Machine Learning-Based ICU Stay Prediction

Master Thesis

submitted by

Tristan Heil

Degree Course: Industrial Engineering M.Sc.

Matriculation Number: 2511802

AIFB, IBT

KIT Department of Economics and Management

Advisor:	Prof. Dr. Harald Sack
Co-Referee:	PD Dr.-Ing. Axel Loewe
Supervisor:	M.Sc. Silvia Becker
Supervisor:	Dr. Anna Jacyszyn
Supervisor:	Dr.-Ing. Genet Asefa Gesese
Date of submission:	May 26, 2026

Assertion / Erklärung

I hereby certify on my honour that I have composed this work independently, have fully and accurately acknowledged all sources and aids used, and have clearly marked all passages taken from the work of others, whether quoted verbatim or paraphrased. I further certify that I have complied with the KIT Statutes on Safeguarding Good Scientific Practice in the version currently in force.

Ich versichere wahrheitsgemäß, die Arbeit selbstständig verfasst, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht zu haben, was aus Arbeiten anderer unverändert oder mit Abänderungen entnommen wurde, sowie die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, May 26, 2026

Tristan Heil

Acknowledgement

I would like to thank my supervisors, M.Sc. Silvia Becker, Dr. Anna Jacyszyn, and Dr.-Ing. Genet Asefa Gesese, for their guidance, constructive feedback, and the many discussions that shaped this thesis throughout the research process.

This work was supported by the state of Baden-Württemberg through bwHPC. This thesis has been prepared as a collaboration within the Leibniz Science Campus "Digital Transformation of Research" (DiTraRe) which is funded by the Leibniz Association (W74/2022).

I am deeply grateful to my family for their encouragement and support.

Abstract

Early risk assessment in the intensive care unit (ICU) is important for clinical orientation, resource planning, and the timely identification of high-risk patients. This thesis investigates whether routinely available admission electrocardiograms (ECG) can support the early prediction of length of stay (LOS) and mortality in the ICU.

Using linked critical care and ECG data from a large retrospective hospital database, an ECG-first pipeline was developed that starts from the first diagnostic 12-lead ECG recorded within the first day in intensive care and later extends this signal with demographic variables, information on the ICU, and structured data from electronic health records. The study compares a classical machine learning baseline with several deep learning architectures in a reproducible multitask setting for length of stay regression and mortality classification. The workflow covers preprocessing, augmentation, multimodal late fusion, and targeted hyperparameter optimisation under a restricted compute budget.

The results show that admission ECGs contain relevant prognostic information, but their predictive value is limited when used alone. For the prediction of LOS, the strongest models converged to a narrow performance range, while a hybrid model combining convolutional processing and long short-term memory units offered the most favourable trade-off between predictive performance, stability, and computational cost. Prediction was most informative in the clinically common range up to ten days, whereas long ICU stays remained difficult and were increasingly underpredicted. For the prediction of mortality, the models achieved moderate discrimination, yet practical classification performance depended strongly on threshold optimisation since default thresholds produced almost no useful mortality predictions. Adding structured clinical context changed the performance for LOS only slightly, but still improved mortality prediction more clearly.

Overall, the thesis provides a systematic and reproducible comparison of ECG-based and multimodal strategies for the prediction of ICU outcomes. It shows that admission ECGs are a meaningful signal source for early risk assessment, especially when combined with compact clinical context. However, it is not yet a standalone solution for clinical decision-making.

Zusammenfassung

Eine frühe Risikoeinschätzung auf der Intensivstation ist wichtig für Ressourcenplanung und die rechtzeitige Identifikation von Hochrisikopatienten. Diese Arbeit untersucht, ob routinemäßig bei Aufnahme verfügbare Elektrokardiogramme (EKG) die frühe Vorhersage der Aufenthaltsdauer auf der Intensivstation und der Sterblichkeit während der intensivmedizinischen Behandlung unterstützen können.

Auf Grundlage verknüpfter Intensiv- und EKG-Daten aus einer großen Krankenhausdatenbank wurde eine EKG-zentrierte Pipeline entwickelt, die vom ersten diagnostischen 12-Kanal-EKG ausgeht, das innerhalb des ersten Tages auf der Intensivstation aufgezeichnet wurde, und dieses Signal anschließend um demografische Variablen, Informationen zur Intensivstation und strukturierte Daten aus elektronischen Gesundheitsakten erweitert. Die Studie vergleicht einen klassischen Machine-Learning-Ansatz mit mehreren Deep-Learning-Architekturen in einem reproduzierbaren Multitask-Setting für die Regression der Aufenthaltsdauer und die Klassifikation der Sterblichkeit. Der Workflow umfasst Vorverarbeitung, Augmentierung, multimodale Late Fusion und eine gezielte Hyperparameter-Optimierung unter begrenztem Rechenbudget.

Die Ergebnisse zeigen, dass Aufnahme-EKGs relevante prognostische Informationen enthalten, ihr Vorhersagewert bei alleiniger Nutzung jedoch begrenzt ist. Für die Vorhersage der Aufenthaltsdauer konvergierten die stärksten Modelle auf einen engen Leistungsbereich. Dabei bot ein hybrides Modell aus Convolutions und Long Short-Term Memory-Einheiten den günstigsten Kompromiss aus Vorhersageleistung, Stabilität und Rechenaufwand. Die Vorhersage war im Bereich bis zu zehn Tagen am aussagekräftigsten, während lange Aufenthalte auf der Intensivstation schwierig blieben und zunehmend unterschätzt wurden. Für die Vorhersage der Sterblichkeit erreichten die Modelle eine moderate Trennschärfe, die praktische Klassifikationsleistung hing jedoch stark von der Optimierung des Schwellenwerts ab, da Standardschwellen nahezu keine brauchbaren Sterblichkeits-Vorhersagen lieferten. Die Ergänzung um strukturierten klinischen Kontext veränderte die Leistung für die Aufenthaltsdauer nur geringfügig, verbesserte die Sterblichkeitsvorhersage jedoch deutlicher.

Insgesamt liefert die Arbeit einen systematischen und reproduzierbaren Vergleich EKG-basierter und multimodaler Strategien zur Vorhersage von Sterblichkeit und Verweildauer von Patienten auf der Intensivstation. Sie zeigt, dass EKGs eine sinnvolle Signalquelle für die frühe Risikoeinschätzung sind, insbesondere in Kombination mit kompaktem klinischem Kontext, derzeit jedoch noch keine eigenständige Lösung für klinische Entscheidungen darstellen.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
APACHE	Acute Physiology and Chronic Health Evaluation
APS	Acute Physiology score
AUC	Area under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
CNS	central nervous system
DBP	Diastolic Blood Pressure
DL	Deep Learning
ECG	Electrocardiogram
EHR	Electronic Health Care Records
FN	False Negative
FP	False Positive
GCS	Glasgow Coma Scale
GRU	Gates Recurrent Units
ICU	Intensive Care Unit
ID	Identifier
LOS	Length of Stay
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAP	Mean Arterial Pressure
ML	Machine Learning
MLP	Multilayer Perceptrons
MPM	Mortality Probability Models
MSE	Mean squared Error

NN	Neural Network
ReLU	Rectified Linear Unit
ResNet	Residual Network
RNN	Recurrent Neural Network
SAPS	Simplified Acute Physiology Score
SBP	Systolic Blood Pressure
SOFA	Sequential Organ Failure Assessment
TN	True Negative
TP	True Positive

1 Introduction

This chapter introduces the clinical problem of early intensive care unit (ICU) outcome prediction. It outlines the limits of established severity scores and of models that rely mainly on structured electronic health record (EHR) data. It then motivates the use of admission-time electrocardiogram (ECG) waveforms together with demographic and structured variables. The chapter places the work in its clinical and stakeholder context, explains why joint prediction of length of stay (LOS) and mortality matters for triage and resource planning, and states the contribution, goal, and research questions of the thesis.

1.1 Problem Definition and Motivation

Accurately predicting ICU LOS is essential to keep limited beds available for patients with more acute conditions and to ensure efficient use of limited and cost-intensive ICU resources. The COVID-19 pandemic has particularly highlighted the critical role of efficient resource utilisation in intensive care medicine [1]. Careful ICU bed allocation is particularly important because a small group of patients accounts for a large share of resource use. In a population-based cohort, the 10% of patients with the highest healthcare costs after ICU admission had an approximately fourfold longer ICU LOS and fivefold higher total healthcare costs than the remaining patients [2]. Conventional scoring systems such as Acute Physiology and Chronic Health Evaluation (APACHE), Simplified Acute Physiology Score (SAPS), and Sequential Organ Failure Assessment (SOFA) are widely used to predict ICU mortality and LOS, but their use is limited by data requirements, calibration problems, and variability in routine application. For example, Polderman et al. (2001)reported substantial variability in APACHE II scores in routine practice, with systematic overestimation relative to expert reassessment [3–5]. These scores also typically rely on extensive data collection, including laboratory and physiological measurements collected at admission and during early ICU care [5, 6].

These limitations create a need for more individualised and data-driven prediction methods. Such methods should better reflect the heterogeneity of patient populations and the fact that a patient's condition evolves during the ICU stay. Subsequent machine learning (ML) models for ICU outcome prediction have largely been built on structured EHR data, while dynamic, high-frequency signals are often only summarised or not explicitly modelled [7, 8]. Many ML models in the ICU act as "black boxes" with limited transparency regarding the characteristics that drive their predictions, and concerns about interpretability, generalisability, and workflow integration have so far limited their adoption in everyday clinical practice, despite strong predictive performance [9, 10]. At the same time, continuous physiological signals like ECG are routinely available at the bed-

side [11]. They provide detailed information on heart rate, rhythm, and morphological abnormalities.

In predictive modelling, high-frequency waveforms such as ECG have more recently gained attention in acute and critical care, compared with earlier approaches based primarily on static EHR variables [12, 13]. As a continuous signal reflecting cardiac activity, the ECG holds great potential for diagnostics and prognosis by capturing subtle, time-dependent patterns that are difficult to detect visually and often precede overt clinical deterioration [12, 13]. Leveraging this information in predictive models may enable more accurate LOS and mortality predictions and improve early risk stratification by identifying patients at elevated risk of deterioration before overt clinical decline. Such predictions could in turn support optimised resource allocation, staff planning, and more structured treatment pathways, as well as earlier escalation of care or more intensive monitoring for those who need it most.

1.2 Contribution and Goal of this Thesis

This thesis has the overarching goal of developing and evaluating a ML framework for early prediction of ICU LOS and mortality based on data available at ICU admission. The core idea is to start with raw ECG waveforms as the primary data source and to incrementally integrate additional modalities such as demographic and structured EHR variables. The evaluation focuses on prediction performance and interpretability to ensure that the developed models are not only statistically sound but also clinically meaningful.

The work was prepared as part of an interdisciplinary collaboration between the Computational Cardiac Modeling group at KIT IBT and the Information Service Engineering group at FIZ Karlsruhe, within the framework of the Leibniz ScienceCampus Digital Transformation of Research (DiTraRe).

The scientific contribution of this work can be summarised as follows. First, it proposes an ECG-first modelling strategy for joint prediction of ICU LOS and mortality and implements a stepwise multimodal pipeline that quantifies the added value of demographic and structured EHR data over ECG-only models. Second, it systematically compares different deep learning (DL) architectures for ECG-based ICU outcome prediction with respect to discrimination, robustness, and calibration. Third, it investigates the interpretability and clinical plausibility of the learned representations and discusses how such models could be embedded into ICU workflows, for example for triage support, resource planning, or early warning systems.

Research Questions

1. What level of discrimination, calibration, and robustness do DL models achieve when predicting ICU LOS and mortality from the first raw ECG recorded at ICU admission?

2. To what extent does integrating ECG with structured EHR and demographic data improve predictive performance compared to ECG-only models?

3. Which ML architectures are most suitable for modelling ECG-based ICU outcome prediction in terms of accuracy and robustness?

1.3 Clinical and Stakeholder Context

Clinical background and prognostic gap

ICUs treat critically ill patients who require close monitoring, organ support, and rapid response to deterioration [1, 14]. In this setting, ICU mortality, in-hospital mortality, and LOS are important because they reflect illness severity, quality of care, and resource use. Prolonged LOS reduces bed availability, increases costs, and is associated with complications and higher mortality [2, 15]. Reliable early prediction of LOS and mortality can therefore support triage, discharge planning, and timely escalation or de-escalation of therapy.

At ICU admission, clinicians already use patient history, vital signs, laboratory values, and scores such as APACHE, SAPS, and SOFA [5, 6, 16]. Continuous monitoring also provides physiological signals such as ECG [11, 13]. However, the ECG is still used mainly for diagnosis and rhythm surveillance rather than for systematic prediction of ICU LOS or mortality [17, 18]. This gap motivates approaches that combine admission ECGs with basic demographic and clinical information to derive early risk estimates [19, 20]. Such estimates could support triage, resource planning, and early identification of patients who may need closer monitoring or more aggressive therapy [21, 22].

Stakeholders and implementation context

Any predictive system for ICU care would operate in a socio-technical environment with several stakeholders. Intensivists, fellows, and nursing staff are the primary users and need outputs that are understandable, clinically plausible, and compatible with routine workflows. Important concerns include black-box behaviour, documentation burden, changing professional roles, and alarm fatigue [9, 10, 23]. Patients and carers are affected indirectly and may worry that artificial intelligence (AI) could replace human clinical judgement. A human-in-the-loop approach is therefore important: AI-supported risk estimates should

inform, not substitute, individual medical decision-making. Used under clinical supervision, they may also improve prognostic communication [23, 24].

Developers, clinical researchers, hospital managers, regulators, and IT governance bodies focus more on interoperability, heterogeneous data quality, implementation effort, organisational impact, responsibility, data protection, bias, and continuous performance monitoring [10, 14, 23]. Early LOS prediction is only useful if the clinical value justifies implementation and training costs [2, 25].

Clinical success criteria

From a clinical perspective, such a model is successful if it provides accurate, early, and actionable risk estimates for ICU mortality and LOS at or shortly after admission. It should be robust across relevant patient groups, tolerate moderate data-quality variation, and add value for triage, treatment decisions, and transfer or discharge planning. Beyond predictive performance, it should rely on routinely available ECG and clinical data, integrate into existing ICU dashboards or EHR systems, and provide outputs that are understandable for ICU staff.

1.4 Project Constraints

This thesis is based on a single retrospective ICU dataset, MIMIC-IV ECG, since it provides a rare combination of raw ECG waveforms, linked ICU stays, and complementary structured clinical variables in a transparent and reproducible research setting [26]. Using one dataset also keeps cohort construction, preprocessing, and label definition consistent across all experiments, which is appropriate for a single-author thesis, although it necessarily limits claims about generalisability. The study is conducted in a multimodal setting and addresses a dual prediction task (LOS and mortality). Detailed information on the cohort, feature definitions, missing data, and preprocessing is provided in Chapter 4. The clinical requirements are not externally mandated within this project but are derived from the stakeholder and implementation considerations outlined in Section 1.3. Accordingly, the models should produce interpretable outputs and remain conceptually plausible for ICU workflows, even though actual deployment is outside the scope of this thesis. Here, “time of admission” refers to the time at which a patient enters the ICU, and all predictions are restricted to information available at or immediately after that admission event in order to support early risk stratification rather than later in-stay reassessment. Methodologically, the focus lies on comparing several ML model families instead of heavily fine-tuning a single state-of-the-art model, given that the thesis aims to assess modality-specific trade-offs, robustness, and reproducibility under realistic project constraints.

2 Fundamentals

This chapter introduces the theoretical foundations that are required for the subsequent analysis. It first outlines core concepts of ML, then summarises the mathematical principles and common architectures of DL, and finally reviews the relevant medical background on cardiac physiology and electrocardiography. Together, these foundations provide the conceptual basis for the methods and experiments presented in the following chapters.

2.1 Fundamentals of Machine Learning

ML is a subfield of AI. It focuses on algorithms that learn predictive or descriptive patterns from data rather than relying on explicitly programmed rules [27–29]. Formally, an ML system is given a set of observed examples, called the training data, with inputs x and associated targets y. From these examples, it learns a mapping $f : x \mapsto y$ that captures the relationship between inputs and outputs. The central objective is generalisation: accurate predictions on new, unseen samples rather than mere reproduction of the training data [27, 28]. Two failure modes limit generalisation. *Overfitting* occurs when a model memorises characteristics of the training set instead of learning the underlying structure, resulting in high training accuracy but poor test performance [28]. *Underfitting* occurs when a model is too simple to capture relevant patterns, leading to poor performance on both training and test data [28, 29]. To evaluate generalisation, the available data are partitioned into three disjoint subsets. The training set is used to fit model parameters. The validation set serves for model selection and hyperparameter tuning without biasing the final evaluation. The test set is accessed only after a model configuration has been fixed. It provides an unbiased estimate of performance on unseen data [28, 30, 31].

