
n_abnormal_vitals	Count of abnormal vitals	Physiological derangement
vitals_per_hour	Monitoring frequency	Care intensity
pulse_pressure	SBP − DBP	Hemodynamic stress
age_over_75	Binary age flag	Elderly vulnerability
any_labs_ordered	Lab ordering indicator	Clinical concern

Table 4.4: Engineered Features

These engineered features were not arbitrary additions. They were constructed to embed clinically meaningful structure into the predictor space, thereby complementing the expressive power of LightGBM with variables that may improve separability of severe outcomes. In a dataset characterized by rare events, heterogeneous measurement patterns, and complex interactions between physiology and care processes, such feature engineering was an important part of the final methodology.

4.3.2 Data Type Separation

After feature exclusion and engineering, variables were separated according to data type and preprocessing requirements. This separation was necessary because the dataset contained heterogeneous information sources, and different categories of variables required different transformations. The preprocessing framework distinguished among numerical variables, missingness indicators, low-cardinality categorical variables, and higher-cardinality categorical variables.

Numerical variables included demographic quantities such as age, process measures such as emergency department length of stay, vital sign summaries over the first six hours, laboratory values, variability-related summaries, and all engineered continuous or count-based features. These variables were handled through imputation and scaling. The numerical group included both original dataset columns and derived features such as shock_index, map_6h, n_abnormal_vitals, vitals_per_hour, pulse_pressure, age_over_75, and any_labs_ordered, provided that they were numeric in representation.

Missingness indicators formed a separate group. These were variables whose names began with the missing_ prefix and encoded whether a corresponding measurement was absent. Since these variables already represented binary informational signals, they were not treated as ordinary continuous variables and were not standardized. Preserving them in an interpretable 0/1 form was important both for model behavior and for later explanation.

Categorical variables were divided by cardinality. gender and arrival_transport were treated as low-cardinality categories suitable for direct one-hot encoding. In contrast, chiefcomplaint and race required more careful handling because they contained substantially larger sets of distinct values. Direct one-hot encoding of all unique values would have produced an unnecessarily sparse and unstable representation, particularly for categories occurring only a small number of times.

The final separation strategy therefore reflected both statistical and practical considerations. It allowed each variable type to be processed in a way consistent with its semantic role in the data while also controlling the dimensionality and stability of the final feature matrix. This structured separation formed the basis of the unified preprocessing pipeline applied consistently across all models.

4.3.3 Handling Missing Values

Missing data were handled explicitly and conservatively. For numerical variables, the preprocessing pipeline applied median imputation. Median imputation was selected because it is robust to skewed clinical distributions and less sensitive to extreme values than mean imputation. In emergency department data, laboratory and physiological variables can exhibit strong asymmetry and outliers, particularly among critically ill patients, making the median a more stable central tendency estimate.

This imputation strategy was fitted using the training data only. The median value for each numerical variable was learned from the training partition and then applied unchanged to the validation and test partitions. This design was essential to prevent information leakage, as even apparently harmless summary statistics can subtly transmit distributional information from evaluation data into the training process if computed globally.

Categorical variables were imputed using a constant placeholder category, MISSING. This approach ensured that missing categorical values were retained as a valid and observable category rather than silently discarded or filled with a potentially misleading existing label. In a clinical setting, missing documentation may itself be associated with operational circumstances or patient acuity, and retaining such cases in a distinct category is therefore preferable to implicit removal.

Missingness indicators were also preserved as separate predictors. This design choice acknowledged that absence of measurement can carry information. For example, the fact that lactate or troponin was not measured may signal lower clinical suspicion or lower observed severity, while the presence of those tests may reflect concern about shock, sepsis, or myocardial injury. By combining median imputation with explicit missingness flags, the methodology ensured that the model could make use of both the observed values and the informational content of their absence.

In this way, missing data handling was not treated merely as a preprocessing necessity. It was incorporated as a substantive modelling consideration, aligned with the realities of emergency care and implemented in a leakage-aware manner.

4.3.4 Scaling and Encoding

Following imputation, numerical variables were standardized using StandardScaler. Standardization was applied to the numerical feature block after median imputation and was

again fitted exclusively on the training data. Although tree-based models such as LightGBM are generally less sensitive to scaling than linear models, the use of a shared preprocessing framework across multiple analyses and explainability procedures made standardized numerical inputs beneficial. Standardization also supported later rule extraction and made transformed thresholds interpretable within a consistent scaled space, even though those thresholds required careful explanation in TE2Rules outputs.

Categorical encoding was performed using one-hot encoding with the handle_unknown="ignore" setting. This ensured that categories appearing in validation or test data but not seen during training would not cause transformation failures. Such robustness was especially important in emergency department data, where textual chief complaints and recorded demographic categories can vary across encounters.

Before one-hot encoding, two higher-cardinality variables were reduced using top-N category retention. For chiefcomplaint, the fifty most frequent values in the training set were retained, while all other complaints were mapped to OTHER. For race, the ten most frequent values in the training set were retained, and all remaining values were mapped to OTHER. This strategy achieved two aims simultaneously. First, it reduced the instability associated with extremely rare categories. Second, it preserved common and potentially informative categories without allowing feature dimensionality to expand uncontrollably.

The use of top-N grouping was particularly appropriate for chiefcomplaint, where the raw data contained highly heterogeneous and inconsistently formatted text values. In such settings, preserving a limited number of common categories while collapsing the long tail offers a pragmatic balance between information retention and regularization. A similar rationale applied to race, where a limited number of common categories captured most observations while rare categories were grouped to avoid fragmented encoding.

Taken together, scaling and encoding transformed the heterogeneous raw data into a dense and consistent model-ready matrix. The transformation preserved clinically meaningful information, limited unnecessary sparsity, and supported both model training and subsequent interpretability analyses.

4.3.5 Preprocessing Pipeline Implementation

All preprocessing operations were implemented through a unified, reusable pipeline. The study did not rely on ad hoc manual transformations for each separate model. Instead, a shared preprocessing class encapsulated data loading, feature engineering, categorical mapping, column grouping, and transformation using a scikit-learn ColumnTransformer wrapped within a pipeline-style interface.

This implementation was important for methodological consistency. The same conceptual preprocessing logic was applied across Model 1, Model 2, and the directly trained multiclass variant of Model 3. The preprocessing object was fitted on the training partition for each task, after any cohort filtering and after target-specific exclusions had been defined. Once fitted, the same object transformed validation and test data without any refitting. This prevented accidental divergence between training and evaluation preprocessing and ensured that the saved preprocessor could later be used together with the serialized model.

The pipeline also supported model persistence. For each trained LightGBM model, both the fitted estimator and the fitted preprocessor were saved as serialized artifacts. This made the workflow reproducible and allowed downstream analyses, including SHAP and TE2Rules, to reconstruct the exact transformed feature space used during training. It also ensured that the cascade formulation for Model 3 could apply the saved Model 1 and Model 2 preprocessors to the shared test set in a deterministic manner.

An additional methodological strength of the implementation was that feature engineering occurred inside the preprocessing flow rather than as a detached manual step. This meant that derived variables were created systematically for every relevant split and every model, reducing the risk of inconsistency. The same applied to category reduction and one-hot encoding, both of which were training-set dependent and therefore appropriately encapsulated within the fitted preprocessing object.

Overall, the preprocessing pipeline served not only as a convenience mechanism but as an essential methodological safeguard. It enforced consistency, supported reproducibility, and operationalized the requirement that all learned transformations be estimated only from training data.

4.3.6 Key Considerations

The preprocessing framework was specifically designed to handle the following datasets characteristics:

- Substantial missingness in laboratory variables
- Skewed numerical distributions and outliers
- Mixed categorical structure requiring encoding
- High-cardinality variables (chiefcomplaint)
- Repeated encounters and multiple mortality outcomes requiring careful exclusion of identifiers and labels

The framework consisted of six main steps:

1. Exclusion of identifiers and outcome labels
2. Removal of the high-cardinality chiefcomplaint variable
3. Separation of variables by data type
4. Imputation of missing numerical and categorical values
5. Standardization of numerical features
6. One-hot encoding of categorical features

This systematic approach ensured the data were prepared in a reproducible, statistically appropriate and clinically defensible manner for subsequent machine learning analysis.

4.4 Developed Models

This study implemented three complementary predictive models to examine mortality rosk in emergency departments (ED) patients using w6_mortalty_multi.csv. While all models share common preprocessing framework and use the same ML algorithms, each addresses a distinct predictive objective.

4.4.1 Model 1: Binary Mortality Timing Within the Death Cohort

Model 1 addressed the question of mortality timing among patients who ultimately died in hospital. The cohort for this model was restricted to encounters with y_death_hosp = 1, yielding 4,587 encounters in total. Within this deaths-only cohort, the target variable was y_death_24h, where the positive class indicated death within 24 hours of emergency department presentation and the negative class indicated death occurring later than 24 hours.

This task was clinically meaningful because it focused specifically on identifying the most rapidly deteriorating patients within the mortality cohort. Distinguishing early from later

death is more refined than overall mortality prediction and may reflect differences in acute instability, physiological reserve, and severity at presentation. Because the cohort was much smaller than the full dataset and the positive class represented roughly 13% of cases, the model required careful handling of both sample size and moderate class imbalance.
The model was implemented using LightGBM in binary classification mode with class_weight="balanced". Hyperparameter tuning was performed using RandomizedSearchCV over the training partition only. The search space included the number of estimators, maximum tree depth, number of leaves, minimum child samples, learning rate, L1 regularization, L2 regularization, column subsampling, and row subsampling. Five-fold stratified cross-validation was used within the training set, and the optimization objective was ROC-AUC.

After training, the model produced class probabilities on the validation set, and the operating threshold was selected using Youden's J statistic. That threshold was then fixed and applied to the locked test set. The final saved model therefore reflected a sequence of cohort restriction, leakage-aware preprocessing, stratified hyperparameter tuning, validation-based threshold optimization, and test-set evaluation.

From a methodological standpoint, Model 1 served as a specialized early-mortality classifier. It did not attempt to identify who would die overall; rather, it focused on when death would occur among those already in the mortality cohort. This role became especially important later in the cascade formulation of Model 3, where the Model 1 probability estimate was used as the conditional early-death component..

4.4.2 Model 2: Binary Full-Cohort In-Hospital Mortality
Model 2 addressed overall in-hospital mortality in the full emergency department cohort. This was the broadest and most operationally relevant task in the study, since it estimated whether a patient would die during the hospital admission regardless of the exact timing. The target variable was y_death_hosp, and the model was trained on all 424,952 encounters.

The principal methodological challenge in Model 2 was extreme class imbalance. Only 4,587 encounters were positive for in-hospital death, corresponding to a prevalence of approximately 1.08%. In such settings, a model can achieve high apparent accuracy by overwhelmingly predicting the majority class, while still failing to identify clinically important deaths. Consequently, the training and evaluation strategy was designed to emphasize discrimination and threshold-aware performance rather than raw accuracy.

Model 2 was implemented as a binary LightGBM classifier. Unlike Model 1, imbalance handling was addressed through tuning of scale_pos_weight rather than a fixed balanced-class setting. The candidate values included moderate and more aggressive positive-class weighting, including the empirical class ratio. This allowed the search process to determine an appropriate trade-off between sensitivity and specificity rather than imposing full inverse-frequency weighting by default.

Hyperparameter tuning was conducted through RandomizedSearchCV using three-fold stratified cross-validation on the training partition. As in Model 1, the search space included the number of estimators, tree depth, number of leaves, minimum child samples, learning rate, regularization parameters, and subsampling parameters. ROC-AUC was used as the tuning objective. Once the best model had been selected, predicted probabilities on the validation set were used to determine the decision threshold via Youden's J statistic. That threshold was then transferred unchanged to the test set.