2.1.1 Mathematical Foundations

In supervised learning, a model with parameters θ produces predictions $\hat{y}_i = f(x_i; \theta)$ for labelled training pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$. Learning aims to reduce the discrepancy between predictions and observed targets through a task-dependent loss function [28, 30]. For regression with continuous targets, the mean squared error (MSE) is widely used:

$$\mathcal{L}_{\mathrm{MSE}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.1}$$

[30]. For classification, cross-entropy loss is the standard choice. For binary classification with $y \in \{0, 1\}$ and predicted probability $\hat{p} = \sigma(z)$, the binary cross-entropy is

$$\mathcal{L}_{\mathrm{BCE}}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \big] \tag{2.2}$$

[25, 30]. For multiclass classification with C classes, the categorical cross-entropy extends this to one-hot targets and a softmax output [25, 30]. Parameters are typically optimised via gradient-based methods. Gradient descent updates parameters by following the gradient of the loss function. In practice, stochastic variants that operate on mini-batches of data are widely used to reduce computational cost [30, 32, 33]. The trade-off between model complexity and fit is formalised by the bias–variance decomposition: overly simple models suffer from high bias (underfitting), while overly complex models suffer from high variance (overfitting) [28].

2.1.2 Learning Paradigms

ML methods are distinguished by the type of feedback available during training [27, 28]. In **supervised learning**, the model is trained on labelled examples where each input x is paired with a target y. Classification assigns discrete labels, while regression predicts continuous values. This paradigm applies to tasks such as predicting clinical outcomes from patient data [28, 30]. **Unsupervised learning**, by contrast, operates on unlabelled data and seeks to discover latent structure, for example through clustering, density estimation, or dimensionality reduction [27, 29]. Applications include patient subgrouping and anomaly detection. This thesis formulates the prediction of ICU outcomes from ECG and EHR data as a supervised learning problem. Subsequent sections therefore focus on supervised methods and architectures.

2.1.3 Data Representation and Preprocessing

ML models operate on diverse data modalities that require distinct representations. Tabular data consist of fixed feature sets per instance, organised as a sample-by-feature matrix. Examples include demographic variables and laboratory measurements [27, 28]. Sequential data are ordered observations where temporal structure is informative. Physiological signals such as ECG recordings fall into this category and therefore require representations that preserve temporal order [27, 31]. Images and text form additional modalities with spatial and linguistic structure, respectively. Preprocessing improves data quality and ensures comparability across samples. Common steps include normalisation, handling of missing values, and signal-specific operations such as resampling, filtering, and artifact removal [30, 31]. A fundamental distinction exists between feature engineering and representation learning. In feature engineering, domain experts manually design informative variables from raw data. In representation learning, the model learns intermediate representations directly from raw inputs, reducing reliance on manual feature design [34, 35]. The specific data modalities and preprocessing steps used in this thesis are described in section 4.

2.2 Fundamentals of Neural Networks and Deep Learning

Feedforward neural networks (NNs) approximate an unknown target function $f(x)$ by combining multiple layers of linear transformations with nonlinear activations [34–36]. Section 2.2.1 introduces the mathematical foundations. Section 2.2.2 then surveys common architectures.

2.2.1 Mathematical Foundations

The basic building block of a feedforward NN is a layer that combines a linear transformation with a nonlinear activation. For an input vector $\mathbf{x} \in \mathbb{R}^d$, a single layer first computes the pre-activation vector

$$\mathbf{a} = \mathbf{W}\mathbf{x} + \mathbf{b} \tag{2.3}$$

and then applies an activation function element-wise to obtain $\mathbf{z} = \sigma(\mathbf{a})$. Here, $\mathbf{W} \in \mathbb{R}^{m \times d}$ denotes the weight matrix, $\mathbf{b} \in \mathbb{R}^m$ the bias vector, and σ an element-wise activation function [32, 37]. This layer-wise pattern is stacked repeatedly to build a deep network. Nonlinear activations are essential because without them, a composition of multiple layers would still collapse to a single linear mapping [34, 37].

Different activations are used for different roles in the network. Rectified linear units (ReLU) are a common default in hidden layers. Sigmoid functions are often used for binary outputs or gating mechanisms. Softmax is used at the output layer for multiclass prediction, where $\text{softmax}(\mathbf{a})_i = e^{a_i} / \sum_{j=1}^{C} e^{a_j}$ [25, 37–39]. The output activation therefore determines how the final layer is interpreted, for example as a probability distribution over classes.

As discussed in section 2.1.1, learning is cast as the minimisation of a task-dependent loss. Typical choices are mean squared error (MSE) for regression and cross-entropy for classification. For multiclass classification, the standard sample-wise cross-entropy loss is

$$\ell_{\text{CE}}(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log \hat{y}_c, \tag{2.4}$$

while binary classification commonly combines sigmoid outputs with binary cross-entropy [25, 30]. In both cases, the loss links the network output to the learning objective by quantifying the discrepancy between the prediction $\hat{y}$ and the target y.

Training can be understood as the supervised learning cycle illustrated in Figure 1. The figure explicitly presents that learning proceeds from left to right during forward propagation and from right to left during the backward and update phase.

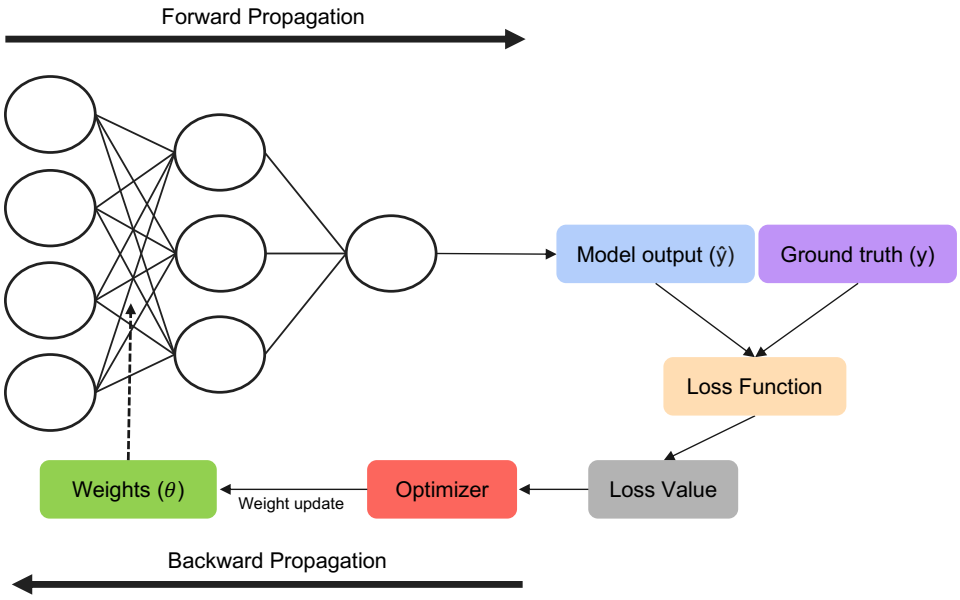

Figure 1: Conceptual supervised training step for a feedforward network.

Starting from an input, the network applies its current parameters θ to produce a model output $\hat{y}$. The loss function then compares $\hat{y}$ with the ground-truth target y and returns a scalar loss value that measures the prediction error. Backpropagation uses the chain rule to compute gradients of this loss with respect to θ, and an optimiser updates the parameters accordingly [32, 40]. In practice, deep networks are commonly trained with stochastic gradient-based methods, often using adaptive optimisers such as Adam [33, 35, 41]. The focus of this thesis is on comparing model architectures and input modalities rather than on tuning one architecture to its fullest possible extent. Therefore, the relevant abstraction is the shared forward-loss-backward training loop. Although the schematic uses a compact feedforward example for clarity, the same logic carries over to the convolutional, residual, and recurrent models introduced below. The discussion above describes the generic training cycle of deep NNs.

To understand DL more broadly, three additional ideas are important. First, deep models learn hierarchical representations by stacking many parameterised transformations, so that early layers capture simple patterns while deeper layers combine them into more abstract features [34, 35, 42]. Second, these transformations need not always be fully connected. Depending on the architecture, the linear part of a layer can exploit structure such as sparsity, locality, or parameter sharing. One important example, discussed in more detail in section 2.2.2, is the one-dimensional (1D) convolution. A 1D convolutional layer applies a learnable kernel k of length L to the input sequence x. The operation is typically implemented as cross-correlation [22, 32]:

$$(x \star k)_t = \sum_{s=0}^{L-1} x_{t+s} \cdot k_s \tag{2.5}$$

This example illustrates how architectural design can encode assumptions about the structure of the input while keeping the number of trainable parameters manageable. Third, increasing depth and expressive power also introduces two practical challenges: optimisation becomes more difficult, and the risk of overfitting increases. As networks become deeper, gradients may attenuate across many stacked transformations. One common response is to introduce skip connections, as in residual learning, where the target mapping is written as $H(x) = \mathcal{F}(x) + x$, equivalently $\mathcal{F}(x) = H(x) - x$ [37, 39]. This formulation facilitates gradient flow and enables the training of substantially deeper models [37, 39].

At the same time, the increased representational capacity of deep models must be controlled to avoid overfitting. Regularisation therefore complements architectural design by constraining effective model complexity through appropriate regularisation strategies [31, 37, 40]. Further optimisation and regularisation methods are discussed in section 2.2.3.

2.2.2 Deep Learning Architectures

DL refers to the use of deep NN with many stacked layers that learn hierarchical representations directly from raw data, thereby reducing the need for manual feature design [34, 35, 42]. Through successive transformations, such models can automatically extract low-, mid-, and high-level features that are useful for a wide range of prediction tasks.

Multilayer Perceptrons (MLP) consist of fully connected layers in which each unit in one layer is connected to all units in the next layer. They are particularly suited for structured or tabular data and can capture complex nonlinear relationships between features, but tend to require many parameters when applied to very high-dimensional inputs [30].

Convolutional Neural Networks (CNN) use convolutional filters with local receptive fields, weight sharing, and often pooling operations to aggregate local information. Although originally developed for image data, the same principles apply to one-dimensional inputs: 1D convolutions are widely used for time series and physiological signals such as ECG, where CNNs learn local patterns and increasingly abstract hierarchical features across layers [31, 36, 37, 43].

Residual Networks (ResNet) extend CNNs with skip connections that allow gradients to flow directly through the network, mitigating the vanishing gradient problem and enabling training of very deep architectures. Residual blocks learn incremental changes (residuals) rather than full transformations [37, 39].

Recurrent Neural Networks (RNN) model sequences by recursively updating a hidden state that is propagated over time and summarises past inputs. Gated variants such as **Long Short-term Memory** (LSTM) networks and **Gated Recurrent Units** (GRU) are specifically designed to better capture long-range dependencies and are widely used in time series and ICU outcome prediction [44–46].

Transformer architectures replace explicit recurrence with self-attention mechanisms that allow each position in a sequence to attend to all other positions [47, 48]. This design scales well to long sequences and parallel computation. It has become the standard in natural language processing, with rapidly growing use in time series, signal processing, and ECG analysis [48, 49].

Hybrid and multimodal architectures combine different building blocks to exploit complementary strengths, for example CNNs for local feature extraction together with RNNs or transformers for temporal or contextual modelling [22, 50, 51]. When combining different data types (e.g., ECG waveforms and tabular EHR data), each modality is processed by its own encoder network. The resulting feature vectors are then combined in a shared output layer that produces the final prediction [7, 12].

2.2.3 Regularisation and Optimisation in Deep Learning

Regularisation and optimisation techniques in DL aim to improve generalisation to unseen data while ensuring that training remains efficient and numerically stable. Modern neural networks rely on a combination of architectural choices, explicit regularisers, normalisation layers, and optimisation algorithms, all of which are controlled by hyperparameters that must be tuned carefully [14, 31].

The behaviour of DL models is influenced by hyperparameters, although their impact is not uniform. In many practical settings, tuning optimisation and regularisation hyperparameters such as learning rate, batch size, weight decay, or dropout rate improves performance mainly incrementally rather than fundamentally. By contrast, architectural hyperparameters such as network depth, width, or kernel size can have a larger effect because they determine model capacity and representational structure. Typical search strategies comprise grid search and random search, with Bayesian optimisation as a more sample-efficient alternative. In all cases, a dedicated validation set is used to compare configurations and select the final model [31, 38]. Regularisation methods constrain model complexity and reduce overfitting. Common approaches include weight decay or L2 regularisation, which penalises large weight norms, dropout, which randomly deactivates a subset of neurons during training, and early stopping, which halts training once performance on a validation set deteriorates [35, 37, 40].

Data augmentation serves as an additional, data-centric regulariser by synthetically in-

creasing the diversity of the training set, for example through segment-based sampling, synthetic oversampling, or signal-specific perturbations [31, 37, 51]. Normalisation layers further stabilise and accelerate training. Batch normalisation and layer normalisation rescale and shift intermediate activations to maintain more consistent distributions across layers, which permits the use of larger learning rates and often improves convergence behaviour [37, 40, 48]. These effects can translate into both faster training and improved generalisation, though the precise impact depends on architecture and training regime. Optimisation algorithms determine how network parameters are updated based on gradients of the loss function. Stochastic gradient descent (SGD) with momentum remains a widely used baseline, combining noisy gradient estimates with an exponentially decayed moving average to accelerate progress in relevant directions. Adaptive methods such as Adam, AdamW, or RMSProp adjust effective learning rates per parameter based on historical gradient information, which can speed up convergence and simplify tuning, even sometimes at the cost of reduced generalisation compared to well-tuned SGD [31, 33, 35, 41].

2.3 Medical Fundamentals

In this subsection I introduce the medical foundations that are required for the subsequent analysis of ECG data and clinically relevant prediction tasks. It therefore summarises the anatomical and physiological properties of the atria, outlines the cardiac excitation system, and explains the basic principles of the electrocardiogram as the central diagnostic signal considered in this thesis.

2.3.1 Cardiac Anatomy and Basic Function

The human heart is a four-chambered muscular pump that maintains blood flow through two connected circuits: the right heart drives the pulmonary circulation through the lungs, while the left heart drives the systemic circulation through the rest of the body [52, 53]. Anatomically, the right atrium receives deoxygenated blood from the superior and inferior vena cava and passes it through the tricuspid valve into the right ventricle, which ejects blood into the pulmonary artery. Oxygenated blood then returns from the lungs via the pulmonary veins to the left atrium, passes through the mitral valve into the left ventricle, and is pumped into the aorta. The right and left sides of the heart are separated by septa, while the valves ensure predominantly unidirectional flow. This structural organisation forms the basis for the coordinated electrical activation described in the next section and for the surface ECG, which reflects the propagation of excitation through atrial and ventricular myocardium [52–54].

Figure 2 summarises this macroscopic organisation by showing the four chambers, the major valves, and the main vessels that connect the pulmonary and systemic circulation.

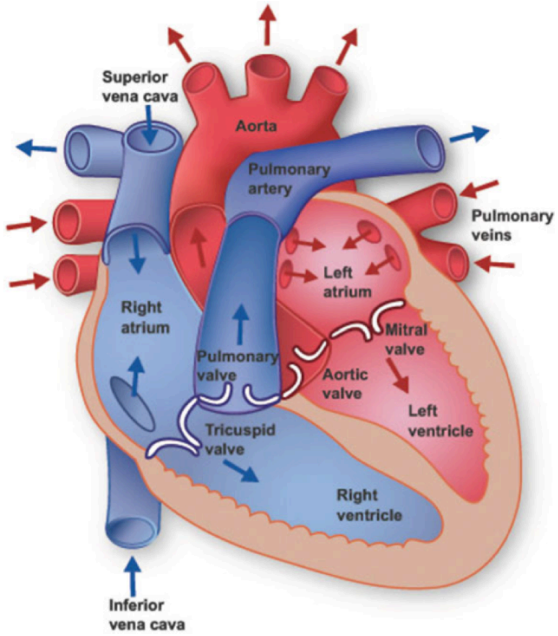

Figure 2: Schematic of the human heart with the main blood-flow directions. Adapted from [55].

2.3.2 The Electrical Excitation System

To pump blood effectively, the heart must contract in a regular, coordinated way. These contractions are triggered by electrical signals. Specialised cardiac muscle cells generate impulses and pass them on so that atria and ventricles activate in the right order and at the right time [56].

At the organ level, the electrical impulse normally starts in the *sinoatrial node* (SA node) in the right atrium. This structure acts as the heart's natural pacemaker and sets the basic rhythm at roughly 60 to 80 beats per minute under resting conditions. Cardiomyocytes are connected through intercalated discs that contain gap junctions, allowing electrical current to spread rapidly from cell to cell and thereby enabling coordinated, nearly synchronous activation of the myocardium [18, 56]. From there, excitation first spreads through the atrial muscle so that the atria contract and push blood into the ventricles. Interatrial conduction pathways help synchronise activation of the left atrium with the right atrium, with Bachmann's bundle representing the best known preferential route for right-to-left atrial conduction [57]. The signal then reaches the *atrioventricular node* (AV node), where conduction briefly slows down. This short delay is physiologically important because it gives the ventricles time to fill before they contract. Afterwards, excitation is transmitted

through the *His bundle*, the left and right bundle branches, and finally the *Purkinje network*, which rapidly distributes the impulse throughout the ventricular myocardium. In this way, activation follows an ordered sequence from atria to ventricles, enabling efficient and synchronised pumping of blood.

At the cellular level, each myocyte maintains a stable resting voltage across its membrane. When neighbouring tissue depolarises, local ion channels open and trigger the cardiac action potential, i.e. the characteristic time course of membrane voltage change in an excitable heart cell. This action potential consists of a rapid depolarisation phase, a plateau phase, and repolarisation back towards the resting state. The plateau phase is particularly important in cardiac muscle because calcium influx through voltage-dependent channels helps couple electrical excitation to subsequent force generation and contraction. After repolarisation, the cell enters a refractory period. During this time it cannot be re-excited immediately, which helps maintain orderly propagation and reduces the risk of self-sustaining re-entry circuits [18, 56].