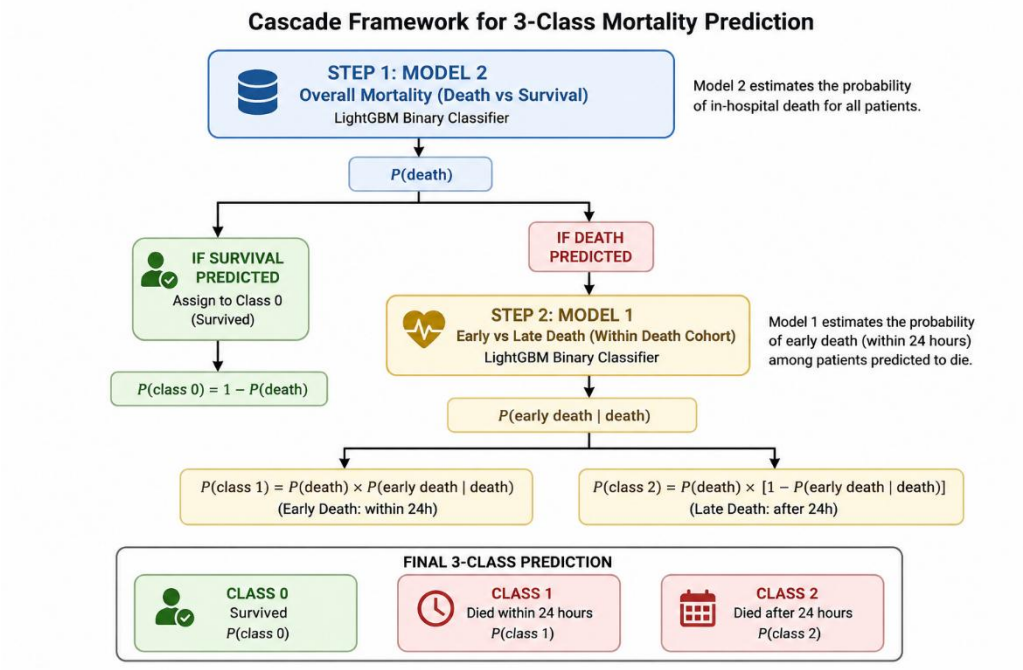
Model 2 functioned as the primary mortality screener in the final framework. Because it was trained on the full cohort and optimized to distinguish death from survival, it provided the foundational mortality probability estimate for the broader dissertation objectives. It was also the first stage of the final multiclass cascade, where its estimated probability of death acted as the gate separating survivors from patients entering the mortality-timing branch.

4.4.3 Model 3: Multiclass Mortality Timing

Model 3 addressed multiclass mortality timing over the full cohort. The underlying clinical objective was to distinguish among three outcome states: survival, death within 24 hours, and death after 24 hours. This task was formulated through a derived target variable, timing_class, where class 0 represented survivors, class 1 represented deaths within 24 hours, and class 2 represented deaths after 24 hours.

The final methodology adopted a two-stage cascade architecture for this task. This choice reflected the structure of the outcome itself. Mortality timing can naturally be decomposed into two sequential questions: first, whether the patient will die at all; and second, conditional on death, whether that death will occur early or later. The cascade therefore aligned the model design with the clinical logic of the task.

In the first stage of the cascade, Model 2 produced the probability of in-hospital death for each encounter in the test cohort. In the second stage, Model 1 produced the probability of early death conditional on belonging to the mortality cohort. These probabilities were then combined algebraically to obtain a three-class probability distribution. Specifically, the probability of survival was defined as one minus the death probability, the probability of early death was defined as the product of the death probability and the early-death-given-death probability, and the probability of later death was defined as the product of the death probability and the complement of the early-death-given-death probability. The resulting three probabilities were normalized so that each row summed to one.



This formulation offered several advantages. First, it decomposed a highly imbalanced multiclass problem into two more tractable binary problems. Second, it exploited the fact that overall mortality and timing within the death cohort are related but not identical tasks. Third, it allowed the study to benefit from a mortality model trained on the full cohort and a timing model trained specifically within the death cohort, where the distinction between early and later death is clinically most relevant.

The final multiclass methodology also included direct multiclass and SMOTE-based analyses as robustness-oriented model development experiments, but the dissertation's operational multiclass framework is the cascade architecture. The direct multiclass LightGBM treated the problem as a single three-class task with balanced class weights, while the SMOTE variant oversampled the rare early-death class in the transformed training space. These analyses were

informative for understanding the difficulty of the rare class, but the final multiclass prediction framework was defined by the cascade because it reflected the most coherent decomposition of the prediction task and yielded the strongest end-to-end performance on the locked test se**t.**

4.4.4 Common Algorithm Considerations

Across the main models, LightGBM was chosen as the primary learning framework because it was particularly well suited to the characteristics of the dataset. Emergency department mortality prediction involves non-linear interactions between vital signs, laboratory values, demographics, and care process variables. It also involves mixed scales, skewed distributions, structured missingness, and severe outcome imbalance. Gradient-boosted decision trees are well adapted to such settings because they can capture non-linear boundaries and higher-order interactions without requiring extensive manual specification of interaction terms.

LightGBM was especially appropriate given the feature engineering strategy. Engineered variables such as shock_index, map_6h, and n_abnormal_vitals were clinically meaningful on their own, but their predictive value also depend on interactions with age, laboratory values, and monitoring intensity. Tree-based boosting can model such relationships more flexibly than a strictly linear method. In addition, LightGBM provides feature importance measures and is compatible with SHAP, which supported the dissertation's emphasis on interpretability.

At the same time, the study retained logistic regression as a baseline comparison for the two binary tasks. This was important because logistic regression remains a familiar and historically influential modeling family in clinical prediction research. By evaluating a regularized linear baseline on the same patient-level splits and the same shared preprocessing pipeline, the study was able to assess whether the more expressive gradient-boosted trees captured useful structure beyond what could be achieved with a simpler linear model.

The models also shared a consistent philosophy regarding leakage control and operating-point selection. Hyperparameter tuning was always confined to the training partition. Validation data were used for threshold selection in binary tasks but not for retraining. Test data were held back for final evaluation only. This consistency in experimental logic was as important as the specific algorithmic choices themselves, because it ensured that performance estimates reflected generalization rather than inadvertent reuse of evaluation information.

4.5 Explainability Framework

Interpretability was treated as a substantive requirement of the modeling framework rather than an optional add-on. Because mortality prediction in emergency care can influence triage, escalation, and clinical prioritization, it is not sufficient for a model to achieve strong predictive performance alone. The study therefore incorporated two complementary post-hoc explanation methods: SHAP and TE2Rules.

SHAP was used to quantify the contribution of individual features to model predictions. For the LightGBM models, SHAP values were computed using TreeExplainer, which is well suited to tree-based ensembles and provides exact or highly efficient explanations in that context. SHAP supported global interpretation through summary plots that ranked variables by mean absolute contribution and visualized the directionality of feature effects. This was methodologically important because it allowed the study not only to identify influential predictors but also to observe how low or high feature values shifted predicted mortality risk.

For Model 1, SHAP was computed over the full test set because the deaths-only cohort was comparatively small. For Models 2 and the direct multiclass Model 3, SHAP was computed on stratified subsamples of the test set to control runtime while preserving class structure. In the multiclass setting, SHAP values were generated separately for the relevant mortality classes, allowing interpretation of the early-death and later-death decision tendencies.

SHAP was particularly valuable in the context of the engineered features. Derived variables such as vitals_per_hour, shock_index, map_6h, and pulse_pressure were not introduced merely as heuristic additions; they were intended to encode clinically relevant signals that could be examined after training. SHAP made it possible to verify whether these variables contributed meaningfully to model behavior and whether their contributions were directionally consistent with clinical expectations.

TE2Rules provided a second layer of interpretability by extracting human-readable IF-THEN rules from the trained LightGBM models. Because TE2Rules does not natively consume LightGBM models, a custom adapter was used to convert LightGBM tree structures into the internal tree representation required by the rule-mining framework. For the multiclass problem, TE2Rules was applied in a one-vs-rest manner for the direct multiclass LightGBM so that separate rule sets could be generated for the early-death and later-death classes.

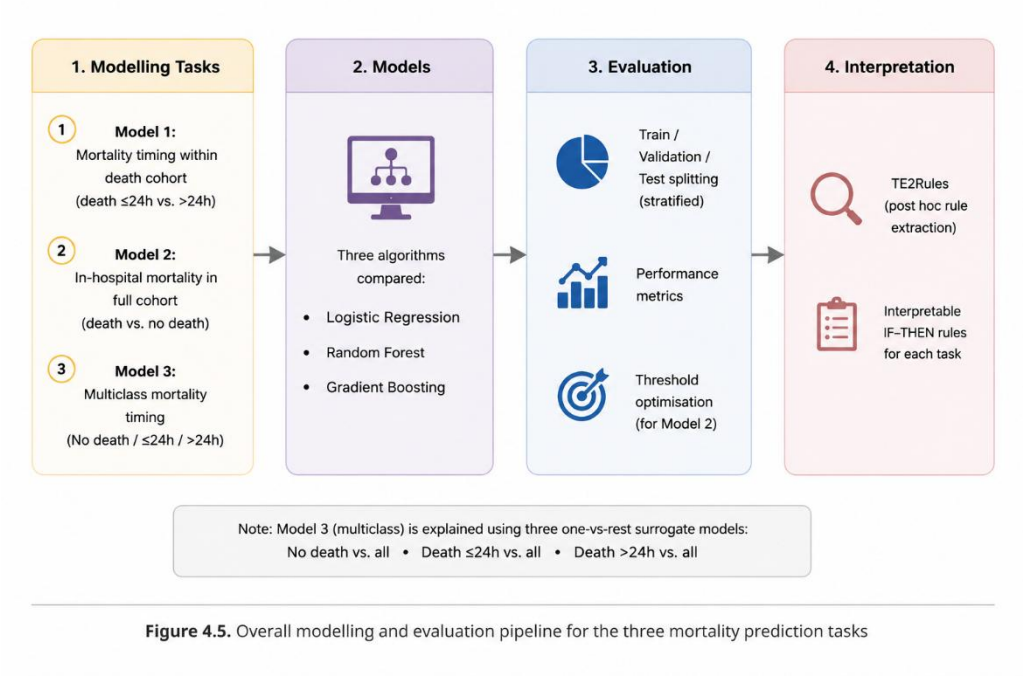


Figure 4.5. Overall modelling and evaluation pipeline for the three mortality prediction tasks

The purpose of TE2Rules was different from that of SHAP. Whereas SHAP provides additive contribution-based explanations at the feature level, TE2Rules generates compact decision rules that summarize patterns associated with positive predictions. Such rules are useful for communicating model behavior to readers who may prefer threshold-based reasoning or who wish to understand which combinations of conditions tend to define high-risk subgroups.

Together, SHAP and TE2Rules formed a layered explainability framework. SHAP offered quantitative and directional interpretation of feature contributions, while TE2Rules offered symbolic and human-readable approximations of ensemble decision logic. This combination strengthened the transparency of the dissertation's predictive framework and supported more clinically interpretable conclusions.

4.6 Experimental Setup

4.6.1 Data Partitioning

A central design principle of the experimental setup was that partitioning had to occur at the patient level rather than the encounter level. The dataset included repeated encounters for some individuals, and therefore a random row-wise split would have allowed the same patient to appear in both training and evaluation partitions. Such overlap would create information leakage because patient-specific characteristics and recurring clinical patterns could be indirectly learned during training and recognized during testing.

To prevent this, all partitioning was based on subject_id. Each patient was assigned entirely to one split only and all encounters associated with that patient followed the same assignment. Stratification was performed at the patient level using the maximum value of the relevant target within each subject, ensuring that the prevalence of the target outcome remained approximately stable across splits.

The final partitioning strategy was a 60/20/20 split into training, validation, and test sets. This split was applied separately within each modeling task after any necessary cohort restriction or target construction. The training set was used for preprocessing fitting and hyperparameter tuning. The validation set was reserved for model selection support, especially threshold optimization in the binary tasks. The test set was kept locked for final performance estimation.

This design balanced several needs simultaneously. It provided a sufficiently large training set for model fitting, maintained an independent validation subset for operational threshold selection, and preserved a sizable and untouched test subset for unbiased evaluation. For a study involving rare outcomes, explicit separation of validation and test roles was especially important because threshold choices can materially alter reported performance under severe imbalance.

4.6.2 Cross Validation

Within the training partition, cross-validation was used for hyperparameter tuning. The binary early-versus-late mortality model used five-fold stratified cross-validation because the deaths-only cohort was comparatively small and benefited from more efficient reuse of available data during model selection. The full-cohort mortality model and the direct multiclass timing model used three-fold stratified cross-validation. This choice reflected the much larger sample sizes in those tasks and reduced computational burden without compromising the stability of model selection.

Stratification was employed so that class proportions remained approximately consistent across folds. This was important for all tasks, but especially for the full-cohort mortality and multiclass timing problems, where the positive or rare classes represented only a small fraction of the data. Without stratification, some folds could become unrepresentative or too sparse to support stable evaluation during hyperparameter search.

The optimization objective differed by task. For the two binary LightGBM models, ROC-AUC was used as the scoring criterion during cross-validation. This was appropriate because

ROC-AUC measures ranking performance independent of any single threshold and remains informative under imbalance when the goal is to assess how well the model separates positives from negatives. For the direct multiclass timing model and the SMOTE-based multiclass variant, macro F1 was used as the cross-validation objective. Macro F1 was chosen because it assigns equal importance to all classes, including the rare early-death class.

Cross-validation was therefore used not as a final evaluation method but as an internal model-selection mechanism within the training data. This distinction is important. The reported dissertation results are based on the locked test set, whereas cross-validation served to identify suitable hyperparameter configurations without exposing the model to evaluation data.

4.6.3 Class Imbalance Handling

Class imbalance was a defining characteristic of the study and influenced both modelling and evaluation choices. The full-cohort mortality task involved a positive class prevalence of approximately 1.08%, while the early-death subgroup was much rarer still. The multiclass timing problem was particularly challenging because the early-death class constituted only a very small proportion of all encounters.

The methodology addressed imbalance differently depending on the task structure. In Model 1, the imbalance was moderate within the deaths-only cohort, and class_weight="balanced" was used within the LightGBM classifier. This allowed the learning algorithm to increase the influence of the minority early-death cases without introducing synthetic data.

In Model 2, imbalance was more severe and was addressed through a tuned scale_pos_weight hyperparameter. Rather than forcing the model to adopt the full empirical inverse-frequency ratio, several candidate values were evaluated during hyperparameter search. This allowed the selected model to balance sensitivity and specificity in a data-driven way rather than if maximal reweighting would necessarily produce the best discrimination.

In the direct multiclass version of Model 3, class imbalance was handled through class_weight="balanced", which increased the loss contribution of the rare mortality classes relative to the dominant survival class. In addition, a dedicated SMOTE-based multiclass experiment was implemented to test whether synthetic oversampling of the early-death class could improve representation of the rarest class in the transformed training space. In that experiment, only the training set was oversampled, and the validation and test sets remained

untouched. This respected the general principle that synthetic data generation must not contaminate evaluation partitions.

The final multiclass methodology, however, relied on cascade decomposition rather than synthetic resampling. This decision was justified not only by empirical performance but by the structure of the prediction task. Decomposing mortality timing into mortality first and timing second offered a principled way to reduce the impact of extreme rarity by solving two more coherent subproblems.

4.6.4 Threshold Optimization

For the threshold selection was treated as an explicit methodological step. After each binary model had been tuned and fitted on the training data, predicted probabilities were generated for the validation set. The optimal operating threshold was then selected using Youden's J statistic, defined as sensitivity plus specificity minus one.

This choice was motivated by the need to derive a clinically meaningful operating point from the validation data rather than from the test data or an arbitrary default. In a heavily imbalanced clinical classification problem, a default threshold can produce poor sensitivity or poor specificity depending on the calibration and score distribution of the model. Youden's J provides a principled way to identify the threshold that maximizes the trade-off between true-positive and true-negative rates.

Once selected, the validation-derived threshold was fixed and applied unchanged to the test set. This separation was essential. If thresholds were tuned directly on the test set, the resulting performance estimates would be optimistic and would no longer represent true out-of-sample generalization. By isolating threshold optimization to the validation set, the methodology preserved the integrity of the final evaluation.

The multiclass models did not use threshold tuning in the same way. The direct multiclass LightGBM and the cascade model both produced class probability distributions, and the predicted class was taken as the one with maximum probability. Thus, threshold optimization was a task-specific component of the binary models and not a universal step applied indiscriminately across all tasks.

4.6.5 Software and Hardware

The workflow was implemented in Python with the following libraries:

- The software environment recorded for the project included NumPy 2.2.2, pandas 2.2.2, scikit-learn 1.5.2, imbalanced-learn 0.13.0, Matplotlib 3.10.0, TE2Rules 1.0.1, joblib 1.4.2, XGBoost 2.1.4, and LightGBM 4.6.0. The final predictive pipeline described in this chapter used LightGBM as the primary modeling framework

The experiments were executed on a CPU based environment (AMD64 architecture) with parallel execution enabled for Random Forest (n_jobs = -1). No GPU frameworks were used, and all preprocessing and model fitting were embedded within scikit-learn pipelines to ensure reproducibility.

4.6.6 Reproducibility Measures

- Fixed random seed (42) for all splits, sampling, and model initialization
- Locked test set remained untouched for evaluation
- All preprocessing, model outputs, and TE2Rules explanations were archived

4.7 Evaluation Methodology

The predictive performance of the three mortality models was assessed using a comprehensive set of evaluation metrics, chosen to capture both overall discrimination and clinically relevant classification performance. Given the severe class imbalance in the dataset, evaluation relied not only on accuracy but also on threshold independent, threshold dependent and class sensitive metrics.

4.7.1 Binary Models (Model 1 & Model 2)

For the two binary tasks, evaluation was performed using:

Threshold-independent metrics:

- ROC-AUC: Measures the model's ability to rank positive cases above negative cases across all thresholds.
- PR-AUC: Focuses on precision-recall trade-off, particularly informative for rare positive classes.

Threshold-dependent metrics calculated from the confusion matrix:

- Accuracy: (TP + TN) / (TP + TN + FP + FN)
- Precision: TP / (TP + FP)

- Recall (Sensitivity): TP / (TP + FN)
- Specificity: TN / (TN + FP)
- F1-score: 2 × (Precision × Recall) / (Precision + Recall)
- Negative Predictive Value (NPV): TN / (TN + FN)

For Model 2, which represents the full ED cohort, additional threshold-dependent measures were considered: balanced accuracy, alert burden, alerts per true case, and PPV lift relative to prevalence. A threshold sweep (0.01–0.99) was performed to select the optimal classification threshold, maximizing F1-score while maintaining PR-AUC as a tiebreaker.

Confusion matrices were used to interpret the classification outcomes, highlighting differences in false positive and false negative rates at the selected operating thresholds.

4.7.2 Multiclass Model (Model 3)

For the multiclass mortality timing task, metrics were chosen to account for dominance of the no-death class:

- Accuracy
- Macro F1-score: Equal weight to each class, preventing majority-class dominance
- Weighted F1-score: Weighted by class prevalence
- One-vs-rest macro-ROC-AUC: Each class treated as positive against the rest

Confusion matrices and classification reports were used to identify misclassification patterns between early death, late death, and no-death classes. Metrics were macro-averaged to ensure fair assessment of rare classes.

4.7.3 Class Imbalance Considerations

- PR-AUC, precision, recall, and F1-score were emphasized to directly assess detection of rare positive events.
- Macro-averaged metrics for Model 3 accounted for class imbalance in multiclass evaluation.
- Threshold selection for Model 2 focused on F1-score rather than accuracy to balance sensitivity and precision under severe imbalance.

4.7.4 Rationale for Metric Selection

- ROC-AUC: Broad ranking ability
- PR-AUC: Informative for rare positive classes
- Precision: Fraction of flagged cases that are truly positive
- Recall (Sensitivity): Fraction of actual positive cases detected
- Specificity: Correct identification of negative cases
- F1-score: Balanced summary of precision and recall
- Macro & Weighted F1: Ensure fair multiclass evaluation

This multi-metric evaluation framework ensured that the models were assessed in a clinically meaningful and methodologically robust manner, reflecting both predictive accuracy and practical utility in rare-event emergency department mortality prediction.

Model	Primary Metrics	Additional Metrics	Notes
Model 1 – Binary mortality timing	ROC-AUC, PR-AUC	Accuracy, Precision, Recall, F1-score, Specificity, NPV	Evaluates early vs. later death within mortality cohort; PR-AUC emphasized due to rare positive class
Model 2 – Binary full-cohort mortality	ROC-AUC, PR-AUC	Accuracy, Precision, Recall, F1-score, Specificity, NPV, Balanced Accuracy, Alert Burden, PPV Lift	Full cohort, extreme class imbalance; threshold tuning applied (final threshold = 0.12)
Model 3 – Multiclass mortality timing	Macro F1-score, Weighted F1-score, One-vs-Rest ROC-AUC	Accuracy, Confusion Matrix	No-death class dominates; macro-averaged metrics ensure fair assessment of rare classes

Table 4.5: Evaluation metrics applied to each model, highlighting the use of PR-AUC and macro/weighted metrics to handle class imbalance and rare-event outcomes.

Chapter 5

Results

This chapter presents the outcomes of the predictive models developed for mortality assessment in emergency department patients. It provides a comprehensive evaluation of model performance across three tasks: binary mortality timing within the death cohort, binary full-cohort in-hospital mortality, and multiclass mortality timing. The results include threshold-independent metrics, threshold-dependent metrics, and multiclass evaluation metrics, highlighting both the ranking ability and clinical utility of the models. The chapter also compares model performance, analyses ROC and PR curves to assess discrimination and precision-recall balance and presents the outputs of the TE2Rules explainability framework. Finally, a summary synthesizes the findings, emphasizing the models' predictive strengths, limitations, and practical implications for clinical decision-making.

5.1 Model Performance Results

Model performance was assessed using the locked test sets and realistic validation outputs. Threshold-independent metrics (ROC-AUC, PR-AUC) and threshold-dependent metrics (accuracy, precision, recall, F1-score, specificity) were reported for the binary models. For the multiclass model, macro F1, weighted F1, and one-vs-rest ROC-AUC were the primary measures.

Model 1 – Binary Mortality Timing

Model 1 aimed to distinguish early death within 24 hours from later in-hospital death within the deaths-only cohort. On the test set, Gradient Boosting achieved ROC-AUC = 0.7822 and PR-AUC = 0.3778. Using the validation-selected threshold of 0.4613, the model obtained:

- Accuracy = 0.77
- Precision = 0.31
- Recall = 0.60
- F1-score = 0.41
- Specificity = 0.79

Interpretation**:** The positive class is small (early deaths), which naturally limits precision and F1. The model is moderately good at ranking high-risk patients (ROC-AUC) and identifying most early deaths (recall), but many false positives occur because the class is underrepresented. Logistic Regression offers higher recall at the cost of lower ROC-AUC, reflecting a trade-off between sensitivity and specificity.

Model 2 – Binary Full-Cohort Mortality

Model 2 predicted in-hospital mortality across all ED patients.

Gradient Boosting trained on the natural cohort achieved:

- ROC-AUC = 0.92
- PR-AUC = 0.21
- Accuracy = 0.83
- Precision = 0.05
- Recall = 0.85
- F1-score = 0.10
- Specificity = 0.83
- Threshold = 0.07

Interpretation: Precision appears low because in-hospital mortality is extremely rare (~1 death per 92 survivors). Even a strong model will produce many false positives under these conditions. Recall is high, meaning the model successfully identifies most deaths, which is crucial for clinical use. F1-score is naturally limited by low precision but is expected in rare-event prediction. Model 2 is considered the most reliable because it uses the full patient cohort and a locked test set, giving realistic generalization.

Model 3 – Multiclass Mortality Timing

Model 3 was designed to predict three possible outcomes for emergency department patients: no death, death within 24 hours, and in-hospital death after 24 hours. Several versions of Model 3 were developed and evaluated to identify the approach that balances predictive performance with generalizability and explainability.

1. Baseline LightGBM Multiclass Model

- This direct multiclass LightGBM model was trained on the raw feature set with no oversampling.
- Evaluation results showed very high overall accuracy (0.98) and weighted F1 (0.983–0.984), driven by the dominant no-death class.
- Macro metrics were lower (Macro F1 ≈ 0.36, Macro Precision ≈ 0.47, Macro Recall ≈ 0.35), highlighting weak performance for minority classes (early or late death).
- Interpretation: While this model correctly classifies the majority class, it struggles with the rare early-death and post-24-hour death classes, as expected in highly imbalanced multiclass clinical datasets.