2.3.3 The Electrocardiogram

The ECG records cardiac electrical activity at the body surface. In clinical routine, the 12-lead ECG is a standard non-invasive tool for rhythm analysis and for detecting abnormalities such as ischemia and conduction disturbances [11, 18]. In hospital datasets, it is also commonly available at admission and therefore highly relevant for data-driven risk modelling [26].

From a physiological perspective, the ECG reflects the propagation of depolarisation and repolarisation waves through the myocardium. From a coarse conceptual mathematical perspective, the electrical activity of the heart can be approximated by a time-varying equivalent dipole or heart vector $\mathbf{p}(t)$, while the thorax acts as a volume conductor. The signal measured by lead k can then be interpreted as a projection of this cardiac vector onto the corresponding lead vector $\mathbf{l}_k$, i.e.,

$$s_k(t) \propto \mathbf{l}_k^{\mathsf{T}} \mathbf{p}(t). \tag{2.6}$$

This relation is a simplified forward-model approximation. Each lead does not measure the entire cardiac source directly, but a lead-specific projection of it. Consequently, waveform amplitude and polarity depend on the orientation of the instantaneous electrical vector relative to the lead axis [18, 53, 55].

In practice, ECG leads are implemented as potential differences between defined electrode configurations. The ECG therefore traces the temporal course of these lead-specific voltages, typically displayed in millivolts over time [18, 53]. A deflection is positive or negative depending on whether the instantaneous cardiac electrical vector points towards

or away from the positive pole of that lead. Its amplitude grows when the vector is more aligned with the lead axis and shrinks when the vector is nearly perpendicular to it [18, 53].

A single heartbeat is characterised by a sequence of deflections. The **P-wave** reflects atrial depolarisation, its duration indicates the time required for the electrical impulse to propagate through the atria. Atrial repolarisation is not visible on the surface ECG because it is masked by the simultaneous depolarisation of the ventricles. The **QRS complex** represents ventricular depolarisation, and its duration reflects the time required for ventricular activation. The **T-wave** reflects ventricular repolarisation. its duration and shape reflect the time course of electrical recovery in the ventricles. The timing, amplitude, and polarity of these waves vary with lead position and between individuals, but their sequence remains consistent [18, 53]. Key ECG features include the PR interval (from P-wave onset to QRS onset, reflecting atrioventricular conduction), the QRS duration, the ST segment (from the end of the QRS complex to the onset of the T-wave), and the QT interval (from the onset of the QRS complex to the end of the T-wave, spanning ventricular depolarisation and repolarisation). In addition, the RR interval between successive beats is commonly used to estimate heart rate [18, 53]. Figure 3 illustrates these deflections and intervals in a schematic single-beat trace.

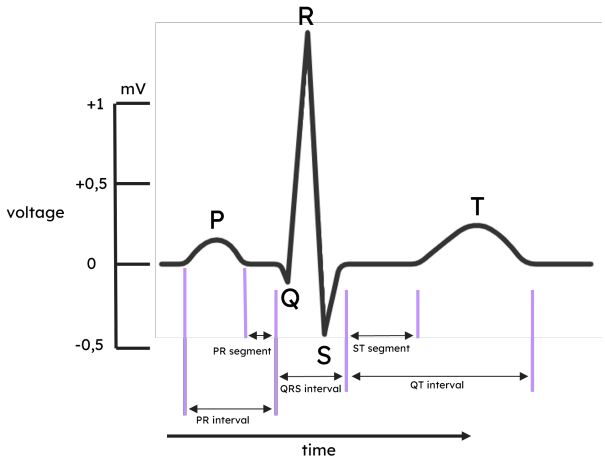

Figure 3: Schematic single-beat ECG with the main waves and intervals. Recreated based on [53].

The standard 12-lead ECG uses ten electrodes to derive twelve distinct views of the heart's electrical activity [18, 53]. The leads fall into two groups: six limb leads, which capture electrical activity in the frontal plane, and six precordial (chest) leads, which capture activity in the horizontal plane.

The three **bipolar limb leads** (I, II, III) were introduced by Einthoven and are derived from the potential differences between pairs of limb electrodes. Denoting the potentials

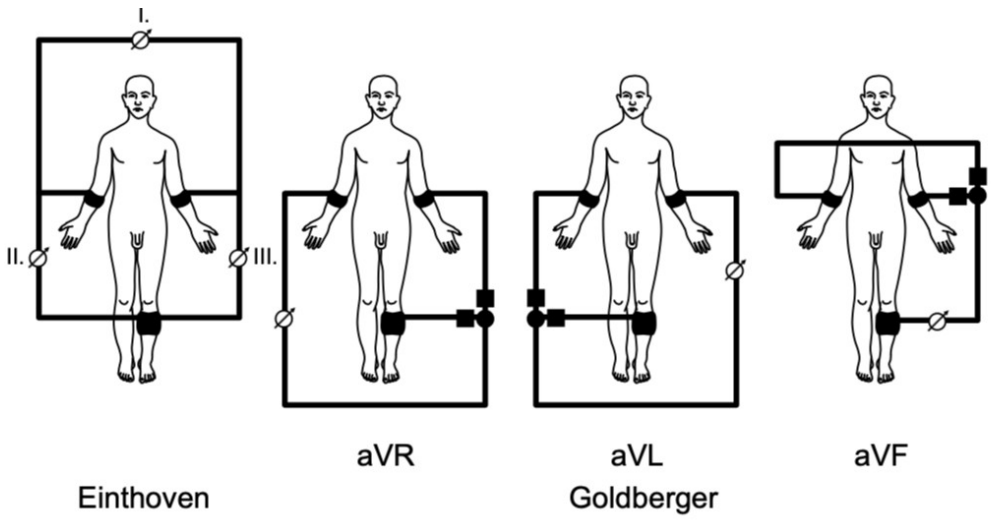

Figure 4: Schematic of the bipolar and augmented limb leads with electrode placement and lead configuration [58].

at the left arm, right arm, and left foot by ϕ_{L}, ϕ_{R}, and ϕ_{F}, respectively, the three leads are defined as

$$\mathrm{I} = \phi_{\mathrm{L}} - \phi_{\mathrm{R}}, \quad \mathrm{II} = \phi_{\mathrm{F}} - \phi_{\mathrm{R}}, \quad \mathrm{III} = \phi_{\mathrm{F}} - \phi_{\mathrm{L}}. \tag{2.7}$$

These leads form Einthoven's triangle and provide views of the heart's electrical axis in the frontal plane [18, 53]. Because these leads are linear combinations of the same limb potentials, they also satisfy Einthoven's law, $\mathrm{II} = \mathrm{I} + \mathrm{III}$, which illustrates that the standard limb leads are not independent measurements [53]. The three **augmented limb leads** (aVR, aVL, aVF), proposed by Goldberger, measure the potential at one limb electrode relative to the average of the other two, thereby providing additional frontal-plane views:

$$\mathrm{aVR} = \phi_{\mathrm{R}} - \frac{\phi_{\mathrm{L}} + \phi_{\mathrm{F}}}{2}, \quad \mathrm{aVL} = \phi_{\mathrm{L}} - \frac{\phi_{\mathrm{R}} + \phi_{\mathrm{F}}}{2}, \quad \mathrm{aVF} = \phi_{\mathrm{F}} - \frac{\phi_{\mathrm{L}} + \phi_{\mathrm{R}}}{2}. \tag{2.8}$$

The six **precordial leads** (V1–V6) are placed on the chest wall. V1 and V2 are positioned in the fourth intercostal space adjacent to the sternum and are oriented towards the right ventricle and interventricular septum. V3 and V4 face the anterior interventricular septum. V4 is typically placed at the midclavicular line in the fifth intercostal space. V5 and V6 extend along the fifth intercostal space towards the left lateral chest and capture electrical activity from the left ventricle and lateral wall [11, 53]. Each precordial lead records the potential difference between the respective chest electrode and the Wilson central terminal $\phi_{\mathrm{WCT}} = (\phi_{\mathrm{R}} + \phi_{\mathrm{L}} + \phi_{\mathrm{F}})/3$, a virtual reference derived from the limb electrodes. Accordingly, the precordial leads are given by

$$V_i = \phi_{V_i} - \phi_{\mathrm{WCT}}, \qquad i \in \{1, \ldots, 6\}. \tag{2.9}$$

Here, ϕ_{V_i} denotes the potential measured at the respective chest electrode [53].

Standard recording settings specify a paper or display sweep speed of 25 mm/s and an amplification of 10 mm/mV, enabling consistent interpretation and comparison across recordings [18, 53]. The 12-lead ECG thus provides a comprehensive, standardised view of cardiac electrical activity from multiple angles. It forms the foundation for both clinical interpretation and automated analysis by ML models [31, 35].

2.3.4 The SOFA score

SOFA is a classical ICU organ-dysfunction score. It was introduced to describe the degree of organ dysfunction or failure over time, rather than to serve as a single admission mortality model [59]. This distinguishes it from admission-oriented severity scores such as APACHE or SAPS, which are mainly designed for early mortality-risk estimation and benchmarking [60].

The score consists of six organ-system components: respiration, coagulation, liver function, cardiovascular function, central nervous system function, and renal function. Each component is assigned an integer value from 0 to 4, where 0 denotes no relevant dysfunction according to the SOFA criteria and 4 denotes the most severe abnormality. The total score is the sum of the six components:

$$\mathrm{SOFA} = S_{\mathrm{resp}} + S_{\mathrm{coag}} + S_{\mathrm{liver}} + S_{\mathrm{cardio}} + S_{\mathrm{cns}} + S_{\mathrm{renal}}, \qquad \mathrm{SOFA} \in [0, 24]. \tag{2.10}$$

Here, S_{resp}, S_{coag}, S_{liver}, S_{cardio}, S_{cns}, and S_{renal} denote the respiratory, coagulation, liver, cardiovascular, central nervous system (CNS) , and renal component scores, respectively. The individual components are based on the arterial oxygen partial pressure to inspired oxygen fraction ratio (PaO_2/FiO_2) for respiration, platelet count for coagulation, bilirubin for liver function, mean arterial pressure (MAP) and vasopressor or inotrope support for cardiovascular function, the Glasgow Coma Scale (GCS) for CNS function, and creatinine or urine output for renal function [59]. In serial ICU use, SOFA is usually calculated per day by assigning each organ system the worst score reached during the assessment window and then summing the component scores. Higher values indicate greater overall organ dysfunction.

3 State of the Art

This chapter reviews prior approaches to ICU outcome prediction for both mortality and LOS and structures the literature along the methodological evolution from classical clinical scoring systems to ML on structured EHR data and, finally, to models operating on temporally resolved time-series data. The aim is to highlight both progress in predictive modelling and persisting limitations: many established approaches still emphasise information available around ICU admission or low-frequency summaries, rather than exploiting the full temporal detail of how a patient's condition evolves during the ICU stay.

3.1 Classical Scoring Systems

Classical ICU scoring systems form the historical backbone of outcome prediction for both mortality and LOS and thus provide the conceptual starting point for ML-based ICU stay prediction [60]. Critical care reviews commonly distinguish between two broad categories of scoring systems [16, 60]. The first comprises admission-oriented severity scores such as the APACHE family, SAPS, and Mortality Probability Models (MPM), which estimate in-hospital mortality risk shortly after ICU admission. The second comprises organ-dysfunction scores, most prominently the SOFA score, which quantify the extent and progression of organ failure during the ICU stay.

Within the APACHE family, APACHE II and APACHE IV are among the most widely used variants. They combine acute physiological abnormalities (worst values of selected vital signs and laboratory measures) with age and chronic health status, and APACHE-based assessment relies on physiological variables collected during the first 24 hours after ICU admission [16, 60].

SAPS II likewise uses an early 24-hour assessment window, whereas SAPS 3 was explicitly designed to use data available at the time of ICU admission [6, 16]. By contrast, SOFA aggregates six organ-system components and is intended for serial reassessment, making it better suited to tracking evolving organ dysfunction over time than to providing a single admission snapshot [59]. In addition, APACHE IV provides a predicted ICU LOS and has been used for severity-adjusted benchmarking of observed versus expected LOS [61].

These systems support bedside risk stratification, severity-adjusted benchmarking of outcomes such as mortality and LOS for hospital and ICU quality assurance, and standardised reporting in clinical studies when applied to populations and settings close to those used for development and validation [60, 61]. Their limitations are structural: scores condense heterogeneous physiology into small fixed sets of variables with *fixed* weights and discrete cut-offs, require time-consuming manual extraction of values from clinical records,

and can show substantial inter-observer variability in real-world APACHE scoring [3, 62]. Calibration and discrimination therefore shift across cohorts, countries, and ICU types. Moreover, they operate on sparse time points rather than on the high-resolution temporal trajectories increasingly available in electronic health records and bedside monitoring [62, 63]. Moreover, conventional score definitions are built around tabular physiology, laboratory values, and diagnostic information. They do not incorporate waveform-centred modalities such as the 12-lead ECG. Modelling outcomes from raw ECG therefore requires a different data and modelling pipeline than classical scoring, which motivates the ECG-centred approach developed later in this thesis.

3.2 Machine Learning on Static EHR Data

Building on classical scoring systems, many ICU prediction studies have applied ML to structured EHR variables rather than to raw waveforms or continuously sampled monitoring data [27, 64]. The corresponding feature sets are typically tabular and comprise demographic characteristics, comorbidities, laboratory values, vital signs, and early treatment-related variables available at admission or summarised over an early observation window. A systematic review by [27] showed that mortality prediction was the dominant use case in this literature, with commonly used models including logistic regression, random forests, support vector machines, gradient-boosting methods, and NNs. Typical prediction targets include ICU or hospital mortality as well as binary or categorised LOS outcomes.

For example, [65] developed models for prolonged ICU stay using structured variables and reported that all tested ML models outperformed a customised SAPS II baseline, with gradient boosting decision trees performing best in both internal and external validation. Similarly, [66] found that logistic regression, LightGBM, and multilayer perceptron models outperformed SAPS II for predicting in-ICU mortality in patients with severe pneumonia. In a separate single-centre study based on admission variables, [8] reported high discriminative performance for random-forest models in ICU mortality as well as short and long ICU stay categories.

At the same time, these results should be interpreted cautiously. The cited studies differ substantially in patient population, predictor definition, outcome formulation, and validation strategy, which limits direct comparison of reported performance gains. Moreover, although such models improve flexibility relative to hand-crafted severity scores, they still rely on static or early aggregated inputs and therefore do not represent the continuous, time-evolving trajectory of critical illness [62]. This limitation motivates the use of ML on time-series data. In the ICU, patient state changes over time, and dynamic models can use repeated measurements rather than only admission snapshots. For example, Baker et al. used variation in vital signs over a 24-hour window for automatic mortality risk

prediction [22]. Recent work on clinical time-series benchmarks and dynamic prediction models has therefore shifted the focus from admission snapshots to temporally resolved EHR and monitoring data, enabling repeated risk estimation throughout the ICU stay [30, 62, 63].

3.3 Machine Learning on Time-Series Data

Time-series modelling has become a central direction in ICU outcome prediction because critical illness is inherently dynamic. Physiological instability, treatment response, and organ dysfunction evolve over hours and days, so that clinically relevant information is often contained not only in absolute values but also in temporal patterns such as trends, variability, and abrupt deviations [14, 62]. Public benchmarks based on MIMIC-III accelerated this line of work by defining reproducible tasks for in-hospital mortality, LOS, physiologic decompensation, and phenotype classification from multivariate clinical time series [30]. In parallel, dynamic prediction studies using high-frequency electronic health record data, i.e. densely time-stamped bedside measurements, laboratory results, and interventions, showed that temporally resolved ICU data can support repeated risk estimation throughout the ICU stay rather than a single forecast at admission [62].

3.3.1 Structured ICU Time Series

The first major strand of this literature models time-stamped but still structured ICU variables, for example vital signs, laboratory values, interventions, and summary measurements sampled over fixed intervals. Early and influential work used recurrent neural networks to update mortality risk over time, illustrating that sequence models can monitor patient state as new observations become available [44]. More recent comparative studies broadened the set of candidate architectures. For example, [21] evaluated temporal convolutional networks (TCNs) on 24-hour clinical time series from MIMIC-III for both mortality and LOS-related tasks and reported that TCNs were competitive with, and in some mortality metrics superior to, several baselines for prediction. Taken together, recent ICU time-series work is no longer defined by static classifiers alone, but by sequence-aware models that explicitly exploit temporal ordering and patterns of absent or irregularly observed measurements [21, 30, 44].

3.3.2 Continuous Monitoring and Waveform-Adjacent Data

A second strand extends beyond tabular snapshots towards continuously sampled physiological monitoring. This is particularly relevant in intensive care, where bedside monitors generate dense streams of heart rate, blood pressure, respiratory, and oxygenation signals.

Such data can add prognostic information that is lost when only admission summaries or coarse aggregates are retained [62, 67]. For example, [67] showed in a pediatric ICU cohort that adding continuous vital-sign information from bedside monitors to static clinical variables improved prediction of prolonged LOS after intubation. Similarly, dynamic ICU models based on high-frequency EHR data indicate that repeated temporal updates can improve discrimination and clinical usability compared with single-time-point approaches [44, 62]. However, much of this literature still operates on derived multivariate clinical time series or monitor summaries rather than on raw waveform signals themselves [62, 67].