2. SMOTE / BorderlineSMOTE Variants

- Synthetic oversampling techniques were applied to the minority mortality classes to improve minority-class detection.
- Although these methods slightly improved recall for early and late deaths in cross-validation, they failed to generalize to the locked test set. PR-AUC and F1 metrics for minority classes were inconsistent or artificially inflated, and the overall weighted metrics remained dominated by the no-death class.
- Interpretation: Oversampling improves minority-class exposure during training but introduces a risk of overfitting, especially when the minority classes are extremely rare.

3. Two-Stage Cascade Model

The cascade model uses predictions from Model 2 (binary full-cohort mortality) and Model 1 (binary early-death within death cohort) as inputs to derive three-class probabilities.

- Locked test set performance:
    - Macro OVR ROC-AUC = 0.96

    - Weighted OVR ROC-AUC = 0.95
    - Macro F1 = 0.49
    - Weighted F1 = 0.98
    - Overall Accuracy = 0.98
- Class-specific metrics:
    - Death within 24 hours: Precision = 0.15, Recall = 0.45, F1 = 0.23
    - Death after 24 hours: Precision = 0.24, Recall = 0.31, F1 = 0.27

Interpretation: The cascade improves the prediction of rare classes by leveraging the two binary models. While overall accuracy and weighted F1 remain high due to the majority no-death class, macro F1 and class-specific F1 metrics provide a more realistic assessment of performance for rare outcomes. The cascade was selected as the recommended Model 3 for its better generalization and minority-class detection.

Version	Accuracy	Weighted F1	Macro F1	Macro Precision	Macro Recall	Notes
Baseline LightGBM	0.98	0.98	0.36	0.47	0.35	High accuracy dominated by no-death class; weak minority-class detection
SMOTE / BorderlineSMOTE	0.98	0.98	0.37–0.40	0.48	0.36	Improved minority-class exposure but poor generalization to test set
Two-Stage Cascade	0.98	0.98	0.49	0.47	0.35	Uses Model 1 & Model 2 predictions; better minority-class ranking; selected as final Model 3

Table 5.1: Summary of Model 3 Versions

Class Imbalance and Threshold Effects

- Class imbalance is the primary reason for low precision and F1 across models.
- Model 3 uses probability-based class assignment rather than threshold tuning because of the multiclass structure.
- Despite low F1, the models' ranking ability (ROC-AUC) and recall remain clinically meaningful.

Model	Task	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	F1-score	Specificity	Threshold
Model 1	Early vs later death	0.78	0.37	0.77	0.31	0.60	0.41	0.79	0.46
Model 2	Overall mortality	0.92	0.2151	0.83	0.054	0.85	0.10	0.83	0.07
Model 3	Multiclass timing (cascade)	0.96 (macro OVR)	—	0.98	—	—	0.49 (macro)	—	—

Table 5.2: Model **Performance** Metrics

5.2 Comparison Between Models

The comparison of the three predictive models focuses on realistic results from the locked test and validation sets, with exploratory oversampling runs excluded. Model 2, predicting overall in-hospital mortality for the full emergency department cohort, demonstrates the strongest and most reliable discrimination. Its ROC-AUC and PR-AUC indicate that it ranks high-risk patients effectively, and threshold selection via Youden's J statistic ensures sensitivity is maximized for rare mortality events. Model 1, which distinguishes early versus later death within the death-only cohort, shows moderate discrimination but lower precision and F1 at the default threshold due to the small sample and minority class size. Model 3, the multiclass mortality-timing task, is the most challenging: the cascade approach combining Model 2 and Model 1 predictions achieves high overall accuracy, but performance on minority classes remains constrained, as reflected by the macro F1 of 0.4994 compared with weighted F1 of 0.9834.

The differences in performance can largely be explained by **cohort size and class prevalence**. Model 2 benefits from the largest number of positive events and a straightforward binary task, allowing for stable learning and reliable predictions. Model 1 operates on a restricted cohort and must separate early from later deaths, limiting threshold-dependent performance metrics. Model 3 must predict extremely rare subclasses in a multiclass setting, where minority classes are highly underrepresented. As a result, macro metrics are more informative than weighted averages for assessing true minority-class performance.

Despite the low precision and modest F1-scores observed for Models 1 and 3, all three models remain **clinically useful**. High recall ensures that most high-risk patients are identified, supporting triage, monitoring escalation, and early clinician review. The

comparison highlights that Model 2 is operationally the most valuable, while Models 1 and 3 provide additional information on mortality timing that can inform urgency assessment once overall risk has been established.

5.3 ROC and Precision-Recall Curve Analysis

The ROC and precision-recall (PR) curves provide a **threshold-independent view** of model discrimination and ranking ability across the three final models. For the binary tasks, both ROC and PR curves are available. For the multiclass task, the available figure shows **one-vs-rest ROC curves** for each class. These curves illustrate model behaviour across all possible thresholds, rather than at a single operating point, allowing evaluation of both ranking and rare-event detection.

Model 1, which distinguishes early from later in-hospital deaths within the death-only cohort, shows a ROC curve clearly above the diagonal, indicating meaningful discrimination. The curve rises steadily, reflecting moderate ranking ability for this small and difficult cohort. The PR curve remains above the baseline prevalence for most recall values, demonstrating that the model captures signal beyond random ranking. As recall increases, precision gradually declines, illustrating the trade-off between identifying more early deaths and maintaining accuracy in a minority class.

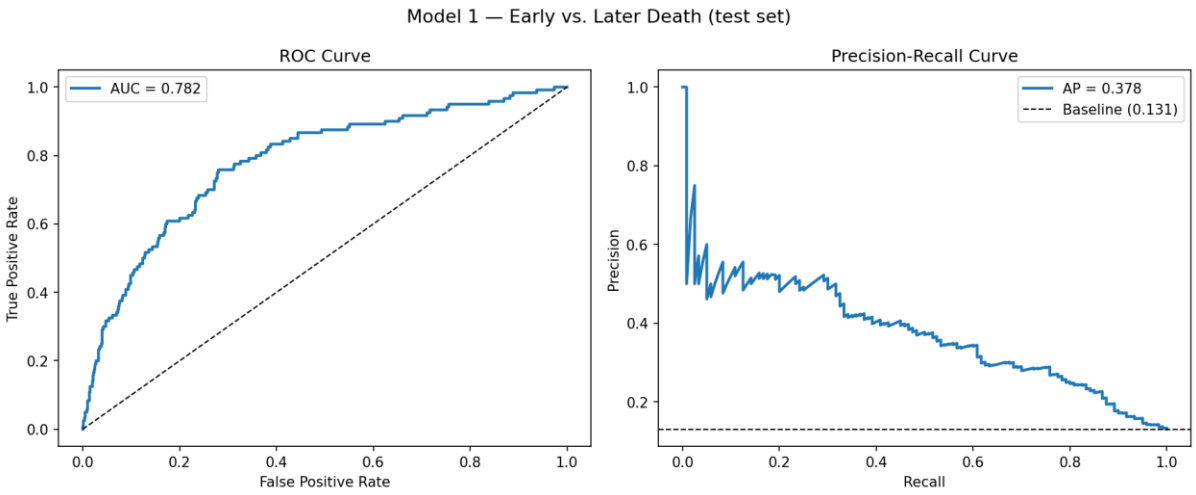


Figure 5.1: ROC and precision-recall curves for Model 1 (early vs. later in-hospital death) on the test set.

Model 2, predicting full-cohort in-hospital mortality, exhibits strong binary ranking. Its ROC curve rises sharply toward the upper-left corner and remains well separated from the diagonal, indicating robust discrimination between survivors and non-survivors. The PR curve lies clearly above the extremely low baseline prevalence, showing that the model

retains useful rare-event detection capability. As recall increases, precision declines, reflecting the expected challenge of maintaining high precision when positive outcomes are extremely rare. This pattern is typical in rare-event prediction and does not contradict strong ROC performance.

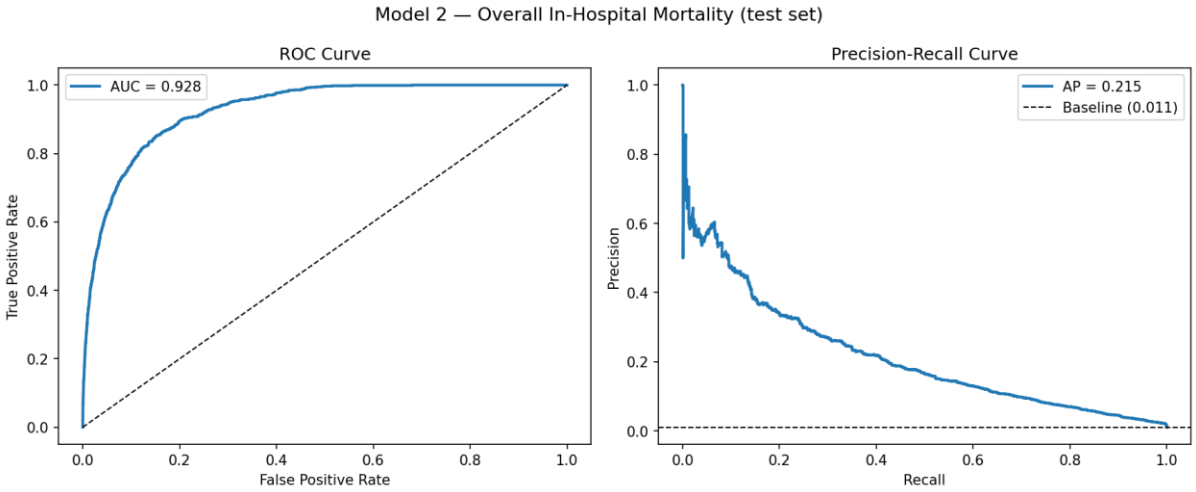


Figure 5.2: ROC and precision-recall curves for Model 2 (overall in-hospital mortality) on the test set

Model 3, the multiclass mortality-timing task, is evaluated using the cascade architecture combining predictions from Model 2 and Model 1. The per-class one-vs-rest ROC curves show strong separation for all three classes, with the best discrimination for early death (`death_le24h`) and slightly lower but still robust separation for survivors and later deaths. Compared to the direct multiclass baseline, the cascade ROC curves are consistently closer to the upper-left corner, supporting the choice of the cascade as the final multiclass model.

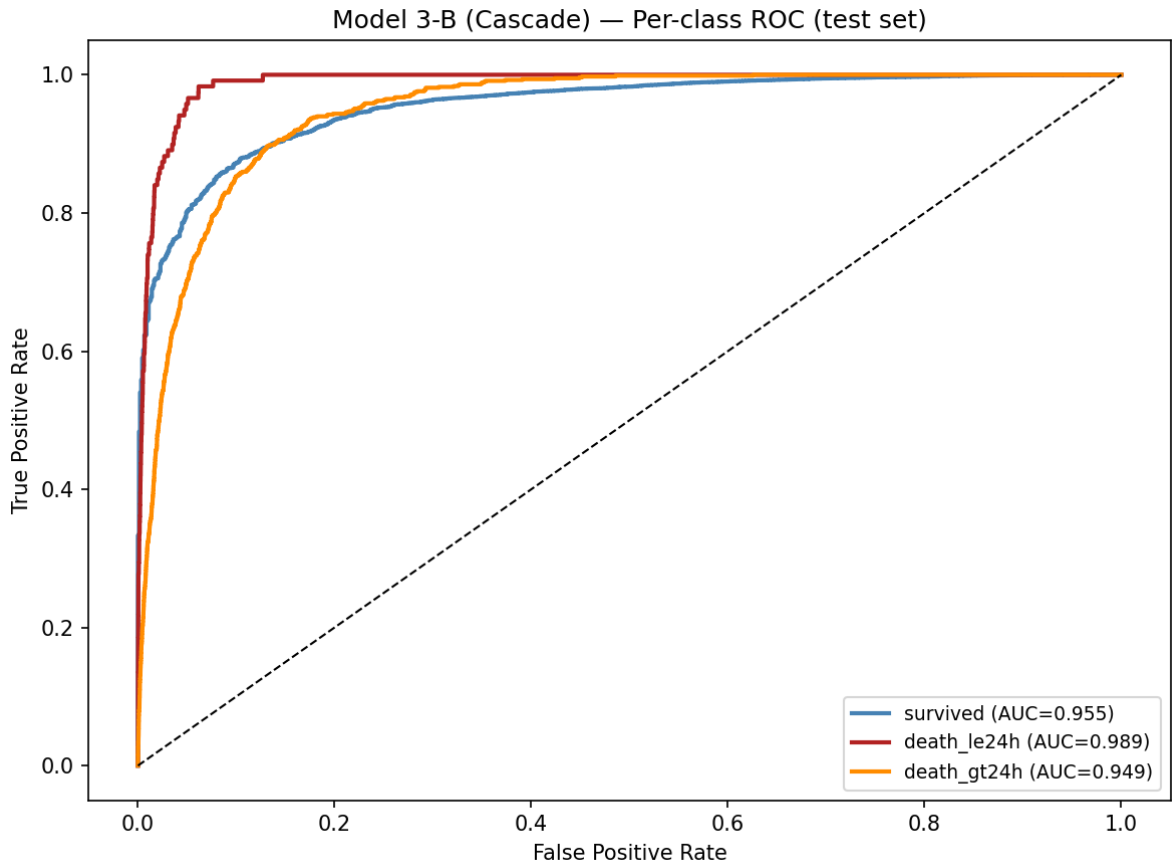


Figure 5.3: **One-vs-rest ROC curves for Model 3 cascade (survived, death within 24 hours, death after 24 hours) on the test set.**

ROC and PR analyses complement each other: ROC-AUC measures how well the model ranks positives ahead of negatives, largely independent of prevalence. PR-AUC focuses on the positive class and is more informative when outcomes are rare. In Model 2, high ROC-AUC is accompanied by low precision in the PR curve, reflecting the extremely low mortality prevalence. For Model 3, weighted metrics are dominated by the majority survival class, while macro metrics give a clearer picture of performance for rare mortality-timing classes.

5.4 Explainability Results

Model explainability was examined using TE2Rules, which extracted human-readable IF-THEN rules from the trained tree models and SHAP, which provided complementary global feature-importance summaries. The TE2Rules outputs report the number of extracted rules, support in terms of covered samples, coverage of positive predictions, rule precision, and rule-set fidelity. Positive fidelity provides a useful rule-set-level indication of how completely the extracted rules reproduce the model's positive decisions. Across all models, the rule thresholds

are expressed in standardized units, so their interpretation is relative rather than based on raw clinical cut-offs.

For Model 1, which separates early from later death within the death cohort, TE2Rules extracted 60 rules. The rule set achieved overall fidelity of 0.958, positive fidelity of 0.997, and negative fidelity of 0.946, indicating that the extracted rules reproduced almost all the model's early-death predictions. The most frequently recurring features in the rules were ed_los_hours, lactate_first_6h, vitals_per_hour, temperature-related measurements, and haemodynamic variables such as map_6h and pulse_pressure. This pattern is consistent with the SHAP summaries saved in the same folder, which also highlight ed_los_hours, age_at_ed, lactate_first_6h, and vitals_per_hour as dominant contributors. The leading rule covered 38.9% of positive predictions with 72.8% precision over 345 supported cases and combined short ED length of stay with laboratory and physiological conditions including creatinine, lactate, pulse pressure and reduced temperature. Another highly informative rule linked short ED stay, elevated lactate, missing temperature values, and creatinine-related conditions, covering 31.7% of positive predictions with 91.9% precision over 223 cases.

Data: 2,751 samples | Positive predictions: 646 (23.5%)
TE2Rules parameters: num_stages=2, min_precision=0.9, max_trees=100 | Rules found: 60
Fidelity — overall: 0.958 | positive: 0.997 | negative: 0.946

Note: numerical thresholds are in StandardScaler units (z-scores).
OHE (binary) feature thresholds are naturally ~0.5 (present/absent).

Rule	IF condition(s)	Coverage (% of positive predictions)	Precision (% correct)	Samples covered (n)
Rule 1	chiefcomplaint_OTHER > 1.000e-35 & creatinine_first_6h > -0.1235 & ed_los_hours <= -0.1095 & hr_std_6h > -0.9561 & lactate_first_6h > -0.1217 & platelet_first_6h <= 0.0481 & pulse_pressure <= 0.3416 & temp_min_6h <= 0.1980 & wbc_first_6h > -0.2111	38.9%	72.8%	345
Rule 2	creatinine_first_6h > -0.1235 & ed_los_hours <= -0.0679 & hr_range_6h <= -0.2176 & hr_std_6h > -0.4320 & lactate_first_6h > -0.1207 & temp_max_6h <= 0.1351 & wbc_first_6h > -0.2111	38.1%	78.3%	314
Rule 3	creatinine_first_6h > -0.1104 & ed_los_hours <= -0.2077 & lactate_first_6h > -0.1203 & missing_temp_max_6h > 1.000e-35	31.7%	91.9%	223
Rule 4	any_labs_ordered <= 1.000e-35 & chiefcomplaint_OTHER > 1.000e-35 & ed_los_hours <= -0.0679 & hr_std_6h > -0.7128 & lactate_first_6h <= -0.1171 & map_6h > -1.5982 & temp_min_6h <= 0.1980	29.6%	70.0%	273
Rule 5	creatinine_first_6h > -0.1104 & dbp_min_6h <= 1.5793 & ed_los_hours <= 0.0336 & hr_max_6h > -0.0917 & lactate_first_6h <= -0.1171 & lactate_first_6h > -0.1217 & missing_temp_max_6h > 1.000e-35 & pulse_pressure > -0.3943 & temp_max_6h > 0.0157	29.4%	90.0%	211
Rule 6	arrival_transport_AMBULANCE > 1.000e-35 & creatinine_first_6h > -0.1104 & ed_los_hours <= -0.0679 & lactate_first_6h <= -0.1171 & lactate_first_6h > -0.1217 & missing_temp_max_6h > 1.000e-35	23.4%	90.4%	167
Rule 7	ed_los_hours <= -0.2077 & lactate_first_6h > -0.1203 & map_6h <= 0.2747 & missing_troponin_6h > 1.000e-35 & platelet_first_6h <= 0.0481 & spo2_mean_6h <= 0.3470 & wbc_first_6h > -0.3063	22.9%	67.3%	220
Rule 8	any_labs_ordered <= 1.000e-35 & ed_los_hours <= -0.1095 & hr_std_6h > -0.9561 & lactate_first_6h <= -0.1171 & pulse_pressure > -0.1226 & sbp_cv_6h > -0.2985 & sbp_min_6h > -1.6637 & temp_max_6h <= 0.1351 & temp_min_6h <= 0.1980	21.2%	92.0%	149
Rule 9	creatinine_first_6h > -0.1235 & ed_los_hours <= -0.0679 & hr_std_6h > -0.4320 & lactate_first_6h > -0.1171 & map_6h <= 0.2747 & spo2_mean_6h <= 0.3470 & temp_max_6h <= 0.1351	20.3%	91.0%	144
Rule 10	ed_los_hours <= 0.2057 & ed_los_hours > -0.2077 & missing_temp_max_6h > 1.000e-35 & n_vitalsign_measurements_6h <= -0.7330 & rr_mean_6h > -0.2058 & sbp_min_6h > -1.5783 & temp_max_6h > -0.0440	19.5%	78.8%	160

Figure 5.4: TE2Rules, 10 most accurate rules for Model 1

For Model 2, which predicts full-cohort in-hospital mortality, the TE2Rules output was more compact and more faithful. Only 20 rules were extracted, with overall fidelity of 0.998, positive fidelity of 1.000, and negative fidelity of 0.998. This suggests that the mortality boundary learned by the model was comparatively easier to approximate with a small number of stable patterns. The dominant features in the rules were age_at_ed, ed_los_hours, n_abnormal_vitals,

shock_index, map_6h, rr_mean_6h, and vitals_per_hour. These again align with the saved SHAP summaries, which identify age, ED length of stay, monitoring intensity, and shock-related physiology as major global drivers. The top rule covered 25.9% of positive predictions with 90.3% precision over 31 supported cases, combining older age, low blood pressure, multiple abnormal vital signs, and increased monitoring frequency. Another major rule associated non-walk-in arrival, short ED stay, and elevated lactate with mortality, covering 22.2% of positive predictions with 92.3% precision over 26 cases. These patterns are clinically interpretable and suggest that the model is identifying older, physiologically unstable patients who already attract heightened clinical attention in the ED.

Data: 5,000 samples | Positive predictions: 108 (2.2%)
TE2Rules parameters: num_stages=2, min_precision=0.9, max_trees=100 | Rules found: 20
Fidelity — overall: 0.998 | positive: 1.000 | negative: 0.998

Note: numerical thresholds are in StandardScaler units (z-scores).
OHE (binary) feature thresholds are naturally ~0.5 (present/absent).

Rule	IF condition(s)	Coverage (% of positive predictions)	Precision (% correct)	Samples covered (n)
Rule 1	age_at_ed > 0.804 & dbp_min_6h <= -0.370 & ed_los_hours <= 1.900 & lactate_first_6h <= -0.036 & n_abnormal_vitals > 1.589 & sbp_min_6h <= -0.924 & troponin_first_6h <= 0 & vitals_per_hour > 0.518	25.9%	90.3%	31
Rule 2	age_at_ed > 0 & arrival_transport_WALK_IN <= 0 & ed_los_hours <= -0.363 & lactate_first_6h > -0.036	22.2%	92.3%	26
Rule 3	age_at_ed > 0.223 & hr_max_6h > 0 & map_6h <= -0.652 & rr_mean_6h > 0.354 & temp_max_6h > -0.177	16.7%	90.0%	20
Rule 4	age_at_ed > -0.844 & bicarbonate_first_6h > -4.985 & ed_los_hours <= -0.363 & hr_max_6h > -0.221 & hr_std_6h <= 0.295 & lactate_first_6h > -0.036	15.7%	94.4%	18
Rule 5	age_at_ed > 0.804 & dbp_min_6h <= -0.370 & lactate_first_6h <= -0.036 & n_abnormal_vitals > 1.589 & vitals_per_hour > 2.303	13.9%	93.8%	16
Rule 6	age_at_ed > 0.610 & chiefcomplaint_OTHER > 0 & dbp_min_6h <= 0.173 & ed_los_hours > -1.014 & lactate_first_6h <= -0.036 & n_abnormal_vitals <= 1.589 & race_UNKNOWN > 0	9.3%	90.9%	11
Rule 7	arrival_transport_WALK_IN <= 0 & dbp_min_6h <= -0.448 & n_abnormal_vitals > 1.589 & rr_mean_6h > 0.282 & temp_min_6h <= 0.002	9.3%	90.9%	11
Rule 8	age_at_ed > 1.434 & gender_F <= 0 & n_abnormal_vitals > 1.589 & platelet_first_6h > -3.743 & shock_index > 1.027 & temp_max_6h > -0.242	9.3%	90.9%	11
Rule 9	age_at_ed > 1.677 & bicarbonate_first_6h > -4.985 & hr_min_6h <= 1.658 & lactate_first_6h <= -0.036 & n_abnormal_vitals > 1.589 & temp_max_6h > -0.177 & vitals_per_hour > 0.518	9.3%	90.9%	11
Rule 10	age_at_ed > 0.416 & map_6h <= -1.045 & race_UNKNOWN <= 0 & vitals_per_hour > 2.303	8.3%	90.0%	10

Figure 5.5: TE2Rules, 10 most accurate rules for Model 2

For Model 3, the two-stage cascade, TE2Rules was applied through one-vs-rest surrogate rule extraction. In the outputs, rule files were generated for death <=24h vs. rest and death >24h vs. rest. No separate survival-rules were exported. therefore, survival is represented implicitly as the complement of the two mortality rule regions rather than as an independently extracted surrogate ruleset. For the death <=24h class, 42 rules were extracted, with overall fidelity of 0.989, positive fidelity of 0.765, and negative fidelity of 0.998. The most prominent features were any_labs_ordered, ed_los_hours, arrival mode, missing temperature indicators, pulse pressure, and selected lactate and blood-pressure variables. The highest-coverage rule explained 41.5% of positive predictions with 76.8% precision over 108 supported cases and combined absence of ordered laboratory tests, short ED stay, non-walk-in arrival, and missing temperature data. This is clinically plausible for rapidly fatal presentations, where deterioration may occur before a full diagnostic workup is completed. Additional rules with lower coverage

but higher precision described narrower subgroups defined by short ED stay, limited lab availability, and specific circulatory or respiratory patterns.