3.3.3 ECG as a Specialized Time-Series Modality

Beyond structured ICU time-series modelling, DL for raw ECG analysis has also advanced rapidly. CNNs established strong baselines for end-to-end ECG modelling, while newer transformer-based and self-supervised approaches aim to capture longer-range temporal structure and improve label efficiency [49]. Recent foundation models such as ECG-FM suggest that large-scale pretraining on very large 12-lead ECG corpora can yield transferable ECG representations that support downstream interpretation tasks and selected screening or risk-prediction applications [68, 69]. Likewise, large population-level studies indicate that raw 12-lead ECGs contain substantial prognostic information for mortality prediction far beyond classical rhythm classification tasks [70]. The important caveat is that this ECG literature is mostly centred on cardiac diagnosis, general cardiovascular risk, or broad hospital populations rather than ICU-specific operational outcomes such as LOS [49, 69, 70].

3.3.4 Current Limitations in the Literature

Despite this progress, the literature remains fragmented across modalities and clinical settings. ICU time-series studies predominantly model structured multivariate EHR trajectories, whereas advanced ECG studies predominantly address interpretation tasks or non-ICU risk prediction [14, 49]. The closest study to the setting considered here is the recent work by [20], which combined an ECG-derived risk score from the first ECG after ICU admission with the Acute Physiology Score (APS) III and age to predict 28-day mortality in ICU patients. This represents an important step towards ECG-informed ICU prognostics, but it does not address ICU LOS, does not compare multiple DL architectures at the raw-waveform level, and does not examine in a systematic way how predictive value changes when ECG information is added to structured clinical data [20].

Accordingly, the literature still lacks a unified evaluation framework for early ICU outcome prediction from the *first raw 12-lead ECG recorded at ICU admission*, especially when both

mortality and LOS are considered. The availability of linked resources such as MIMIC-IV-ECG now makes this open question empirically accessible by connecting diagnostic ECG waveforms with detailed ICU clinical records [20, 26].

Building on this opportunity, this thesis addresses these gaps by establishing a unified evaluation framework for early ICU outcome prediction that starts from the first raw 12-lead ECG at ICU admission and then evaluates which additional structured clinical information available within 6 hours of ICU admission is meaningful to include. It systematically compares multiple DL architectures and examines the incremental value of adding structured clinical information to ECG-based prediction for both mortality and LOS.

4 Data Understanding

This chapter describes the data basis of the study. It first introduces the underlying MIMIC-IV clinical database and the linked MIMIC-IV-ECG waveform subset and then characterises the resulting study cohort by means of exploratory analysis and a subsequent summary of the most relevant dataset properties. The aim is to make explicit which data modalities are available, how records are linked to ICU stays and outcomes, and which empirical properties characterise the final cohort used in the remainder of this work.

4.1 Dataset Overview

The overview is divided into two parts: the source databases and identifiers used for linkage, followed by the cohort construction steps.

4.1.1 Data Sources and Identifiers

The empirical basis of this thesis is the combination of the general MIMIC-IV clinical database and the linked MIMIC-IV-ECG waveform subset. MIMIC-IV is a large deidentified critical-care resource derived from Beth Israel Deaconess Medical Center and organised into a hospital-wide `hosp` module and an ICU-specific `icu` module [71]. The database contains data on more than 65,000 ICU patients and more than 200,000 patients with emergency department admissions, together with demographics, admissions, transfers, laboratory results, diagnoses, medications, charted events, and outcomes [71]. The ECG component is provided through MIMIC-IV-ECG, a matched subset that contains approximately 800,000 diagnostic 12-lead ECGs across nearly 160,000 unique patients [26]. Each recording is 10 seconds long and sampled at 500 Hz, which makes the dataset well suited to waveform-based ML on raw ECG signals [26]. Importantly, patients in MIMIC-IV-ECG are matched to the MIMIC-IV clinical database, which enables joint analyses of ECGs and structured EHR variables [26].

At the raw-data level, each ECG is provided as a WFDB record consisting of a header file (`.hea`) and a signal file (`.dat`). In the original resource, the files are organised hierarchically by patient and study, with directory names following the pattern `p{subject_id}` and `s{study_id}` [26]. This structure is convenient for archival access, but it is not yet directly suited for model training. In the project pipeline, the raw WFDB records are therefore indexed and linked to ICU metadata to enable subsequent cohort construction and analysis.

The linkage between both resources relies on deidentified patient and visit identifiers. In MIMIC-IV, the identifiers `subject_id`, `hadm_id`, and `stay_id` connect patients, hospital

admissions, and ICU stays across tables [71]. In MIMIC-IV-ECG, each diagnostic recording is assigned a `study_id`, and the released metadata allows direct linkage of ECGs to the corresponding `subject_id` and ECG timestamp [26]. This makes it possible to align diagnostic ECGs with ICU stays, outcomes, and selected EHR-derived covariates. At the same time, the source documentation highlights two important constraints: first, date-time information is deidentified by subject-specific date shifting, and second, ECG machine clocks may not be perfectly synchronised with the timestamps of the clinical database [26, 71]. Consequently, temporal matching has to be handled conservatively. Because the cohort construction combines waveform data, ECG metadata, and ICU tables, several identifiers recur throughout the remainder of this chapter. Table 1 therefore provides a compact reference for the most important IDs and timestamps used in the dataset description.

ID	Description
`subject_id`	Deidentified patient identifier
`hadm_id`	Identifier of a hospital admission
`stay_id`	Identifier of a specific ICU stay
`study_id`	Identifier of a specific diagnostic ECG recording
`waveform_path`	File path pointing to a WFDB waveform record
`ecg_time`	Timestamp of the ECG recording
`intime`, `outtime`	Start and end time of an ICU stay
`deathtime`	Documented time of death, if available

Table 1: Central identifiers and timestamps for ECG–ICU cohort construction.

4.1.2 Dataset Construction

The analysis cohort is constructed from clinical ICU tables and ECG metadata. On the clinical side, `icustays.csv` must provide at least `subject_id`, `hadm_id`, `intime`, and `outtime`. Rows with missing ICU start or end times are excluded from cohort construction. On the ECG side, the metadata index must contain at least `subject_id`, `study_id`, and `waveform_path`, together with a valid timestamp parsed from the first available column among `ecg_time`, `charttime`, and `record_time` . In practice, these csv-based metadata are more suitable for reproducible cohort construction than the raw waveform headers alone. This distinction matters because timestamp quality is one of the central limitations of MIMIC-IV-ECG. The WFDB headers include `base_date` and `base_time`, but these values originate from ECG machine clocks whose synchronisation with the clinical information system is not guaranteed [26]. For this reason, matching steps in the project prioritise csv-derived ECG timestamps where available and use header informa-

tion only as a fallback. This choice is important for constructing reproducible ICU-linked ECG cohorts, especially when restricting the analysis to early recordings after ICU admission.

For this thesis, the cohort construction is described in three stages. The full MIMIC-IV-ECG resource serves as the reference set for percentage calculations. From this resource, an ICU-linked subset is constructed first, followed by a more restrictive early-ICU subset containing ECGs recorded during the first 24h after ICU admission. Figure 5 illustrates this progressive restriction. As indicated by the figure, the reduction from the full MIMIC-IV-ECG resource to the final analysis cohort follows two consecutive filtering steps.

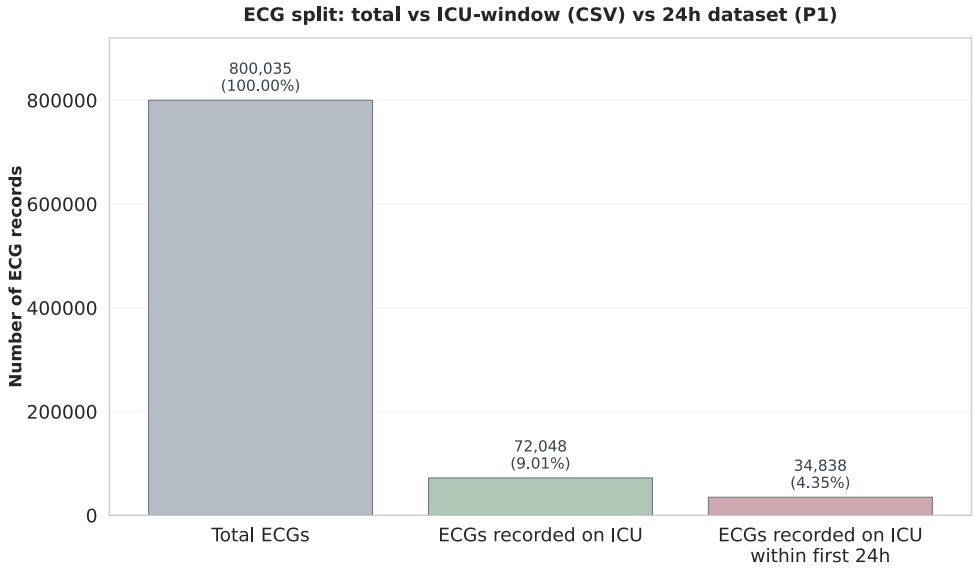

Figure 5: Successive restriction of the available ECG data from the full MIMIC-IV-ECG resource to ICU-linked ECGs and the subset recorded within the first 24 h of ICU stay.

In the first step, the full MIMIC-IV-ECG resource is reduced to ECGs recorded during an ICU stay. An ECG enters this ICU-linked subset ("ECGs recorded on ICU") only if the following linkage conditions are satisfied:

- `subject_id` in the ECG metadata matches `subject_id` in the ICU stay data.
- The ECG timestamp satisfies `intime` $\leq$ `ecg_time` $\leq$ `outtime`.
- If `hadm_id` is present in the ECG metadata, it must match the ICU `hadm_id` exactly. If `hadm_id` is missing, the ECG is kept only if exactly one ICU stay remains after temporal matching. Ambiguous matches are discarded.

In addition, two data-integrity requirements are imposed before an ECG is kept in the ICU-linked subset:

- Duplicate entries are removed on `waveform_path`, so that each waveform record appears at most once.
- An ECG is only kept and copied if both waveform files, `.hea` and `.dat`, are present.

In the second step, the ICU-linked subset is further restricted to ECGs recorded within the first 24 h after ICU admission. Starting from ECGs that already satisfy the ICU-linkage criteria above, an ECG is retained in this early-ICU subset ("ECGs recorded on ICU within first 24 h") only if the following additional conditions are met:

- The ECG satisfies the early-window condition $\texttt{intime} \leq \texttt{ecg\_time} < \texttt{intime} + 24$ h.
- If multiple ICU stays satisfy the condition, the stay with minimal absolute time distance $|\texttt{ecg\_time} - \texttt{intime}|$ is selected.
- An ECG is only kept and copied if both waveform files, `.hea` and `.dat`, are present.

Clinically, restricting the cohort to the first 24 h after ICU admission focuses the analysis on information that is available early in the patient trajectory and therefore better reflects an early risk stratification setting than ECGs recorded later during the stay. This progressive filtering also leads to a substantial reduction in cohort size. Starting from 800,035 ECGs in the full MIMIC-IV-ECG resource, the ICU-linked subset comprises 72,048 recordings (9.01 %), while the subset restricted to the first 24 h of ICU stay contains 34,838 recordings (4.35 % of the full resource). This reduction is expected because not every diagnostic ECG can be linked to an ICU stay, and not every ICU-associated ECG falls into the early time window relevant for the prediction task. Based on this cohort, the target variables considered in this work are ICU mortality and ICU LOS. LOS is defined at the ICU-stay level as the difference between `outtime` and `intime`, as represented in the linked `icustays` data [71]. ICU mortality is defined by checking whether a documented `deathtime` falls within the ICU interval [`intime`, `outtime`]. Under this operational definition, ICU mortality is assigned only when a death timestamp is explicitly available and lies within the ICU stay. Stays with missing `deathtime` are therefore not labelled as ICU mortality from this field alone, and deaths after ICU discharge are not counted as ICU mortality. Since both outcomes are defined at the patient or stay level, whereas ECGs are recorded at study level, the central data-engineering task is to construct a coherent ECG-centric cohort in which each waveform can be assigned to a clinically meaningful ICU context. To ensure reproducibility, this thesis uses MIMIC-IV-ECG version 1.0 (published Sept. 15, 2023) and MIMIC-IV version 3.1 (published Oct. 11, 2024). These release versions together with the extraction date should be reported with the experimental setup. The detailed preprocessing and model-specific preparation steps are described in chapter 5.

4.2 Exploratory Data Analysis

After establishing the ICU-linked ECG cohort, an exploratory analysis was carried out to describe the distribution of demographic variables and target variables in the data. Given that ECG-based prediction tasks are constructed at the level of individual recordings, the descriptive statistics in this section are reported primarily at the ECG-record level.

For the 24 h-ICU subset, the resulting cohort contains $n = 34,838$ ECGs recorded within the first 24 h after ICU admission. This subset is smaller than the overall MIMIC-IV-ECG resource and is defined by the availability of diagnostic ECGs and their temporal relation to ICU admission [26, 71]. These records originate from 21,408 unique patients, corresponding to an average of 1.63 ECGs per patient in the 24 h cohort.

The unit of analysis in this section is the ECG record rather than the patient. A single patient may contribute more than one diagnostic ECG to the ICU-linked cohort, and a single hospitalisation may contain multiple ECGs with different temporal relations to ICU admission. The descriptive statistics therefore portray the ECG-based cohort at the record level, even though outcomes such as LOS and mortality originate from stay-level or patient-level clinical information. Accordingly, patient-level variables such as age and sex are attached to each ECG record and are therefore usually repeated across multiple rows for the same patient rather than collapsed to a single patient-level observation. This record-level representation is appropriate for ECG-centred prediction, but it also creates a potential leakage risk if ECGs from the same patient are assigned to different data splits. To avoid this form of leakage, dataset partitioning in the present work was performed at the patient level, so each patient is assigned to a single data split.

The age distribution of the whole cohort is shown in Figure 6. The mean age was 66.8 years and the median age was 68.0 years. The distribution spanned a broad age range, with a smaller proportion of ECG records in younger adult age groups.

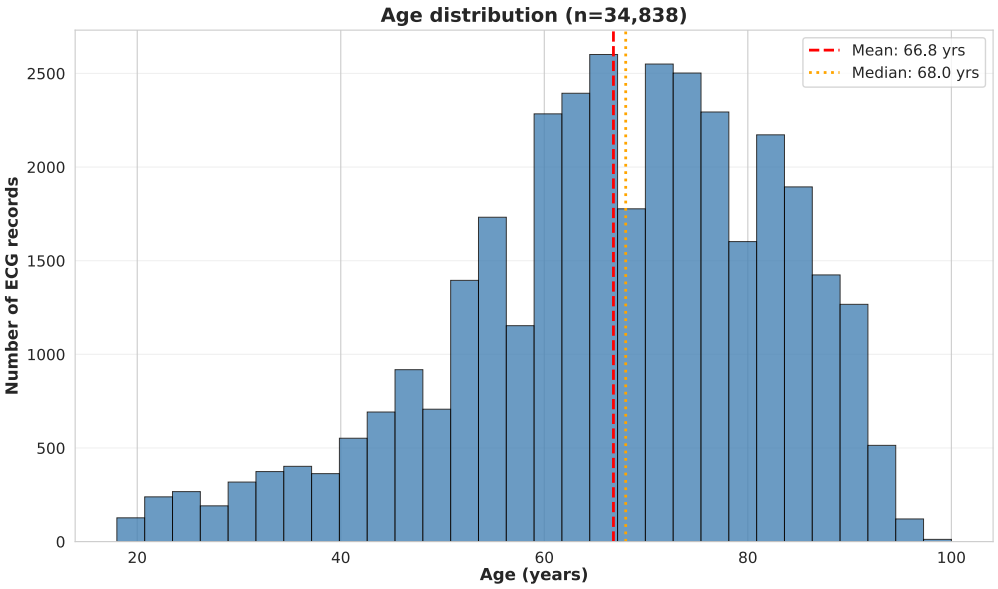

Figure 6: Age distribution of ECG records in the subset restricted to recordings obtained within the first 24 h of ICU stay.

Sex was distributed as follows: male ECG records accounted for 20,802 observations (59.7 %), and female ECG records for 14,036 observations (40.3 %), as shown in Figure 7. The mortality distribution is shown in Figure 8. A total of 2,700 ECG records were associated with patients who died (7.8 %), whereas 32,138 ECG records were linked to survivors (92.2 %).

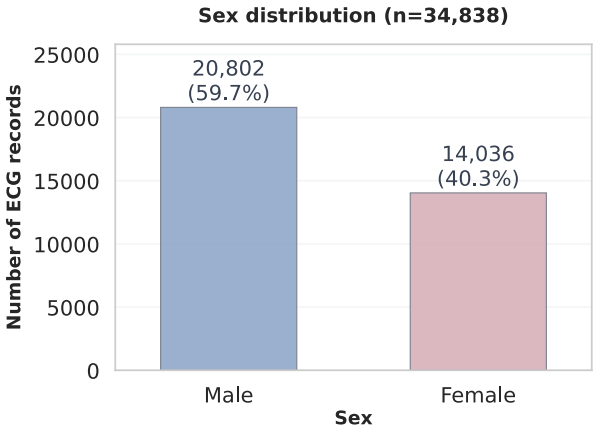

Figure 7: Sex distribution of ECG records in the 24 h ICU subset.

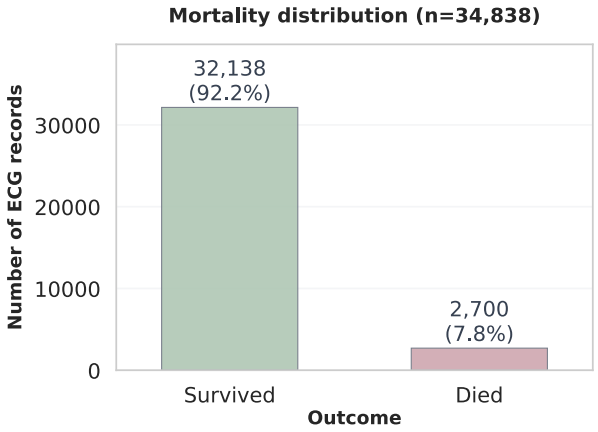

Figure 8: Mortality distribution in the 24 h ICU ECG cohort.

LOS is distributed asymmetrically. As shown in Figure 9, short ICU stays accounted for the largest share of the cohort, with the highest proportion observed at one day (32.1 %), followed by two days (18.2 %) and stays shorter than one day (13.6 %). The proportion decreased with increasing duration, while stays of 15 days or more accounted for 3.4 % of ECG records.

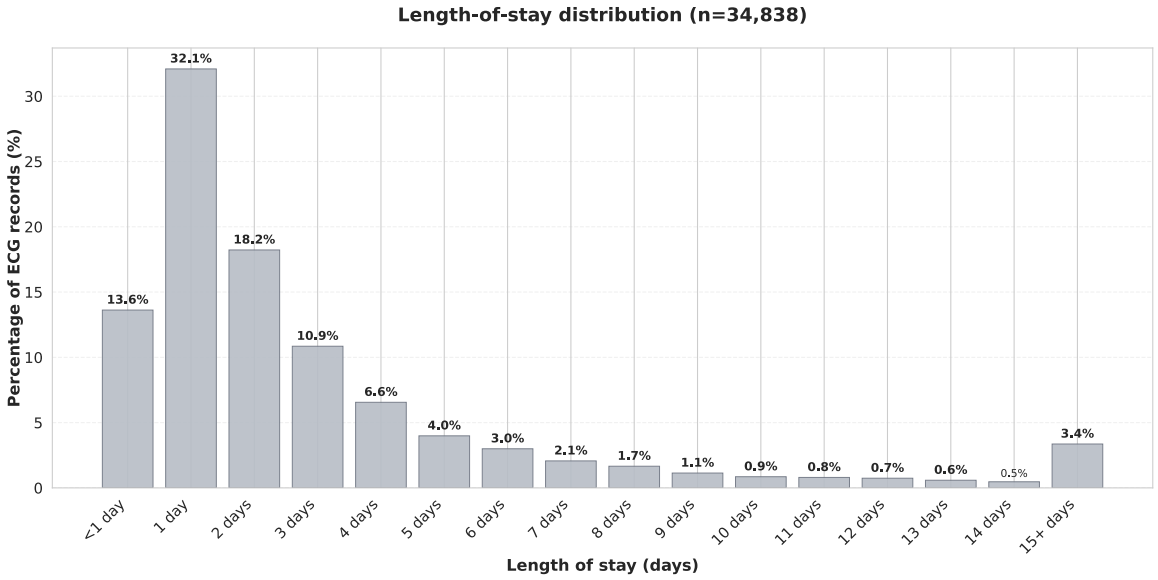

Figure 9: Distribution of ICU LOS in the 24 h ECG cohort.

In addition to waveform data, the dataset used in this work also contained structured EHR covariates describing the clinical state around ICU admission. These variables cover the main measurement domains required to represent acute organ status, including renal function, coagulation and haematology, liver function, cardiovascular status, respiratory status, and neurological status. At the dataset level, these variables complemented the ECG signals with tabular clinical information that was aligned to the ICU episode and ECG context and was assigned to the corresponding ECG records. Together, these covariates formed the structured basis for clinical severity representations such as SOFA-derived features. As with the demographic variables, the model pipeline retains these covariates at the ECG-record level rather than collapsing them to one row per patient.

Another relevant categorical property of the cohort was ICU location. This information is represented by the stay-level variable `first_careunit` and is assigned to each ECG via `stay_id`. Figure 10 shows the distribution of the ten most frequent ICU units in the 24 h cohort. In the current analysis, these top-10 units account for 100.0 % of ECG records. The most frequent categories are cardiovascular, medical, coronary, and surgical intensive care units, including labels such as `CVICU`, `MICU`, `CCU`, `MICU/SICU`, `SICU`, and `TSICU`. In addition, rare label variants occurred as separate categorical strings.

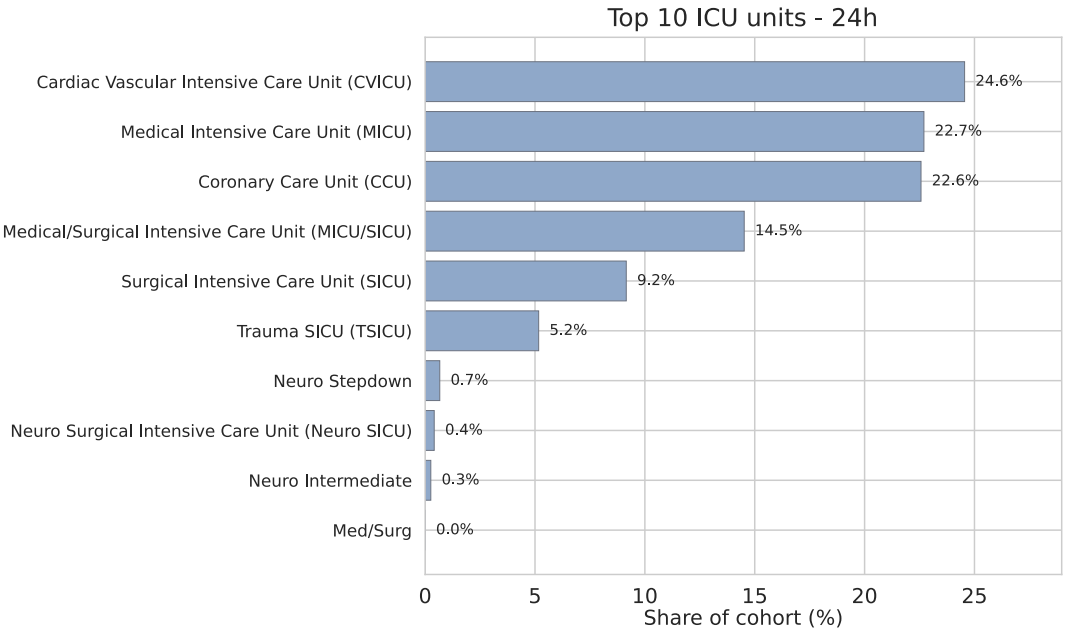

Figure 10: Top-10 ICU locations in the 24 h ECG cohort based on `first_careunit`.

The descriptive findings summarised in this section were taken into account in the following subsection.

4.3 Data Summary

Chapter 4 establishes the data basis for the subsequent modeling work. MIMIC-IV-ECG provides the diagnostic 12-lead ECG records, while MIMIC-IV provides ICU-stay information, outcomes, demographics, ICU-unit labels, and structured clinical variables. Linking both resources made it possible to derive an ECG-centred ICU cohort, to restrict it to recordings from the first 24 h after ICU admission, and to assign LOS, mortality, demographic, ICU-unit, and EHR covariates to each retained ECG.

Starting from 800,035 ECGs in MIMIC-IV-ECG, the ICU-linked subset contained 72,048 recordings, and the final first-24 h subset contained 34,838 ECG records from 21,408 unique patients. The resulting cohort is multimodal and record-level rather than purely patient-level. It is centred on older adults, with a mean age of 66.8 years and a median age of 68.0 years, and contains more male than female ECG records (59.7 % vs. 40.3 %). Mortality is rare at 7.8 %, whereas ICU LOS is concentrated in short stays with a long right tail. ECG records are also concentrated in a limited set of ICU unit categories. These properties define the dataset context for the model development and evaluation in the following chapters.

5 Methods

This chapter describes the methodological setup of the thesis. It first outlines the experimental setup and implementation context, followed by the preprocessing and augmentation steps applied before model training. The chapter then presents the baseline and DL architectures considered in this work, explains why specific model families were selected, and documents which approaches were pursued further during model development. In addition, this chapter describes the multimodal fusion strategy and the training and hyperparameter optimisation protocol used for model selection.

5.1 Experimental Setup and Implementation

All experiments in this thesis were developed within a shared technical infrastructure that was designed to support consistent implementation, execution, and documentation across different model families. The project was implemented primarily in Python, with `PyTorch` as the central framework for DL components. At a structural level, the repository was organised into dedicated directories for configurations, scripts, source code, data, and generated outputs. This modular organisation separates experiment definition, implementation, execution, and artifact storage, and thereby provides a stable basis for systematic experimentation. The corresponding code repository is available at `https://github.com/pr014/Master-thesis`.

A central design principle of the project is configuration-driven experiment management. Experimental settings are specified through YAML configuration files, while the underlying implementation logic remains shared across runs. This makes it possible to compare different architectures and model variants within one coherent framework without repeatedly modifying the training code itself. In addition to this technical separation, the project used a compact identifier system to organise experiments and related components consistently across scripts, outputs, and result tables. These identifiers denote the role of a component within the experimental pipeline and make it possible to refer to preprocessing variants, augmentation schemes, architecture choices, hyperparameter sets, and dataset definitions in a compact and standardised way. Table 2 summarises the prefix categories used throughout the project.

Prefix	Category	Meaning
`A`	Architecture	Refers to a model architecture or model family, e.g., CNN, LSTM, hybrid models, or pretrained foundation models.
`P`	Preprocessing	Describes a preprocessing pipeline or preprocessing stage applied before model training.
`H`	Hyperparameters	Identifies a predefined hyperparameter configuration used for a given model family.
`AG`	Augmentation	Refers to an augmentation strategy or augmentation bundle used during training.
`D`	Dataset Variant	Denotes the dataset definition or feature combination used in an experiment.

Table 2: High-level structure of experiment identifiers used throughout the project.

This identifier scheme was used to describe experiments as modules rather than as isolated training runs. A concrete run can therefore be viewed as the combination of several components, for example `A7 + P1 + AG2 + H3 + D8`, where each identifier refers to one defined element of the pipeline. This made it possible to document large numbers of experiments in a compact form and to compare related runs systematically across model families. The table is therefore not meant as a result summary, but as an overview of the organisational structure used to manage the experimental workflow.

The same codebase was used for local development, smaller-scale test runs, and large-scale training. Computationally demanding experiments were executed on GPU resources of the BW UniCluster 3.0 via SLURM-based job submission, while local entry scripts were mainly used for development, debugging, and preparation. This setup allowed the project to scale from prototyping to larger experiment series without changing the underlying implementation logic.

The infrastructure was also designed with reproducibility in mind. Configuration files, run scripts, logs, checkpoints, and other artifacts were stored in a structured manner so that individual experiments can be reconstructed and compared more easily. In this way, the technical setup forms the foundation for the modelling and training procedures described in the following sections. Detailed training settings and hyperparameter choices are discussed in Section 5.7, while quantitative results are reported in Chapter 6.

5.2 Preprocessing and Augmentation

Two complementary mechanisms shape the ECG inputs used in this work. *Preprocessing* transforms raw recordings into a fixed and comparable representation that is shared

across training, validation, and test data. *Augmentation* introduces stochastic perturbations only during training in order to increase robustness and reduce sensitivity to nuisance variability. Preprocessing therefore defines the common input space of the models, whereas augmentation acts as a training-time regularisation mechanism.

Preprocessing was used to standardise the signal representation before model training. Clinical ECG recordings can differ in signal quality, noise profile, and scaling. Therefore, prior work discussed filtering, normalisation, and sampling rate as explicit design choices in ECG ML pipelines, while also showing that no single recipe is universally optimal across datasets and tasks [72, 73]. In addition, the diagnostic ECGs in MIMIC-IV-ECG are provided as 12-lead recordings of 10 seconds at 500 Hz, which makes fixed-length input processing natural for this dataset [26]. For this reason, several preprocessing specifications were computed in advance and stored as fixed dataset variants.

Band-pass filtering was applied as the first step, with a lower cutoff of 0.5 Hz and an upper cutoff of 50 Hz. This choice was motivated by the need to suppress baseline drift and high-frequency noise before the signals are forwarded to the models. Prior ECG literature shows that filter design and cutoff choice can affect signal quality and clinical interpretation [73], while more recent work suggests that the empirical benefit of filtering is task-dependent rather than universal [72]. The use of a Butterworth filter was chosen because it provides a smooth passband response and is widely used in ECG preprocessing. The exact order and cutoff values should therefore be read as project-specific implementation choices rather than as literature-established optima. A visual example of this transformation is shown in figure 11.

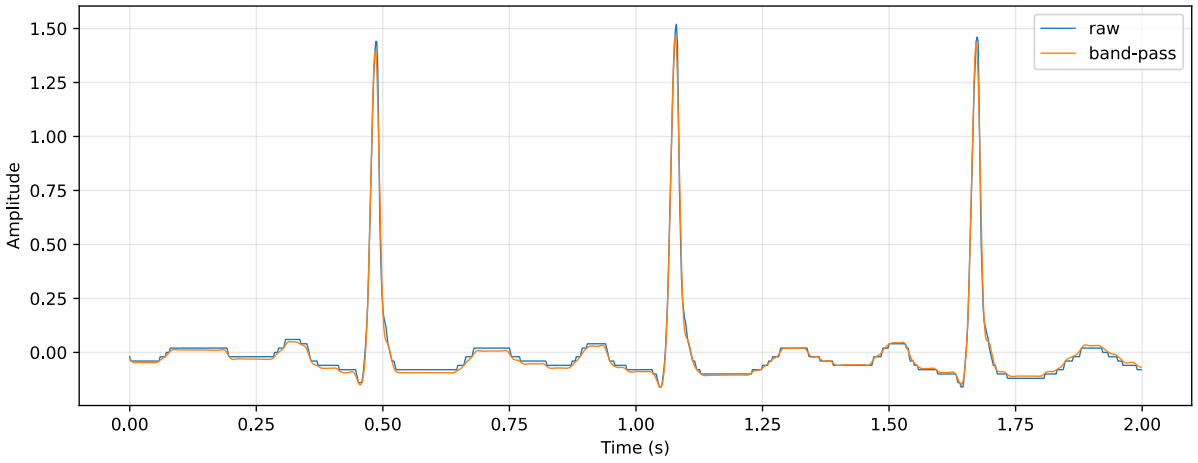

Figure 11: Example of band-pass filtering on an ECG signal.

After band-pass filtering, **notch filtering** was applied to attenuate interference at 60 Hz. This step was relevant because power-line interference is a common ECG noise source in general. The fact that the underlying data source is US-based mainly determined the target frequency of 60 Hz. In a German or other 50 Hz setting, the same notch-filtering step would have been equally relevant, but it would have targeted 50 Hz instead. Survey literature identified power-line interference as a common ECG noise source and described its removal as part of ECG denoising workflows [74]. In this study, the explicit notch stage was therefore treated as a conservative implementation decision to reduce acquisition-related artifacts before model ingestion.

Resampling and fixed-length segmentation then resampled the filtered signals to 500 Hz and segmented them into fixed 10-second windows, corresponding to 5000 samples per lead. This choice was aligned with the structure of MIMIC-IV-ECG, where diagnostic recordings are also represented as 12-lead, 10-second ECGs sampled at 500 Hz [26]. In practical terms, this step ensured that all models receive inputs of identical shape.

As a final optional step, **per-lead normalisation** was applied. This variant was included because prior work has shown that normalisation is an influential preprocessing decision in ECG ML, but not one with uniformly positive effects across settings [72]. At the same time, the non-normalised variant was retained because the absolute amplitude may still contain task-relevant information. For this reason, normalisation was treated as an explicit preprocessing choice encoded through the preprocessing ID rather than as a mandatory default.

Augmentation was applied only during training. The purpose was to expose the models to plausible perturbations that reflect real acquisition variability and to reduce the risk that small recording differences shift the model prediction undesirably. This is especially relevant in ECG modelling because recording conditions, motion artifacts, and lead-specific disturbances vary across clinical settings [74]. Recent large-scale ECG models likewise used ECG-specific augmentation as part of robust representation learning [68].

Additive **Gaussian noise** was used to simulate small-scale sensor noise and recording variability during training. This transform perturbs the signal locally while preserving the assigned target and is therefore a natural robustness-oriented augmentation for ECG waveforms. Both ECG augmentation surveys and task-specific augmentation studies describe noise injection as a standard perturbation family in ECG model training [74, 75].

Baseline wander was implemented as a low-frequency perturbation. The rationale for this choice was that low-frequency drift is a common artefact in ECG acquisition, shifting the isoelectric line and thereby affecting not only waveform morphology but also baseline position, apparent amplitudes, and overall signal quality. Including this perturbation was therefore intended to improve robustness to realistic recording conditions [74].