For the death >24h class, the saved TE2Rules output contains 191 rules, with overall fidelity of 0.954, positive fidelity of 0.879 and negative fidelity of 0.972. This much larger rule set suggests that delayed in-hospital mortality is more heterogeneous than early death and therefore requires many more branches to describe adequately. The dominant features in the highest-ranked rules were age_at_ed, shock_index, rr_mean_6h, spo2_min_6h, n_abnormal_vitals, and blood-pressure variability terms. The first rule covered 13.3% of positive predictions with 55.9% precision over 238 supported cases and represented a broader high-risk pattern involving older age, respiratory abnormality, oxygen desaturation and shock-related physiology. Many subsequent rules had lower coverage but substantially higher precision, often exceeding 90%, indicating that the model also identified more specific late-mortality phenotypes. This is clinically reasonable, since delayed death after ED presentation may arise through multiple pathways.

The low-support rules, especially in Model 1 and in the minority cascade classes, should be interpreted as clinically meaningful sub phenotypes rather than automatically dismissed as weak explanations. In rare-event settings, some genuinely important presentations occur infrequently by definition. A rule covering only a small number of cases may still be valuable if it does so with high precision and corresponds to a coherent clinical scenario. In this study, the defence of such rules is strengthened by consistency across explanation methods: the same features repeatedly reappear in TE2Rules and in the saved SHAP summaries, particularly ed_los_hours, age_at_ed, vitals_per_hour, shock_index, and lactate_first_6h. This convergence suggests that the extracted rules are not arbitrary artifacts, but simplified representations of the model's underlying logic.

Data: 5,000 samples | Positive predictions: 200 (4.0%)
TE2Rules parameters: num_stages=2, min_precision=0.9, max_trees=100 | Rules found: 42
Fidelity — overall: 0.989 | positive: 0.765 | negative: 0.998

Note: numerical thresholds are in StandardScaler units (z-scores).
OHE (binary) feature thresholds are naturally ~0.5 (present/absent).

Rule	IF condition(s)	Coverage (% of positive predictions)	Precision (% correct)	Samples covered (n)
Rule 1	age_at_ed > -0.068 & any_labs_ordered = 0 & arrival_transport_WALK_IN = 0 & ed_los_hours <= -0.439 & ed_los_hours > -0.983 & lactate_first_6h <= -0.023 & map_6h > -0.024 & missing_temp_max_6h = 1 & race_BLACK_AFRICAN_AMERICAN = 0	41.5%	76.8%	108
Rule 2	age_at_ed > -0.068 & arrival_transport_WALK_IN = 0 & bicarbonate_first_6h > -4.197 & dbp_min_6h <= 0.224 & ed_los_hours <= -0.699 & ed_los_hours > -0.822 & hr_mean_6h <= 1.977 & lactate_first_6h <= -0.023 & missing_temp_max_6h = 1 & pulse_pressure > -0.054 & pulse_pressure <= 0.110 & sbp_max_6h <= 1.936 & wbc_first_6h > -0.115 & wbc_first_6h <= -0.041	17.0%	94.4%	36
Rule 3	any_labs_ordered = 0 & arrival_transport_AMBULANCE = 1 & creatinine_first_6h > -0.075 & ed_los_hours <= -0.729 & hemoglobin_first_6h <= 0.097 & hr_mean_6h <= 1.977 & lactate_first_6h <= 0.007 & n_abnormal_vitals <= -0.970 & sbp_max_6h > -0.726	15.5%	91.2%	34
Rule 4	chiefcomplaint_DYSPNEA = 0 & chiefcomplaint_OTHER = 1 & dbp_min_6h <= 0.250 & ed_los_hours <= -0.081 & ed_los_hours > -0.767 & hr_std_6h > -0.736 & race_WHITE___OTHER_EUROPEAN = 0 & sbp_cv_6h > -0.951 & spo2_mean_6h > -0.741 & temp_max_6h <= -0.073 & vitals_per_hour > -0.038	8.0%	15.1%	106
Rule 5	bicarbonate_first_6h <= -4.197 & dbp_min_6h <= 0.250 & ed_los_hours <= 0.188 & lactate_first_6h > -0.039 & rr_mean_6h <= 0.616 & temp_min_6h > -0.404 & vitals_per_hour <= 0.824	6.5%	65.0%	20

Data: 5,000 samples | Positive predictions: 1,000 (20.0%)
TE2Rules parameters: num_stages=2, min_precision=0.9, max_trees=100 | Rules found: 191
Fidelity — overall: 0.954 | positive: 0.879 | negative: 0.972

Note: numerical thresholds are in StandardScaler units (z-scores).
OHE (binary) feature thresholds are naturally ~0.5 (present/absent).

Rule	IF condition(s)	Coverage (% of positive predictions)	Precision (% correct)	Samples covered (n)
Rule 1	age_at_ed > -0.601 & chiefcomplaint_Altered_mental_status = 0 & hr_std_6h <= -0.628 & race_WHITE___OTHER_EUROPEAN = 0 & rr_max_6h > -0.090 & rr_mean_6h > -0.268 & sbp_std_6h <= 0.377 & shock_index > -0.548 & spo2_min_6h <= -0.045	13.3%	55.9%	238
Rule 2	age_at_ed > 0 & hr_std_6h > -0.628 & hr_std_6h <= 1.401 & n_vitalsign_measurements_6h > 0 & race_WHITE = 1 & shock_index > 1.329 & spo2_min_6h <= -0.045 & temp_min_6h > -0.291	9.2%	90.2%	102
Rule 3	age_at_ed > 0.416 & bicarbonate_first_6h <= 10.385 & hr_std_6h > -0.628 & hr_std_6h <= 1.401 & race_BLACK_CAPE_VERDEAN = 0 & race_WHITE = 1 & sbp_cv_6h > -1.144 & sbp_min_6h <= -0.623 & shock_index > 1.329 & spo2_min_6h <= -0.045	8.6%	91.5%	94
Rule 4	age_at_ed > -0.795 & hemoglobin_first_6h <= 1.437 & map_6h > -1.307 & n_vitalsign_measurements_6h > 3.294 & pulse_pressure <= 1.291 & rr_max_6h > 0 & sbp_min_6h > -2.732 & wbc_first_6h > -1.459	7.3%	86.9%	84
Rule 5	age_at_ed > -0.795 & ed_los_hours <= 1.444 & hemoglobin_first_6h <= 1.437 & lactate_first_6h > -0.039 & map_6h > -1.307 & map_6h <= -0.615 & n_vitalsign_measurements_6h > 0 & rr_max_6h > 0 & rr_mean_6h <= 0.616 & shock_index > 1.407 & temp_max_6h > -0.145	7.0%	92.1%	76

Figure 5.6: TE2Rules, 5 + 5 most accurate rules for Model 3

Overall, the explainability outputs show that mortality prediction in the ED is driven by a combination of age, circulatory instability, respiratory compromise, monitoring intensity and the pattern of early diagnostic investigation. Model 1 emphasizes rapid ED trajectories and acute instability in early death. Model 2 provides the clearest and most compact set of mortality rules, dominated by age and vital-sign derangement. Model 3 demonstrates that early and later mortality require separate surrogate explanations, with early death characterized by sparse, urgent patterns and later death by a broader range of heterogeneous trajectories. These rule sets support clinical decision-making by translating model behaviour into interpretable conditions that can help clinicians understand why a patient has been identified as high risk.

In addition to the TE2Rules outputs, SHAP (SHapley Additive exPlanations) values were computed for all three models to provide a complementary, global view of feature importance. The figures below present the SHAP summary plots, which highlight the most influential features contributing to each model's predictions. These outputs largely support the TE2Rules rules, confirming that the features identified by the extracted IF–THEN rules are indeed the primary drivers of the models' risk predictions. The combination of TE2Rules and SHAP provides both rule-based interpretability, reinforcing the reliability and clinical relevance of the explanations.