Lead dropout was introduced to simulate partial signal corruption or temporary degradation of individual leads. This is relevant for multilead ECG modelling because model performance can become fragile if uniformly clean information from all channels is assumed. By randomly removing leads during training, the model is encouraged to rely on distributed information across the available signal channels rather than over-specialising to single leads. This is conceptually close to dropping or masking operators in ECG augmentation surveys and to random lead masking in recent ECG foundation-model training pipelines [68, 74]. In implementation terms, affected leads were not removed from the tensor; instead, entire leads were set to zero across the full time axis, so the input shape remained unchanged. If all leads would have been zeroed simultaneously, one random lead was restored so that the model never received a completely silent multichannel input.

Amplitude scaling was included only in `AG2`. Conceptually, amplitude scaling multiplies the full waveform by a global factor, thereby changing overall magnitude while leaving temporal structure unchanged. In this project, such a transform was only considered meaningful when the non-normalised preprocessing variant `P2` was active, because `P1` already applied per-lead z-score normalisation and therefore largely removed global scale differences. The transform itself is nevertheless well represented in the ECG augmentation literature, where scaling is a common value-domain perturbation [74, 75].

In `AG5`, **oversampling** complemented the augmentation bundle through list-based duplication of minority-class training examples. More specifically, additional training-list entries were created for ECGs with mortality label 1, while each duplicate still pointed to the same underlying waveform file rather than to a separately generated copy. The number of additional entries per positive case was either fixed in the configuration or chosen automatically so that positive cases appeared approximately as often as negative ones in the training list. For these duplicate entries, a dedicated time-segment-shuffle augmentation was then applied using a deterministic seed derived from the base record and child index, so that each copy received its own reproducible segment permutation during training. Although oversampling is not a signal transformation in the narrow sense, it was grouped with augmentation in the project-wide identifier logic because it changes the effective training distribution. This decision is also consistent with prior ECG classification work that used oversampling strategies such as SMOTE to improve robustness under class imbalance [51].

5.3 Experiment Identifiers and Configuration Registry

To keep the experimental workflow reproducible across many runs, all configurations were encoded through compact identifiers. Rather than describing each experiment as a separate case, runs were defined as combinations of architecture, preprocessing, augmentation,

hyperparameter, and dataset identifiers. In compact form, one experiment can therefore be written as, for example, `A5 + P1 + AG4 + H3 + D8`. The following tables summarise the identifier system used throughout this thesis. Further details on preprocessing and augmentation are provided in Section 5.2, while training-related hyperparameter choices are discussed in Section 5.7.

Architecture IDs specify which model family, model variant, or pretrained backbone was used in an experiment. They therefore provide the central reference for distinguishing between classical baselines, neural architectures trained from scratch, and transfer-learning approaches based on pretrained ECG foundation models. In the experiment registry, the architecture ID defines the representation-learning backbone to which preprocessing, augmentation, hyperparameter, and dataset choices are attached.

ID	Model	Params.	Notes
`A1`	XGBoost	–	Gradient-boosted tree baseline
`A2`	CNN from scratch	46k	Fully trained from scratch
`A3`	Unidirect. LSTM	267k	Fully trained from scratch
`A4`	Bidirect. LSTM	662k	Fully trained from scratch
`A5`	Hybrid CNN-LSTM	714k	Fully trained from scratch
`A6.1`	DeepECG-SL	90M	Pretrained on `WCR_77_classes`
`A6.2`	DeepECG-SL	90M	Pretrained on `wcr_lvfe_equal_under_40`
`A6.3`	DeepECG-SL	90M	Pretrained on `wcr_lvfe_under_50`
`A6.4`	DeepECG-SL	90M	Pretrained on `wcr_afib_5Y`
`A7.1`	HuBERT ECG	92M	Base model
`A7.2`	HuBERT ECG	30M	Small model
`A7.3`	HuBERT ECG	92M	Large model

Table 3: Architecture identifiers used in the experiment registry.

Preprocessing IDs refer to the fixed ECG preprocessing pipelines introduced in Section 5.2. Table 4 summarises the frozen preprocessing variants used in the experiment registry.

ID	Combination
`P1`	Band-pass filtering, notch filtering, resampling to 500 Hz, and standardisation to a fixed 10 s model input (5000 samples). Signals shorter than this were zero-padded at the end, whereas longer signals were centre-cropped. Per-lead z-score normalisation was then applied. This was the default precomputed dataset variant for most experiments
`P2`	Same processing chain as `P1`, but without per-lead z-score normalisation. Signal amplitudes therefore remain in their pre-normalised physical units until any later model-specific scaling

Table 4: Preprocessing identifiers and corresponding operator combinations.

Hyperparameter IDs refer to predefined training configurations that were reused across repeated experiments within a model family. Because recurrent models, hybrid models, and pretrained encoders differ substantially in optimisation behaviour, the same hyperparameter label does not imply the same numerical values across all architectures. Instead, hyperparameter IDs were defined relative to the corresponding model family. At the same time, the label `H1` consistently denotes the respective baseline starting configuration within that model family. It therefore serves as the reference setting used for the controlled comparison stage before more extensive tuning is introduced. Table 5 lists the available identifiers by model family, while the exact parameter values of these configurations are discussed in Section 5.7.

Model family	Available IDs
Unidirect. LSTM(`A3`)	`H1-H4`
Bidirect. LSTM (`A4`)	`H1-H4`
Hybrid CNN–LSTM (`A5`)	`H1-H5`
DeepECG-SL(`A6.1-A6.4`)	`H1-H3`
HuBERT-ECG(`A7.1-A7.3`)	`H1-H2`

Table 5: Hyperparameter identifiers by model family.

Augmentation IDs encode the set of training-time perturbations applied to the ECG signals. Table 6 summarises the final augmentation bundles tested in the project. These bundles were introduced to study the effect of robustness-oriented signal transformations in a structured and reproducible way. Importantly, augmentation IDs only affect the

training data and were not applied during validation or testing, so that model selection and final evaluation remained based on unaugmented inputs.

ID	Combination
AG0	No augmentation
AG1	Gaussian noise (random noise with standard deviation 3% of signal std)
AG2	Gaussian noise and amplitude scaling (random scaling factor 0.85–1.15)
AG3	Gaussian noise and lead dropout (randomly zeroing out leads with $p = 0.1$ per lead)
AG4	Gaussian noise, lead dropout, and baseline wander (sinusoidal drift with frequency 0.15–0.4 Hz, amplitude 0.03–0.08, and random phase per lead)
AG5	Same signal-space bundle as AG4, plus oversampling of mortality classes during training

Table 6: Final augmentation bundles tested in the project.

Dataset IDs define which ECG subset was used and which additional non-ECG information was made available to the model. Table 7 summarises the final dataset IDs on which all experiments in this thesis were based. They therefore play a central role in separating unimodal ECG-only experiments from multimodal settings that incorporate demographics, ICU-unit information, or extended EHR features. In the context of this thesis, the dataset ID is also the main mechanism for encoding which late-fusion input combination belongs to a given experiment.

ID	Dataset	Late fusion - additional information
D1	icustay_ecgs_24h	
D2	icustay_ecgs_24h	demographics (age + sex)
D3	icustay_ecgs_24h	icu_unit
D4	icustay_ecgs_24h	demographics and icu_unit
D5	icustay_ecgs_24h	EHR_feature_data, icu_unit, demographics
D6	icustay_ecgs_24h	EHR_feature_data_small, icu_unit, demographics
D7	balanced_mortality	mortality prediction only

Table 7: Final dataset identifiers and associated late-fusion input combinations.

5.4 Baseline Model

The classical baseline in this work was `XGBoost` (`A1`). It provided a fast and reproducible reference against the end-to-end DL architectures described in Section 5.5. The purpose of this baseline was not to define an upper performance bound, but to test how far a conventional feature-based ECG pipeline could go before learned waveform representations and multimodal fusion were added. Tree-based and gradient-boosting models are commonly used as transparent reference methods in ICU outcome prediction and ECG-based mortality studies [8, 20, 65, 70].

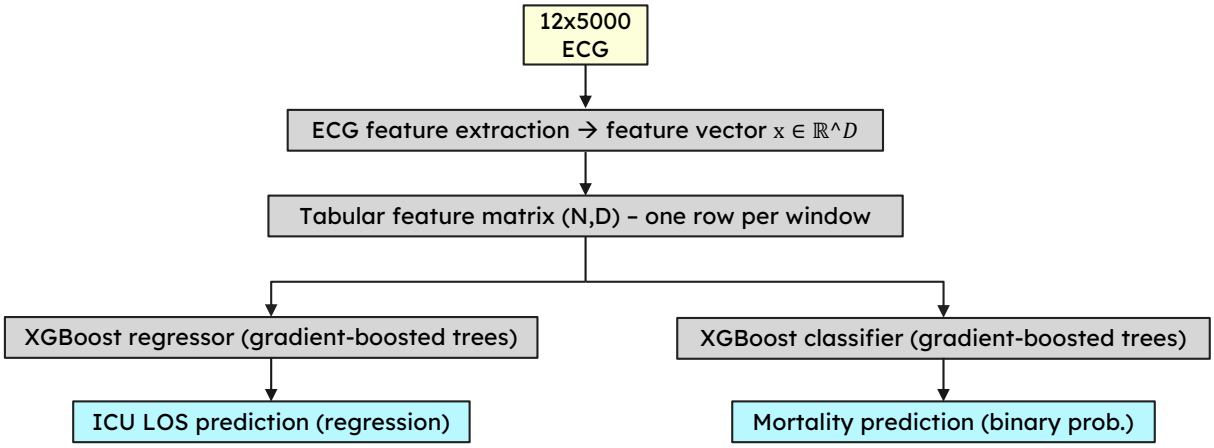

Figure 12: Schematic ECG-to-feature-to-`XGBoost` pipeline. B denotes batch size, N the number of ECG windows, and D the feature dimension.

Figure 12 summarises the baseline architecture. Unlike the neural models, `XGBoost` did not process the ECG as a raw time-series tensor. Each 10-s 12-lead ECG window sampled at 500 Hz was first transformed into a fixed tabular feature vector. The same preprocessing basis as in the main pipeline was used: band-pass filtering, per-lead z-score normalisation, and the `P1` ECG input format. In the standard ECG-only configuration `D1`, demographic and EHR features were not included.

The handcrafted ECG representation had approximately 366 dimensions per recording. For each of the 12 leads, 25 descriptors were extracted, covering time-domain summaries such as mean, standard deviation, minimum, maximum, median, skewness, kurtosis, quantiles, interquartile range, signal range, energy, mean absolute value, root mean square, and zero-crossing rate. Frequency-domain descriptors include dominant frequency, low-, very-low-, and high-frequency power, total power, spectral centroid, and spectral roll-off. In addition, 66 cross-lead correlations from the upper triangle of the 12-by-12 lead-correlation matrix captured simple inter-lead relationships. This converted the waveform into a compact descriptor vector containing amplitude, variability, spectral, and cross-lead information.

Two separate XGBoost models were trained on this shared feature table: an `XGBRegressor` for LOS in days and an `XGBClassifier` for binary in-ICU mortality. Thus, the baseline used the same ECG windows, temporal split, and labels as the DL experiments, but it does not perform end-to-end representation learning. Its role was therefore to provide a strict ECG-only classical ML reference for assessing the added value of the neural waveform models and the later multimodal extensions.

5.5 Deep Learning Models

The DL model family represents a central methodological pillar of this thesis. In contrast to the classical baseline from section 5.4, these models do not rely on manually engineered features, but learn task-relevant representations directly from the ECG waveform. Methodologically, the family is structured as a progression from simpler to more complex architectures: from compact convolutional and recurrent models trained from scratch, through hybrid designs, to large pretrained ECG foundation encoders. This ordering defines the experimental backbone and keeps increases in capacity and pretraining explicit rather than ad hoc. All DL models received the same preprocessed 12-lead ECG input with shape 12×5000, corresponding to a 10 s recording at 500 Hz. Details of the shared signal preparation are given in section 5.2.

Although the architectures differ substantially in complexity and inductive bias, the downstream training logic is intentionally kept as consistent as possible. Where enabled, models were trained in a multitask setting with a shared ECG backbone and two prediction heads: one for LOS regression and one for mortality prediction. This follows the general idea that clinically related targets can regularise a shared latent representation and improve data efficiency, especially when the available cohort is limited relative to model complexity [30]. At the same time, optimisation details remain family-specific where required by the underlying architecture or pretraining strategy, as described in section 5.7.

The following subsections specify each architecture in ascending identifier order from `A2` to `A7`.

5.5.1 A2: CNN from Scratch

The `A2` model defines the lowest-complexity end-to-end waveform architecture in the DL model family. Its encoder consists of three successive `Conv1D` blocks with increasing channel width $12 \rightarrow 32 \rightarrow 64 \rightarrow 128$, each followed by batch normalisation, `ReLU` activation, and max pooling. The implemented kernel sizes decrease from 7 to 5 to 3, which allows the network to first capture the broader local waveform structure and then refine shorter-scale details. A global average pooling layer compresses the temporal dimension into a fixed

128-dimensional representation, which is passed to a compact fully connected prediction head. Figure 13 shows the corresponding layer structure and end-to-end data flow.

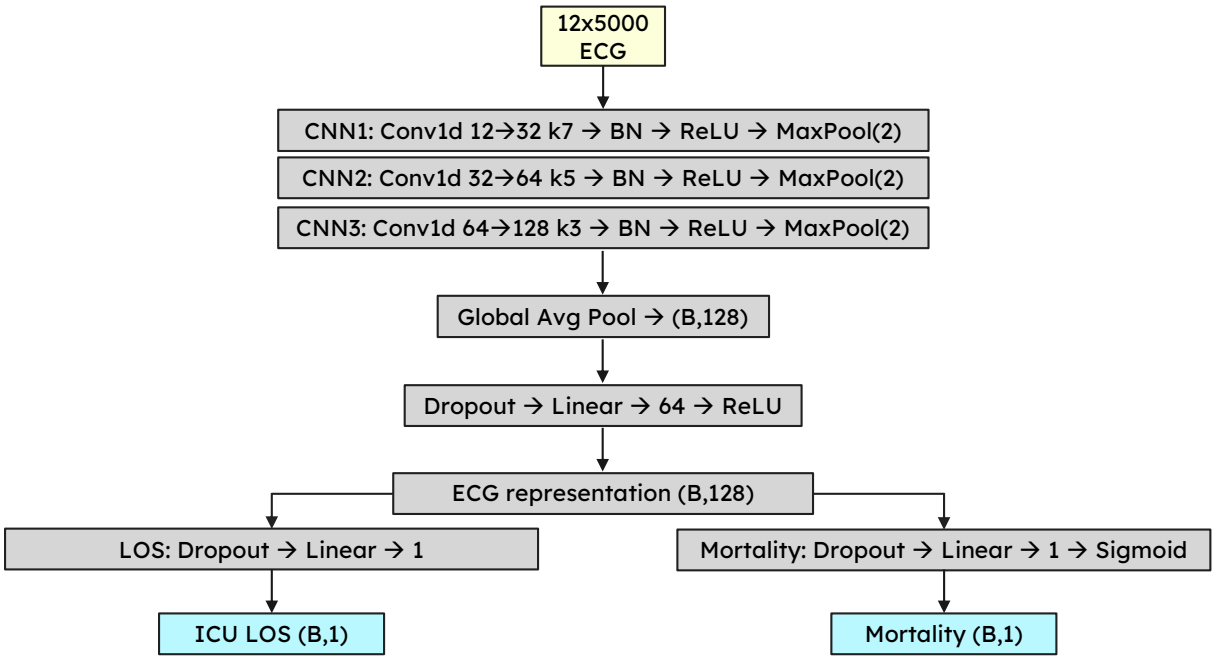

Figure 13: Schematic architecture of the `A2` CNN model with three `Conv1D` blocks and global average pooling.

From an architectural perspective, this model is intentionally simple. It does not include residual connections, attention, recurrence, or large hidden projections. Its purpose is not to maximise capacity, but to provide a controlled and interpretable starting point for end-to-end ECG modelling. In particular, the stacked `Conv1D` blocks can be understood as progressively learning local morphological detectors: earlier layers operate close to the raw waveform and are expected to respond to small-scale signal shapes, while later layers integrate these local responses into more abstract lead-wise descriptors. The use of global average pooling further enforces a compact representation and prevents the classifier head from depending on a large flattened temporal tensor. This architecture was included because CNNs are among the most commonly used and best established DL architectures in ECG analysis [36, 76]. For one-dimensional ECG signals, convolutional filters are well suited to learn local morphological motifs such as P-waves, QRS-complexes, ST-changes, and T-wave patterns directly from raw or lightly processed waveforms [36]. Large ECG benchmark studies have also shown that convolutional models form strong reference architectures for end-to-end ECG classification [31, 35]. Within this thesis, `A2` therefore serves as the simplest DL reference model and provides the basic convolutional baseline against which the recurrent, hybrid, and pretrained architectures are compared.

5.5.2 A3: Unidirectional LSTM

The `A3` architecture interprets the ECG as an ordered sequence rather than a set of local convolutional patches. After rearranging the input from $(B, 12, 5000)$ to $(B, 5000, 12)$, the 12-lead amplitudes at each timestep are optionally projected into a higher-dimensional embedding space. The sequence is then processed by two stacked LSTM layers with a hidden dimension of 128 each and with dropout between the two layers. Temporal information is aggregated by mean pooling over all timesteps, followed by batch normalisation, dropout, and a linear prediction layer. Because the recurrent processing is strictly forward in time, each hidden state depends only on current and past samples within the ECG segment. Figure 14 shows the schematic architecture of the `A3` model.