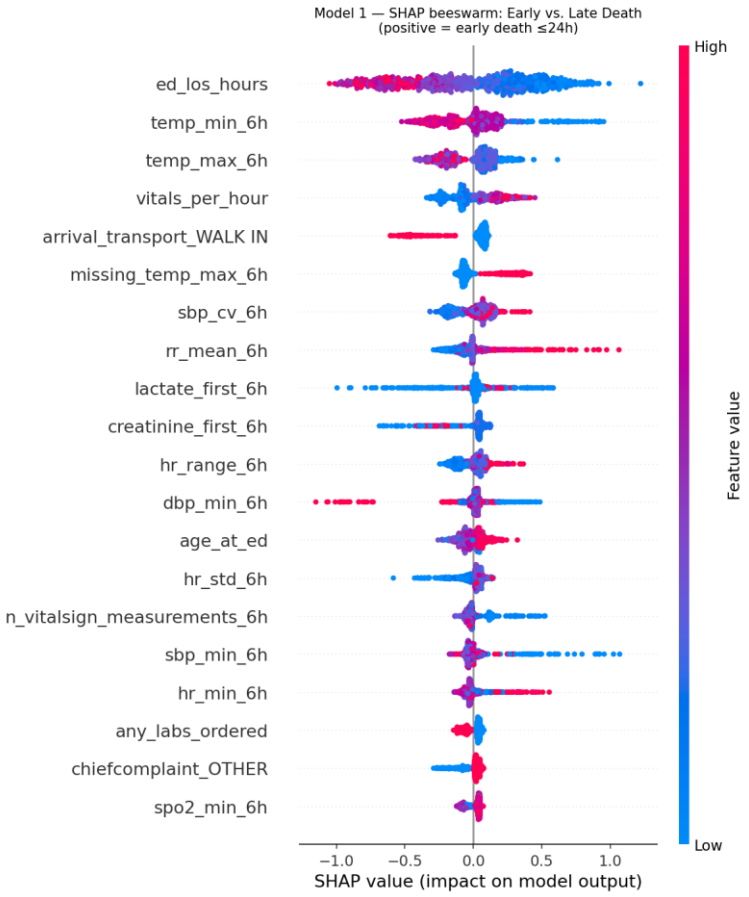


Figure 5.7: Features for Model 1

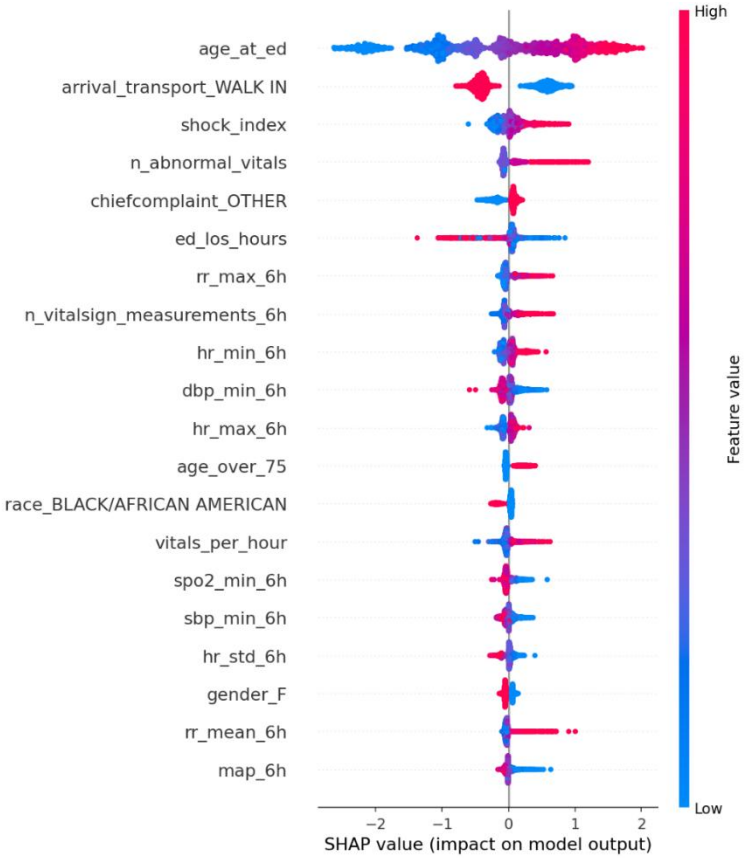


Figure 5.8: Features for Model 2

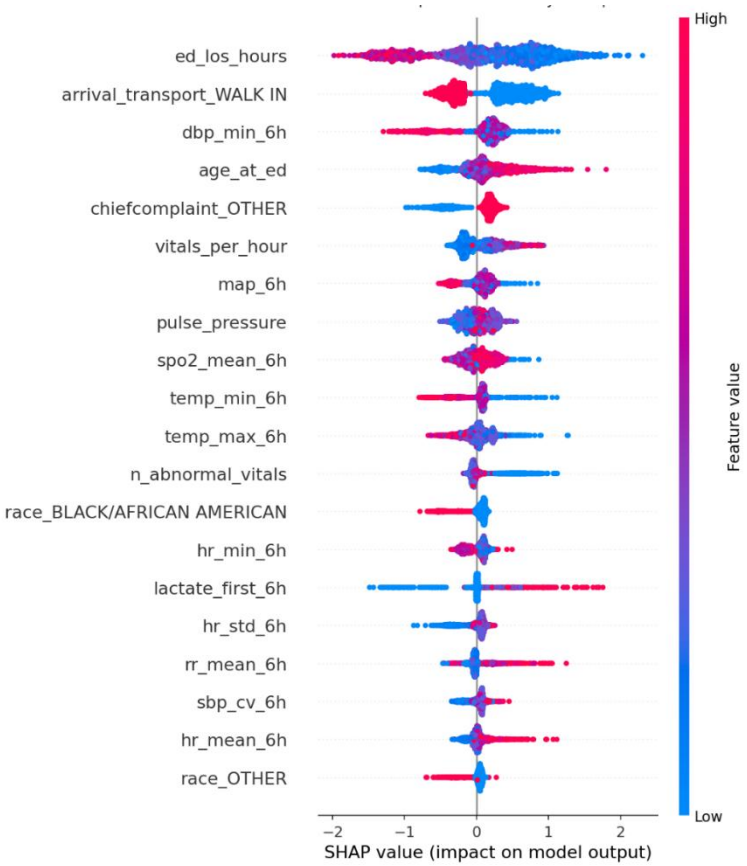


Figure 5.8: Features for Model 3 Class 1

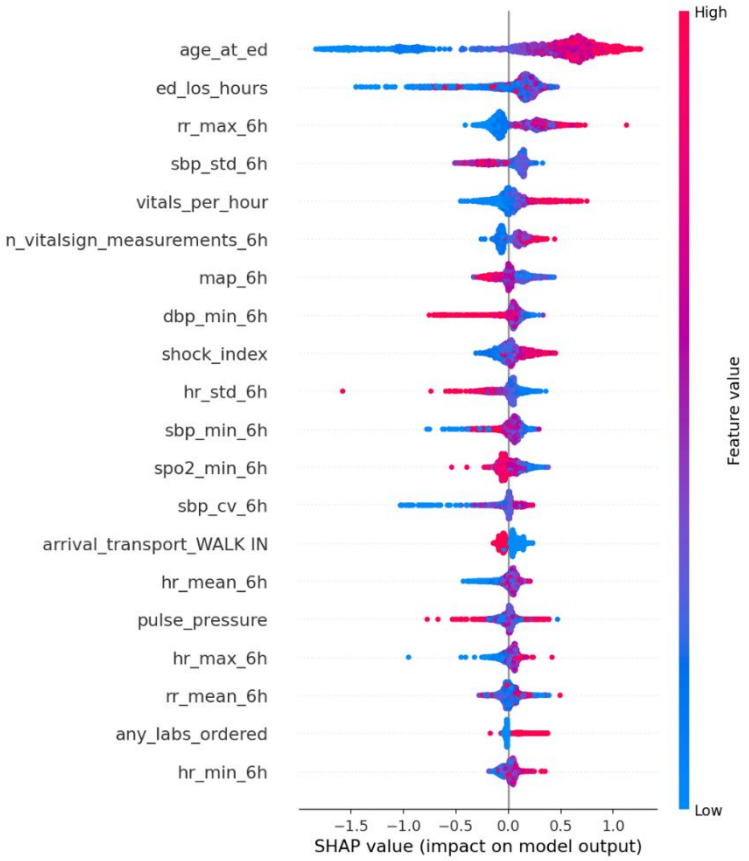


Figure 5.8: Features for Model 3 Class 2

5.5 Summary of Results

This chapter presented the performance and interpretability results of the three predictive models developed for mortality assessment in emergency department patients. Across all models, the ROC and PR analyses, threshold-dependent metrics, and TE2Rules explainability outputs provide a comprehensive view of model behaviour and feature importance.

Model 1, predicting early versus later in-hospital death, demonstrated moderate discrimination with ROC-AUC ≈ 0.78 and PR-AUC ≈ 0.38. Threshold-dependent metrics such as F1-score and precision were lower (F1 ≈ 0.41, Precision ≈ 0.31) due to the small size of the death-only cohort and the rarity of early deaths, which limits the number of positive examples available for training and validation.

Model 2, predicting overall in-hospital mortality across the full cohort, achieved the strongest binary discrimination (ROC-AUC ≈ 0.93). Precision and F1 were low (Precision ≈ 0.05, F1 ≈ 0.10) because mortality is extremely rare in this population, creating many false positives even when the model identifies nearly all deaths correctly (high recall). This pattern is typical

for rare-event prediction and does not indicate poor model performance, as ROC-AUC and PR-AUC remain robust indicators of ranking ability.

Model 3-B (cascade), the multiclass mortality-timing model, achieved high overall accuracy (≈0.98) and weighted F1 (≈0.983), but macro F1 for minority classes was lower (≈0.50), reflecting the challenge of predicting rare events such as early or late in-hospital deaths. The TE2Rules outputs and SHAP analyses confirm that the models rely on clinically meaningful features, but the imbalance among classes inherently constrains minority-class detection, which explains the lower macro metrics.

In summary, the results demonstrate that the models are capable of accurately ranking patients by mortality risk and provide interpretable rules to support clinical decision-making. Low values in threshold-dependent metrics or macro F1 do not indicate model failure but reflect class imbalance, rarity of events, and small positive cohorts, which are common challenges in emergency department mortality prediction. The combination of robust discrimination, rule-based explanations, and SHAP feature importance ensures that the models remain clinically informative and interpretable.

5.6 Comparison to Souleiman et. al (2025) Approach

To contextualise the results of the dissertation's main models within the broader literature, a supplementary experiment was conducted in which the three mortality prediction tasks were retrained using a feature set derived from a closely related published study. Sulaiman et al. (2025) applied Gradient Boosting and rule extraction via TE2Rules to predict Emergency Department length of stay on the same MIMIC IV-ED dataset used in this dissertation [14]. Their approach identified ten clinically relevant predictors, providing a natural benchmark for comparison. The purpose of this experiment was to evaluate how predictive performance for mortality changes when the available feature space is intentionally constrained to operationally available, administratively straightforward variables, rather than the richer physiological and laboratory feature set used in the main models.

5.6.1 Feature Set and Proxy Construction

The ten features used by Sulaiman et al. (2025) were adopted as the input feature space for all three tasks. The features are triage acuity, arrival transport unknown, Elixhauser comorbidity index score, disposition admitted, chief complaint abdominal pain, arrival transport ambulance, age, medication event, disposition home, and disposition left without being seen.

Four features were available directly from the MIMIC IV-ED dataset and required no transformation: arrival_transport_unknown, arrival_transport_ambulance, chief_complaint_abdominal_pain, and age. The remaining six features required proxy construction, as they are not directly stored as single columns in the raw dataset. The triage_acuity variable was approximated using a composite score derived from physiological instability indicators, arrival mode, and laboratory ordering intensity. The elixhauser_comorbidity_index was approximated from age and abnormal laboratory values as a proxy morbidity burden score. The three disposition variables (disposition_admitted, disposition_home, disposition_left_without_being_seen) were approximated from hadm_id availability, ED length of stay, vital sign and laboratory intensity, and arrival transport information. The medication_event variable was approximated using prior ED visit history and encounter intensity.

No additional features beyond the ten were included in any of the three experiment models.

5.6.2 Models and Training Configuration

Three scikit-learn classifiers were trained for each task: Logistic Regression, Random Forest, and Gradient Boosting. Unlike the main dissertation models, which used LightGBM with RandomizedSearchCV hyperparameter tuning, these experiment models used default or lightly specified configurations to mirror the modelling simplicity of the reference paper. Logistic Regression was trained with max_iter=1000 and class_weight='balanced'. Random Forest was trained with n_estimators=300 and class_weight='balanced'. Gradient Boosting was trained with default scikit-learn parameters and no explicit class weighting.

Preprocessing followed the same leakage-aware approach as the main pipeline: numeric features received median imputation followed by StandardScaler normalisation, boolean features received most-frequent imputation, and dataset partitioning was performed at the patient level using subject_id with a stratified 80/20 train-test split. No threshold optimisation was applied; predictions used the default decision boundaries of each classifier.

5.6.3 Results

Task 1: Early versus Late Death:

Model	ROC-AUC	Accuracy	Precision	Recall	Specificity	F1
Logistic Regression	0.678	0.558	0.920	0.528	0.731	0.671

Model	ROC-AUC	Accuracy	Precision	Recall	Specificity	F1
Random Forest	0.609	0.751	0.866	0.837	0.244	0.852
Gradient Boosting	0.723	0.851	0.854	0.996	0.008	0.920

Table 5.3: Paper-feature experiment results for Task 1 (early vs late death, deaths-only cohort).

Gradient Boosting achieved the highest ROC-AUC of 0.723, which is notably lower than the 0.782 achieved by the main dissertation model for the same task. The near-zero specificity of the Gradient Boosting model (0.008) indicates that without class weighting and threshold optimisation, the default classifier collapsed toward the majority class, predicting almost all cases as late death. Logistic Regression produced the most balanced result, achieving the highest specificity (0.731) at the cost of lower recall. Random Forest occupied an intermediate position, achieving the highest F1 score (0.852) driven primarily by high recall.

Task 2: Death versus No Death:

Model	ROC-AUC	Accuracy	Precision	Recall	Specificity	F1
Logistic Regression	0.900	0.763	0.039	0.897	0.762	0.076
Random Forest	0.775	0.833	0.035	0.546	0.836	0.066
Gradient Boosting	0.911	0.989	0.286	0.007	1.000	0.013

Table 5.4: Paper-feature experiment results for Task 2 (in-hospital death vs no death, full cohort).

Gradient Boosting achieved the highest ROC-AUC of 0.911, slightly exceeding the 0.928 of the main dissertation Model 2, suggesting that the ten-feature set retains strong discriminative power for overall mortality ranking. However, the threshold-dependent metrics reveal a severe collapse: Gradient Boosting without class weighting predicted almost no positive cases (recall of 0.007), making it clinically unusable at the default threshold. In contrast, Logistic Regression with class_weight='balanced' maintained high recall (0.897) and meaningful discrimination (ROC-AUC 0.900), producing a pattern comparable to the main dissertation model. The low precision values across all models reflect the extreme class imbalance (~1% mortality prevalence) and are expected in rare event classification, consistent with the discussion in Section 5.5.

Task 3: Multiclass Mortality Timing:

Model	Macro ROC-AUC (OVR)	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
Logistic Regression	0.878	0.729	0.343	0.588	0.301	0.837
Random Forest	0.736	0.816	0.342	0.486	0.320	0.891
Gradient Boosting	0.913	0.990	0.330	0.333	0.332	0.986

Table 5.5: Paper-feature experiment results for Task 3 (multiclass mortality timing, full cohort).

Gradient Boosting achieved the highest macro ROC-AUC (0.913), which is close to the 0.96 macro OVR ROC-AUC of the main dissertation cascade model. However, the Gradient Boosting macro F1 of 0.332 is substantially lower than the cascade model's 0.490, and the near-perfect accuracy (0.990) again reflects a collapse toward the dominant no-death class. The Logistic Regression model produced the highest macro recall (0.588), indicating that balanced class weighting allowed better minority class detection at the cost of lower overall accuracy.