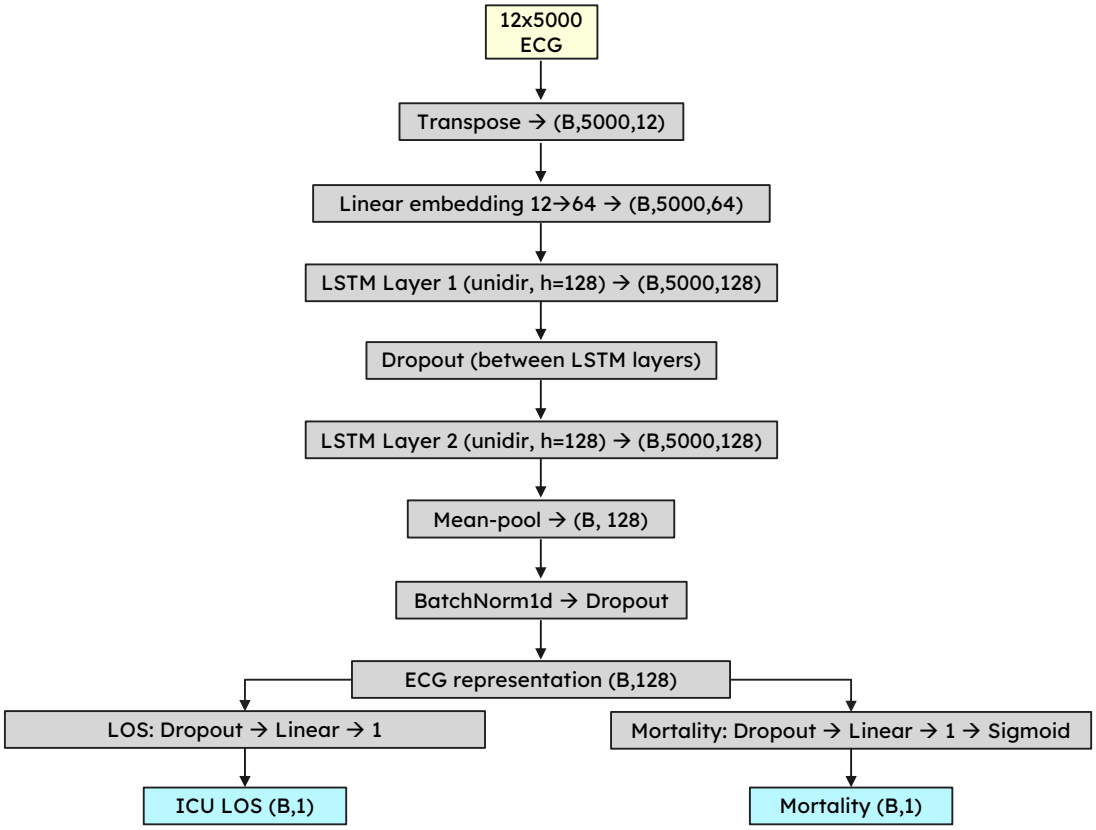

Figure 14: Schematic architecture of the `A3` unidirectional LSTM model with two recurrent layers and temporal mean pooling.

The key modelling assumption behind `A3` is that the ECG should be treated as a sequence of multivariate observations evolving over time. Whereas the CNN baseline primarily emphasises local receptive fields, the LSTM explicitly maintains a recurrent state and is therefore able to accumulate temporal evidence over longer ranges. In the context of ECG signals, this is relevant because certain clinically meaningful characteristics are not purely local but depend on temporal regularity, rhythm, beat-to-beat dynamics, and

the interaction between morphology and timing. This model was chosen as the causal recurrent reference architecture. LSTM networks remain a standard tool for clinical and physiological time-series modelling because they can summarise temporal dependencies more explicitly than shallow feed-forward models [30, 36]. They have also been used successfully in ICU outcome modelling and monitoring settings [44, 46]. Within this thesis, the unidirectional variant is methodologically important because it provides a sequence model that is closer to a causal or online interpretation of the signal. Although the present evaluation is retrospective and based on complete 10-second windows, the strict left-to-right information flow of `A3` still offers a useful conceptual contrast: it tests what can be learned when the model only has access to the "past" of the sequence at each internal step. This makes `A3` the unidirectional recurrent reference model in the architecture family. It was included to represent a strictly sequential ECG encoder, to contrast recurrent modelling with the purely convolutional baseline `A2`, and to provide the causal counterpart to the bidirectional recurrent model `A4`.

5.5.3 A4: Bidirectional LSTM

The `A4` model retains the same general pipeline as `A3`—two recurrent layers, temporal pooling, and a simple head—but replaces the recurrent core with bidirectional LSTMs. Each layer therefore processes the sequence in both forward and backward direction and concatenates both hidden states, yielding a richer contextual representation at every timestep. In the default configuration, this doubles the effective hidden output dimension relative to the unidirectional variant before mean pooling is applied across the full 10-second window. Figure 15 shows the schematic architecture of the `A4` model.

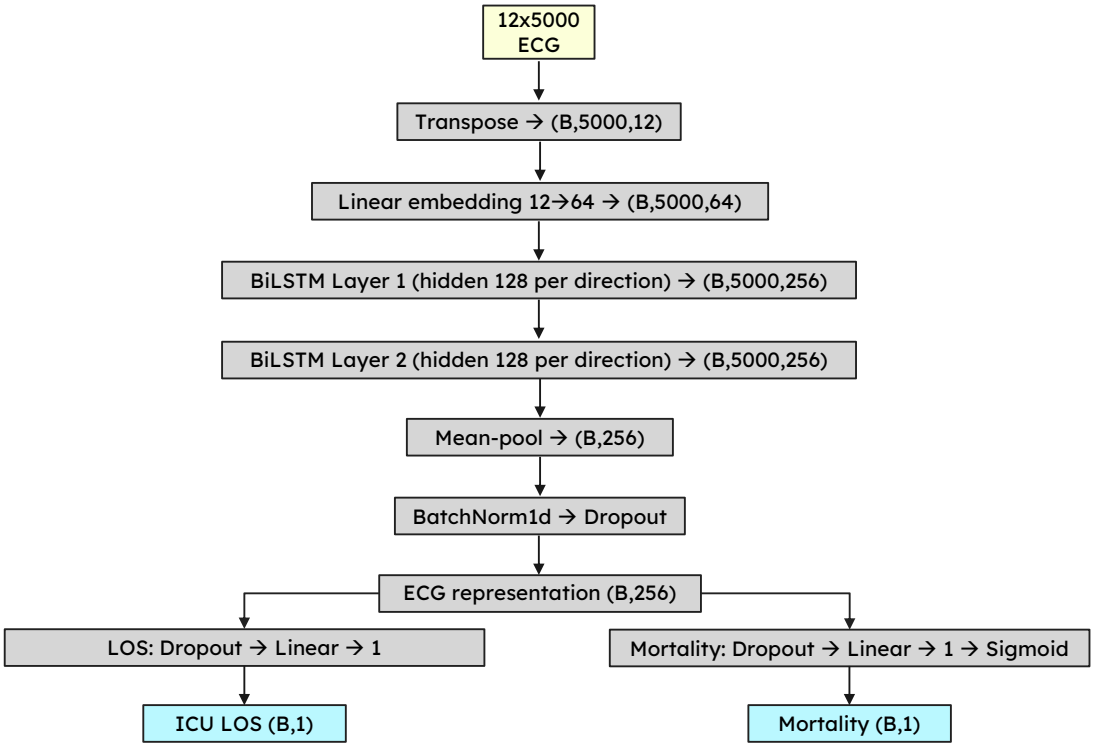

Figure 15: Schematic architecture of the `A4` bidirectional LSTM model with bidirectional recurrent encoding and temporal mean pooling.

The conceptual difference from `A3` is therefore not merely a larger hidden size, but a different way of defining temporal context. In a bidirectional recurrent encoder, each internal representation can integrate evidence from both earlier and later portions of the signal. For fixed ECG windows this is a meaningful design choice, because the entire waveform is available at inference time and there is no requirement to preserve strict causality. As a result, the model can construct a more globally informed representation of each segment than a unidirectional recurrent network. The main reason for including `A4` is this offline evaluation setting. Because the downstream task is performed on complete 10-second ECG windows rather than on streaming signals, information from both temporal directions can be used during encoding. Bidirectional recurrent modelling is therefore a natural sequence baseline for this setup [31, 36]. Within the model family, `A4` serves as the bidirectional recurrent counterpart to `A3` and as the recurrent reference architecture for the hybrid model `A5`.

5.5.4 A5: Hybrid CNN–LSTM

The `A5` architecture combines a convolutional front-end with a recurrent back-end. First, three `Conv1D` blocks extract local morphological ECG features while progressively reducing temporal resolution. Compared with `A2`, the first convolution is kept slightly narrower

and dropout is already applied after the convolutional blocks. This reflects the more specialised role of the front-end in the hybrid model: the CNN extracts local features, whereas the bidirectional LSTM performs the later temporal integration, and the higher-capacity architecture benefits from earlier regularisation. The resulting feature sequence is then processed by a two-layer bidirectional LSTM, so that the recurrent module no longer operates on raw amplitudes but on already transformed waveform descriptors. After mean pooling over time, the model applies a shared fully connected head with dropout and a dense hidden layer before the final output layer. Figure 16 illustrates this processing chain in schematic form, from the initial 12×5000 ECG input through convolutional feature extraction and bidirectional recurrent modelling to the pooled representation and final task-specific output heads.

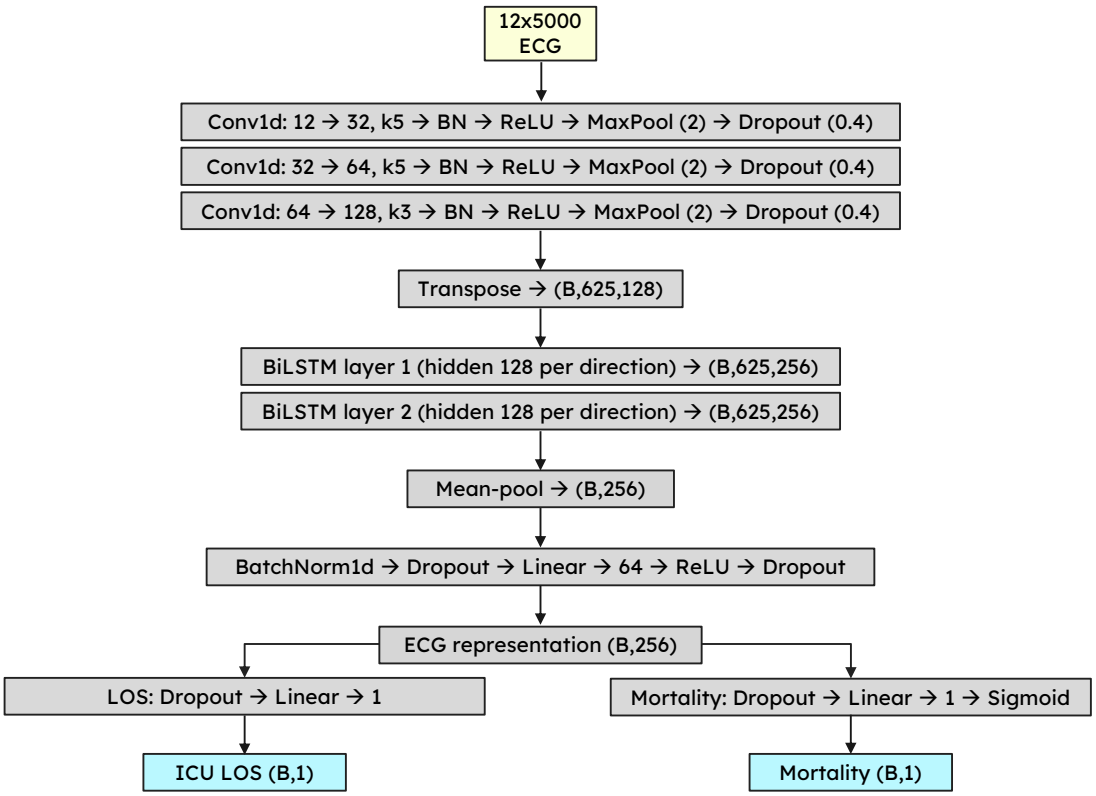

Figure 16: Schematic architecture of the `A5` hybrid CNN–LSTM model with convolutional feature extraction followed by bidirectional recurrent encoding.

This design can be interpreted as a two-stage decomposition of the ECG analysis problem. The convolutional stage first converts the raw waveform into a higher-level feature sequence, ideally emphasising morphological structures that recur within and across leads. The recurrent stage then models temporal dependencies on this transformed representation rather than on the original amplitudes. In other words, the model first asks "what local waveform patterns are present?" and then "how do these learned patterns evolve over time?" This is a more structured approach than forcing a single module to solve both

problems simultaneously. This hybrid design was selected because it explicitly combines two inductive biases that are both plausible for ECG analysis: CNN layers are effective at capturing local morphology, whereas LSTM layers are better suited to model longer temporal dependencies and rhythm. Recent ECG studies repeatedly used hybrid CNN–LSTM or CNN–BiLSTM designs for this reason [38, 50, 77]. In the present work, `A5` occupies the middle point of the architecture ladder: it is considerably more expressive than the standalone from-scratch baselines `A2`–`A4`, yet still substantially smaller and easier to train than the foundation models in `A6` and `A7`.

This intermediate position is methodologically important because `A5` combines the two principal design ideas of the smaller from-scratch models within a single network. It was included as the hybrid reference architecture between the standalone convolutional and standalone recurrent models on the one hand and the substantially larger pretrained foundation models on the other.

5.5.5 A6.1–A6.4: DeepECG-SL Variants

The `A6` models share the same downstream wrapper and differ only in the pretrained initialisation of the ECG backbone. In all four cases, an input adapter maps the 10-second ECG segment to the expected encoder format, the pretrained encoder produces a sequence of ECG features, these features are aggregated by global average pooling, and the resulting representation is passed to a shared dense block before the LOS and mortality heads are attached. The four variants therefore allow a controlled comparison of different pretrained starting points under an otherwise fixed downstream setup.

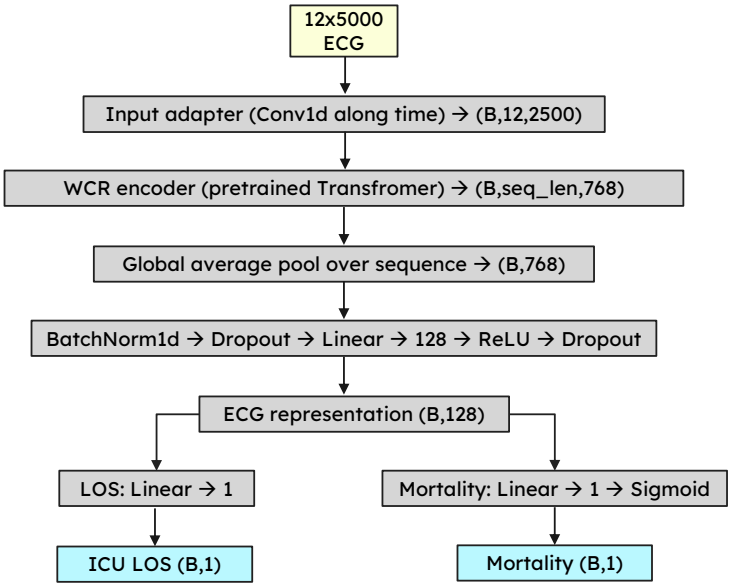

Figure 17: Schematic architecture of the `A6` DeepECG-SL model family with a shared downstream wrapper around pretrained ECG backbones.

The use of this model family, whose schematic architecture is shown in Figure 17, is grounded in the recent DeepECG literature. Nolin-Lapalme et al. introduced DeepECG-SL and DeepECG-SSL in 2026 as open-source ECG foundation models trained on more than one million ECGs and validated across multiple public and private external datasets. In the evaluation reported by Nolin-Lapalme et al., DeepECG-SL was selected as the best supervised ECG-interpretation architecture after an extensive search across transformers, CNNs, and Mamba-based models. The final supervised model was based on a modified EfficientNet-V2-S optimised for one-dimensional ECG signals and reached an AUROC of 0.992 on the internal multi-label ECG interpretation task [69]. In direct comparisons with shared diagnostic labels, the DeepECG models also showed modest but consistent superiority over ECGFounder and ECG-FM on the UK Biobank [78], the Canadian Longitudinal Study on Aging, and PTB datasets [69]. For this reason, the DeepECG line currently constitutes one of the strongest described benchmark families for transferable ECG representation learning. In the present thesis, this family was operationalised through four pretrained checkpoints from the same ecosystem. Using several checkpoints while keeping the downstream head fixed makes it possible to compare different pretraining targets under identical ICU-task conditions. This is methodologically relevant because recent reviews and benchmark studies emphasise that the downstream utility of ECG foundation models depends strongly on pretraining data, objective, and adaptation strategy [19, 68, 79].