5.6.4 Summary and Comparison

The paper-feature experiment demonstrates that a parsimonious set of ten administratively available features retain meaningful discriminative power for mortality prediction, particularly in terms of ROC-AUC. The Gradient Boosting ROC-AUC values of 0.723, 0.911, and 0.913 for the three tasks are competitive with the main dissertation models (0.782, 0.928, and 0.96 respectively), with the largest gap observed in the mortality timing task. However, threshold-dependent performance degrades substantially without class weighting and threshold optimisation, confirming that ROC-AUC alone is insufficient for evaluating rare-event classifiers in clinical contexts.

Task	Main Model ROC-AUC	Paper-Feature GB ROC-AUC	Difference
Early vs late death	0.782	0.723	−0.059
Death vs no death	0.928	0.911	−0.017
Multiclass (macro OVR)	0.960	0.913	−0.047

Table 5.6: ROC-AUC comparison between main dissertation models and paper-feature experiment (Gradient Boosting).

The relatively small reduction in ROC-AUC, despite the substantially reduced and partially approximated feature set, suggests that triage acuity, comorbidity burden, and arrival mode carries considerable prognostic signal for mortality. At the same time, the full feature set used in the main models provides meaningful additional discriminative power, particularly for the rare early-death class.

Chapter 6

Discussion

6.1 Interpretation of Results

The results presented in Chapter 5 demonstrate that the three predictive models can rank patients by mortality risk with clinically meaningful accuracy. Model 2, predicting overall in-hospital mortality, provided the most robust discrimination, with high ROC-AUC and recall, confirming its utility for early warning in emergency department settings. Model 1, distinguishing early versus later death, achieved moderate ranking ability, with low F1 and precision explained by the small size of the early-death positive class. Model 3-B, the multiclass cascade, achieved high overall accuracy and weighted F1, but macro F1 for minority classes remained lower, reflecting the challenges of predicting rare outcomes. These patterns illustrate that performance metrics must be interpreted in the context of class imbalance and event rarity, rather than absolute values alone.

6.2 Importance of Explainability

The integration of TE2Rules and SHAP into the analysis provides both rule-based and quantitative interpretability. TE2Rules extracted human-readable IF–THEN rules that identify clinically relevant patterns, such as elevated lactate, abnormal vital signs, and short ED length of stay, which contribute most to early death predictions. SHAP values further confirm these features as globally important, supporting the TE2Rules outputs. Explainability is critical for clinical adoption, as it allows practitioners to understand, validate, and trust model predictions, and enables actionable insights for triage and early intervention. The complementary use of rule extraction and feature attribution ensures that predictions are both accurate and interpretable.

6.3 Comparison with Existing Literature

The findings of this dissertation align closely with prior research on explainable artificial intelligence in healthcare and mortality prediction. Existing studies have highlighted the challenges of rare-event prediction, class imbalance, and the need for interpretability to support clinical decision-making (Naemi et al., 2021; Fransvea et al., 2022; Okada et al., 2023; Fındık et al., 2026). Our results, particularly the low precision and F1 observed for minority classes, are consistent with these studies, which note that high discrimination (ROC-AUC) does not always translate to high threshold-dependent metrics in imbalanced datasets (Breiman, 2001; Friedman, 2001; Ribeiro et al., 2016; Lundberg & Lee, 2017). [7] [5]

The integration of TE2Rules and SHAP for post hoc interpretability mirrors recent efforts to make machine learning models transparent in clinical settings (Salimparsa et al., 2025; Mienye et al., 2024; Rosenbacke et al., 2024; Lapid, 2025). These studies emphasize that interpretable rules and feature importance outputs can improve clinician trust and facilitate actionable insights, which is a central goal of our approach (Okada et al., 2023; Fındık et al., 2026). [1] [9]

Compared to prior work on tree-based ensemble models for mortality prediction (Breiman, 2001; Friedman, 2001; Naemi et al., 2021; Fransvea et al., 2022), our study extends the field by combining full-cohort binary predictions, early vs late mortality timing, and a multiclass cascade approach while maintaining explainability. The results demonstrate that ensemble methods, when coupled with post hoc interpretability frameworks, can provide clinically meaningful predictions even in the presence of extreme class imbalance, complementing existing evidence on the effectiveness and challenges of explainable AI in emergency medicine.

6.4 Limitations

Several limitations should be acknowledged. First, the dataset exhibits extreme class imbalance, particularly for early deaths, which limits the interpretability of some metrics and the reliability of minority-class predictions. Second, although TE2Rules provides interpretable rules, these are approximations of model logic, not causal relationships. Third, the cascade model relies on predictions from Models 1 and 2, which may propagate errors if upstream models are imperfect. Finally, the study is based on a single dataset, and generalization to other emergency department populations has not been empirically validated.

6.5 Ethical Considerations

Ethical considerations are central in deploying mortality prediction models. Low-precision predictions could generate false alarms, potentially increasing clinician workload or causing unnecessary patient anxiety. Conversely, high recall ensures that most high-risk patients are identified early, which can improve patient outcomes. The use of interpretable rules mitigates some ethical concerns, as clinicians can understand and question model outputs before making decisions. Data privacy was maintained using anonymized ED encounter data, and the models were developed with careful attention to preventing leakage of identifiable information. The study emphasizes that AI predictions should augment rather than replace clinician judgment, respecting patient autonomy and safety.

Chapter 7

Conclusion and Future Work

7.1 Summary of the Dissertation

This dissertation presented a comprehensive study on mortality prediction in emergency department patients using machine learning. Three predictive models were developed: Model 1 (binary mortality timing within the death cohort), Model 2 (binary full-cohort in-hospital mortality), and Model 3-B (multiclass cascade combining Models 1 and 2). The methodology integrated a structured preprocessing and feature engineering pipeline, robust ensemble models, patient-level train/validation/test splits, threshold optimization, and post hoc explainability through TE2Rules and SHAP. The analysis focused on realistic, locked-test evaluation to ensure generalizable findings and interpretability for clinical application.

7.2 Main Findings

- Model Performance: Model 2 demonstrated the most reliable predictive performance across the full cohort, achieving high ROC-AUC and recall despite extreme class imbalance. Model 1 showed moderate discrimination for early versus later deaths, while Model 3-B provided a viable multiclass formulation, although minority-class performance remained limited.
- Explainability: TE2Rules successfully extracted human-readable IF–THEN rules that captured clinically meaningful patterns, supported by SHAP feature importance. The rules highlighted features such as lactate levels, ED length of stay, creatinine, shock index, and abnormal vitals.
- Class Imbalance Effects: Low F1 and precision values for minority classes were explained by the rarity of events and small cohort sizes, demonstrating the need to interpret threshold-dependent metrics in context.

- Clinical Relevance: The combination of predictive accuracy and interpretability enables early identification of high-risk patients, supporting triage decisions and timely interventions.

7.3 Contributions

This dissertation makes several key contributions to the field of explainable AI in healthcare:

1. Integrated Predictive Pipeline: Combines preprocessing, feature engineering, ensemble modelling, and threshold optimization for mortality prediction.
2. Interpretability Framework: Applies TE2Rules and SHAP to generate clinically meaningful, human-readable explanations for binary and multiclass outcomes.
3. Rare-Event Modelling: Demonstrates effective modelling strategies for extremely imbalanced mortality events in emergency department populations.
4. Cascade Architecture: Introduces a multiclass cascade combining binary predictions to improve minority-class discrimination in mortality timing tasks.
5. Clinical Applicability: Provides a methodology that balances predictive performance with explainability, enabling actionable insights for healthcare providers.

7.4 Future Work

Several avenues exist for extending this research:

- External Validation: Apply the models to datasets from other hospitals to assess generalizability.
- Expanded Feature Sets: Incorporate additional physiological, laboratory, and operational features, including longitudinal time-series data.
- Advanced Explainability: Explore complementary interpretability approaches, such as counterfactual explanations or attention-based models.
- Deployment Studies: Evaluate the impact of real-time mortality prediction and TE2Rules outputs on clinical workflow, decision-making, and patient outcomes.
- Imbalance Mitigation: Investigate adaptive oversampling or cost-sensitive learning strategies to improve minority-class prediction in multiclass tasks.

The proposed future work aims to further enhance both the accuracy and interpretability of mortality prediction models, ultimately supporting safer and more effective clinical decision-making in emergency settings.

Bibliography

[1] Salimparsa, Mozhgan, et al. “Explainable AI for Clinical Decision Support Systems: Literature Review, Key Gaps, and Research Synthesis.” *Informatics*, vol. 12, no. 4, 28 Oct. 2025, p. 119, www.mdpi.com/2227-9709/12/4/119, https://doi.org/10.3390/informatics12040119.

[2] Rosenbacke, R., Melhus, Å., McKee, M. and Stuckler, D., 2024. *How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review*. JMIR AI, 3, e53207. Available at: JMIR AI

[3] Mienye, Ibomoiye Domor, et al. “A Survey of Explainable Artificial Intelligence in Healthcare: Concepts, Applications, and Challenges.” *Informatics in Medicine Unlocked*, vol. 51, 16 Oct. 2024, p. 101587, www.sciencedirect.com/science/article/pii/S2352914824001448, https://doi.org/10.1016/j.imu.2024.101587.

[4] Lapid, Nancy. “Health Rounds: AI Can Have Medical Care Biases Too, a Study Reveals.” *Reuters*, 9 Apr. 2025, www.reuters.com/business/healthcare-pharmaceuticals/health-rounds-ai-can-have-medical-care-biases-too-study-reveals-2025-04-09/.

[5] Breiman, L., 2001. “Random forests.” *Machine Learning*, vol. 45, no. 1, pp. 5–32, https://doi.org/10.1023/A:1010933404324.

[6] Friedman, J.H., 2001. “Greedy function approximation: a gradient boosting machine.” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, https://doi.org/10.1214/aos/1013203451.

[7] Naemi, A., Schmidt, T., Mansourvar, M., Wiil, U.K. and Karstoft, H., 2021. *Machine learning techniques for mortality prediction in emergency departments: a systematic review*. BMJ Open, 11(11), p.e052663. Available at: BMJ Open [Accessed 14 May 2026].

[8] Fransvea, P., Costa, G., Lepre, L., Grimaldi, S., Balducci, G. and Sganga, G., 2022. Study and validation of an explainable machine learning-based mortality prediction following emergency surgery in the elderly. Journal of Critical Care, 72, p.154126. Available at: ScienceDirect [Accessed 14 May 2026].

[9] Okada, Y., Kimura, T. and Nishimura, T., 2023. Explainable artificial intelligence in emergency medicine. Clinical and Experimental Emergency Medicine, 10(3), pp.219–228. Available at: CEEM Journal [Accessed 14 May 2026].

[10] Fındık, H., Yılmaz, A., Kaya, M. and Demir, E., 2026. Emergency Department Prediction of In-Hospital Mortality in Suspected Pulmonary Embolism: An Explainable Machine Learning Approach. Journal of Clinical Medicine, 15(4), p.1340. Available at: MDPI [Accessed 14 May 2026].

[11] Okada, Y., Kimura, T. and Nishimura, T., 2023. Explainable artificial intelligence in emergency medicine. Clinical and Experimental Emergency Medicine, 10(3), pp.219–228. Available at: CEEM Journal [Accessed 14 May 2026].

[12] Ribeiro, M.T., Singh, S. and Guestrin, C., 2016. ""Why should I trust you?": Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.1135–1144. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, pp.4765–4774.

[13] Fındık, H., Yılmaz, A., Kaya, M. and Demir, E., 2026. Emergency Department Prediction of In-Hospital Mortality in Suspected Pulmonary Embolism: An Explainable Machine Learning Approach. Journal of Clinical Medicine, 15(4), p.1340. Available at: MDPI [Accessed 14 May 2026].

[14] Sulaiman, W.A., Stylianides, C., Nikolaou, A., Antoniou, Z., Constantinou, I., Palazis, L., Vavlitou, A., Kyprianou, T., Kyriacou, E., Kakas, A., Pattichis, M.S., Panayides, A.S. and Pattichis, C.S., 2025. Leveraging machine learning and rule extraction for enhanced transparency in emergency department length of stay prediction. Frontiers in Digital Health, 6, p.1498939. Available at: Frontiers in Digital Health [Accessed 14 May 2026].

Appendix A

The complete source code developed for this dissertation, including all preprocessing pipelines, machine learning models, explainability frameworks, and the paper feature experiment, is publicly available at the following GitHub repository:
https://github.com/csavva2809/Explainable_AI_in_the_Assesment_of_Mortality