5.5.6 A7.1–A7.3: HuBERT-ECG

The `A7` models use HuBERT-ECG backbones and share the same downstream logic while varying in model size. Internally, the 12-lead input is flattened and processed by a pretrained transformer encoder. The encoder output is mean-pooled over the sequence dimension and forwarded to a lightweight prediction head. In the experiment registry, the three variants correspond to `small`, `base`, and `large` versions.

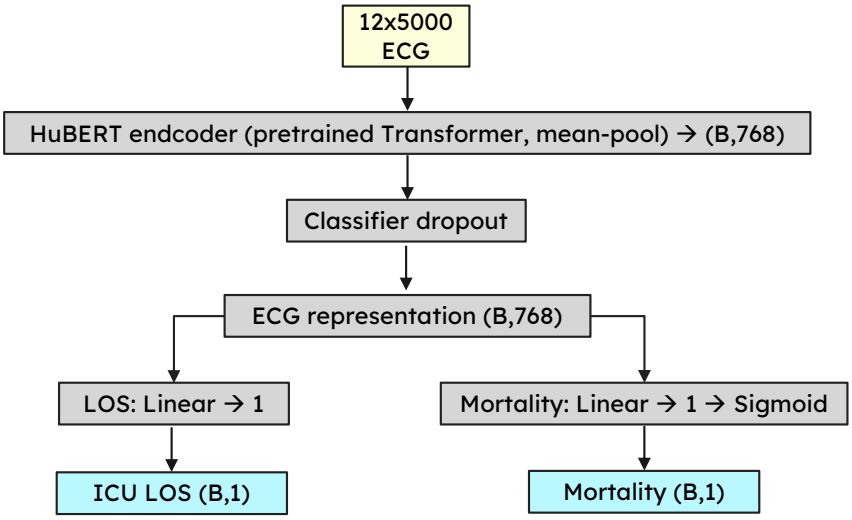

Figure 18: Schematic architecture of the `A7` HuBERT-ECG model family with a pretrained transformer backbone and lightweight prediction head.

HuBERT-ECG was included because it is one of the most prominent recent self-supervised transformer-based ECG foundation models; the schematic architecture of the `A7` model family is shown in Figure 18. Coppola et al. introduced HuBERT-ECG in 2024 as a HuBERT-inspired ECG foundation model that learns representations through feature extraction, clustering, and masked modelling. The model was pretrained on 9.1 million 12-lead ECGs from four countries and evaluated on 17 downstream scenarios, with the explicit goal of broad transferability across diagnostic and prognostic ECG tasks [41]. The original paper also released three model sizes, which makes the architecture particularly suitable for a controlled scale comparison within the same model family [41]. In the context of this thesis, HuBERT-ECG provides a second pretrained foundation-model paradigm next to the DeepECG family. This is methodologically important because it extends the transfer experiments beyond one single pretraining strategy. Recent reviews and benchmarking work emphasise that ECG foundation-model performance is strongly task-dependent and that no single model family dominates all downstream settings [19, 79]. Therefore, including HuBERT-ECG ensured that the project evaluates not only transfer learning in general, but also a concrete self-supervised transformer approach that was specifically developed for broad ECG generalisation.

Taken together, the DL model family forms an architecture ladder with increasing representational capacity. The from-scratch models `A2`–`A5` test which inductive bias is already useful on the target cohort, ranging from local convolutional structure to explicit sequence modelling and hybrid designs. The pretrained models `A6` and `A7` then address whether large-scale ECG pretraining offers additional gains in robustness, data efficiency, and transferability for the ICU prediction tasks considered in this thesis. Taken together, these models do not simply constitute a list of different architectures. Rather, they opera-

tionalise the central experimental comparison of the project: from handcrafted baseline to small end-to-end neural models, from explicitly sequential inductive biases to hybrid designs, and finally from task-specific training to large-scale ECG foundation model transfer. For this reason, section 5.5 is not merely a catalogue of models, but the methodological backbone of the thesis' main empirical study.

5.6 Multimodal Fusion

This section addresses the research question: *To what extent does integrating ECG with structured EHR and demographic data improve predictive performance compared to ECG-only models?* Methodologically, this question is not only about whether additional modalities are fused with the ECG, but also about which variables enter the model, how they are temporally aligned to the early ICU prediction window, and how the resulting multimodal cohort differs from the broader ECG-only population in sample size, coverage, and clinical composition.

5.6.1 Concept and Feature Selection

Multimodal fusion in this thesis denotes the joint use of several information sources for the same prediction task, rather than the post hoc combination of separate unimodal models. In the present project, the central modality is the 12-lead ECG, which is complemented by structured clinical variables derived from the linked EHR data. This design is enabled by the fact that MIMIC-IV provides hospital- and ICU-level clinical information, while MIMIC-IV-ECG links 10-second diagnostic ECGs to the broader clinical database through shared identifiers and timestamps [26, 71].

In the terminology used throughout this thesis, the project follows a late-fusion strategy: modalities are processed separately at first, and fusion is carried out only after the ECG backbone has produced a fixed-dimensional embedding. Operationally, the pooled ECG embedding is concatenated with vectorised tabular inputs before the shared fully connected layers and the task-specific output heads. Reviews of multimodal learning for biomedical data emphasise that the fusion stage is a key design choice and that heterogeneous modalities typically require separate preprocessing and feature extraction before they are merged [80–82]. In the present setting, this is particularly relevant because ECG waveforms and tabular clinical variables differ substantially in scale, dimensionality, and semantic meaning. Early fusion at the raw-input level would therefore impose an arbitrary shared representation too early in the pipeline. Preserving a dedicated ECG encoder allows each modality to be represented in an appropriate form before the modalities are combined for joint prediction. Compared with direct raw-input fusion, this yields a more

modular design and keeps comparisons across CNN, LSTM, hybrid, and pretrained backbones focused on the ECG representation rather than on changing the structure of the tabular pathway.

The multimodal feature set remains intentionally compact and clinically interpretable. In this thesis, the tabular extension of the ECG models was restricted to three feature blocks: demographics, ICU-unit context, and EHR. Demographics and ICU unit act as fixed tabular components whenever multimodal late fusion is enabled, whereas the EHR block is the only variable component. More specifically, the experiments compared two alternative EHR variants, denoted here as full EHR table and small EHR table. This distinction is deliberate: the full EHR variant includes bedside variables together with selected laboratory and output-derived measurements, whereas the small EHR variant excludes laboratory values and keeps only a compact bedside-focused subset that is easier to aggregate consistently. The methodological goal was therefore to compare a stable ECG-plus-context setup under two EHR variants that differ in scope and complexity at the fusion point.

The final EHR feature sets were selected from a larger pool of candidate EHR variables through an internal coverage screening on matched ECG samples. For an ECG sample i, with ICU admission time $t^{(i)}_{\text{stay\_intime}}$ and ECG acquisition time $t^{(i)}_{\text{ecg}}$, the admissible EHR window was defined as

$$\mathcal{W}_i = \left[t^{(i)}_{\text{stay\_intime}}, \min\!\left(t^{(i)}_{\text{ecg}},\, t^{(i)}_{\text{stay\_intime}} + 6\text{h}\right) \right].$$

Only measurements recorded after ICU admission, within the first six hours of the stay, and no later than the ECG timestamp were eligible for feature construction. Availability was then evaluated per variable across the cohort of matched ECG samples. A variable was retained only if it was available in at least 70 % of matched ECG samples, providing a pragmatic balance between retaining clinically informative variables and avoiding features dominated by missingness. This threshold was applied at the variable level, not at the sample level. Individual ECG samples could still contain missing values, but variables missing in most matched samples were excluded.

Variant	Included variables
Full EHR	Respiratory rate, oxygen saturation, mean arterial pressure (MAP), systolic blood pressure (SBP), diastolic blood pressure (DBP), body temperature, Glasgow Coma Scale (GCS) eye and motor subscores, platelets, creatinine, urine output, and weight
EHR-small	Respiratory rate, oxygen saturation, SBP, DBP, body temperature, and weight

Table 8: EHR feature sets used for multimodal late fusion. Detailed feature descriptions are provided in Appendix in table 12 and 13.

Compared with the full EHR block, EHR-small omits laboratory measurements, urine output, and GCS-derived items, and is therefore easier to aggregate consistently from the early ICU window.

In both variants, repeatedly measured point variables were aggregated per ECG over $\mathcal{W}_i$, mainly through the median. By contrast, urine output was summarised through the cumulative sum over the same window because it is an interval quantity rather than a point measurement. Unlike the other chart-derived variables, weight was constructed from admit- and daily-weight ITEMIDs in `chartevents`. All documented weight values inside $\mathcal{W}_i$ were then aggregated through their median.

Demographic variables, in particular age and sex, enter the model because they provide compact baseline risk context and are typically available independently of the ECG waveform itself. Prior ICU prediction work and recent multimodal clinical prediction studies commonly include such demographic variables in structured prediction models [20, 25]. The ICU care context enters through the first ICU unit, which represents the care setting and associated case mix of the matched ICU stay. This is methodologically plausible because ICU outcomes and resource use vary across units and institutional structures, so the care unit may explain part of the heterogeneity in LOS-related outcomes [83]. The EHR block comprises aggregated measurements from the early ICU window defined above. Its concrete content depends on whether a run uses the full EHR table or the reduced EHR-small table.

5.6.2 Temporal Alignment and Leakage Control

A central methodological concern in multimodal ICU prediction is information leakage. In the present thesis, leakage denotes the use of information that is not yet available at the intended prediction time. It also includes variables that reveal the target directly or indirectly because they are generated only after the intended early prediction window.

This distinction matters both during model development and during final evaluation. If future information enters the training or validation process, performance estimates become artificially optimistic because the model learns shortcuts that do not exist in a real deployment scenario. If the same happens in the held-out test set, apparent generalisation is also overstated. In EHR-linked ECG studies, typical sources of leakage are measurements and procedures whose timestamps lie outside the intended observation window, as well as administrative variables that summarise the completed hospital course rather than the state that is knowable early in the ICU stay.

Avoiding such leakage is critical for three reasons. First, it preserves clinical validity by aligning the prediction task with the actual information state around ECG acquisition rather than with knowledge accumulated much later in the hospitalisation. Second, it improves external validity, because models that exploit future-information shortcuts often break down outside the original dataset or in prospective use. Third, it is necessary for fair architecture comparison, since any difference in temporal cut-off confounds the comparison between ECG encoders more strongly than the model design itself. More broadly, the critical care ML literature repeatedly emphasises that rigorous validation and clinically coherent information cut-offs are necessary for trustworthy AI systems in high-stakes settings [12, 14].

The temporal inclusion rule follows the early-prediction information state. For demographic variables and the ICU unit, the relevant values must already be known at or very near ICU admission. For EHR features, admissibility is defined by the window $\mathcal{W}_i$ introduced above. Equivalently, for a candidate feature value x with observation time t_x, the temporal requirement is $t_x \in \mathcal{W}_i$. For static variables without a direct measurement time, such as age, sex, or the first ICU unit, the feature must be definable from information that is already available at or before the matched stay. This rule is especially important because MIMIC-IV-ECG explicitly notes that ECG timestamps are machine-recorded and do not always align perfectly with ICU or hospital events, and that some ECGs occur outside the relevant ICU or emergency-department encounter window [26]. The multimodal pipeline therefore applies the same conservative ECG-to-stay matching logic used for stay-level labels whenever non-waveform variables are attached to ECG-centred samples. Demographic features are linked at the record or patient level, ICU-unit features are linked through the matched stay, and EHR features are additionally restricted to $\mathcal{W}_i$ in order to match the intended early-risk scenario.

5.6.3 Feature Construction

The three tabular blocks differ in source and preprocessing, but all are attached to the same ECG-centred sample definition. Demographic information is the most static block.

Age and sex were taken from the MIMIC-IV patient-level tables and linked to each ECG sample through the project-specific record table. In model input form, demographics were represented as a small fixed-length vector with age as a numeric value and sex as a binary or categorical variable, depending on the respective experiment configuration.

The ICU-unit feature adds organisational context that is not contained in the ECG waveform itself. Operationally, the feature was defined as the first assigned ICU care unit of the matched stay, i.e., `first_careunit` from the MIMIC-IV `icustays` table. This variable captures part of the clinical context in which the ECG was recorded. In practice, it reflects differences in treatment focus and local care routines. Such contextual information is relevant because ICU outcomes and resource use vary between units and structures even after severity adjustment [83]. In the project pipeline, the unit enters only when the ECG can be matched to a valid ICU stay at `ecg_time`, and the resulting categorical value is encoded as a one-hot vector over a predefined `icu_unit_list`. Appendix 14 summarises the ICU-unit categories used in this encoding and provides descriptive plots for the ten most frequent units, see Table 14, Figure 33, and Figure 34. These summaries show that different ICU units are associated with different median LOS and mortality profiles. For example, median LOS is lower in the medical/surgical ICU than in Neuro Intermediate, while mortality is lower in the cardiovascular ICU than in the Neuro Surgical ICU and the Medical ICU. Because the first care unit is already known at ICU admission, its use does not introduce future-information leakage. At the same time, ICU unit is not a causal explanation of outcome. It should be interpreted as descriptive clinical context.

The EHR block is methodologically different from the demographic and ICU-unit blocks because it is composed of time-stamped observations rather than near-static admission context variables. Its inclusion therefore depends directly on the temporal window defined above. The full EHR table is aggregated from the early ICU course using three MIMIC source families: `chartevents` for monitored and charted physiological variables, `labevents` for laboratory variables, and `outputevents` for output-related quantities such as urine measurements. Stay metadata from `icustays` were used to enforce the ICU-centred time window. Detailed feature lists are provided in Appendix 12 and Appendix 13. Their construction followed the scheme introduced above: repeatedly measured charted or laboratory variables were aggregated per ECG, typically by the median, whereas output-type variables were summarised by interval sums. As a result, each ECG was associated with exactly one fixed-length EHR feature vector.

Methodologically, this allowed a controlled comparison between a broader early-EHR representation and a less restrictive variant that emphasises lower-dimensional and predominantly non-invasive bedside information. From the modelling perspective, technical integration remained deliberately simple: the ECG backbone first produces a compact waveform representation, and the active tabular features are then concatenated to this

representation in a fixed order before the shared prediction layers. In the multitask setting, this single fused representation feeds both LOS regression and mortality classification, i.e., one common fully connected trunk is followed by separate task heads rather than by a separate EHR-only pathway [30].

The integration of tabular features also has direct consequences for cohort definition. Demographic features can be attached at the record level and therefore do not by themselves require a stay-linked restriction beyond the ECG cohort definition. ICU-unit and EHR features, by contrast, require a valid ICU-stay match at the ECG timestamp. In addition, the EHR variants require sufficient eligible observations within the six-hour post-admission window and before the ECG time so that the aggregated feature vector can be assembled. ECGs without a valid stay match, with missing temporal alignment, or without sufficient eligible EHR observations were therefore excluded from the corresponding multimodal subset. As a result, models that used the EHR extension may be trained and evaluated on a more restricted population than pure ECG-only models or multimodal variants that only add demographics and ICU unit. Direct comparisons between ECG-only and multimodal variants therefore require caution, because they were not tested on the same data set. In the remainder of the thesis, this restricted population is best understood as an *early-EHR-eligible cohort*, i.e., the subset of ECGs for which multimodal feature attachment was temporally and clinically defensible.

Overall, the multimodal design of this thesis was motivated by the goal of giving the models as much relevant early information as possible. Late fusion provided a controlled and reusable framework for comparing ECG architectures under an identical non-waveform context block, while the selected tabular modalities added information at different levels of abstraction: demographics contribute baseline risk, ICU-unit indicators capture part of the organisational and case-mix context, and the EHR block contributes early physiological context from the ICU stay. This is consistent with recent review and benchmark literature on multimodal clinical prediction in ICU and related settings, which likewise combined heterogeneous but temporally anchored data sources instead of treating clinical prediction as a strictly single-modality task [7, 12, 14]. At the same time, the multimodal feature design remained focused on information that was broadly available and clinically plausible within the early ICU window. The full EHR variant already covered a substantial part of the routinely documented information available for most patients, while demographics and ICU unit added stable contextual signals. This kept the multimodal extension tightly coupled to the ECG prediction setting while preserving a clear experimental distinction between fixed context features (demographics and ICU unit) and variable EHR feature alternatives.

5.7 Training Protocol and Hyperparameter Optimisation

Model development followed a staged decision pipeline. The candidate DL model families were first evaluated under a common baseline configuration in order to compare architectures, feature sets, and fusion variants under controlled conditions. After this baseline stage, selected model families entered a limited hyperparameter refinement stage with predefined architecture-specific settings (`H2`–`H4`). This intermediate step was used to test whether the main model families reacted to plausible changes in learning rate, regularisation, patience, and loss formulation. It should be distinguished from the final automated Optuna search reported in Section 6.4, which was carried out only for the retained multimodal hybrid CNN–LSTM line. Threshold selection for mortality was a separate validation-based decision step and was not treated as model hyperparameter optimisation. This staged separation was mainly a feasibility choice, since tuning every architecture and every feature variant would not have been computationally practical. Table 9 summarises the reference baseline profile denoted by `H1`, which defined the initial comparison regime across model families while allowing family-specific implementations where required by the respective architecture.

5.7.1 Fixed baseline configuration and shared training objective

All deep learning experiments used multitask training, so LOS and ICU mortality were optimised within one shared training run. The model therefore exposed one regression head for LOS and one binary head for mortality, and the scalar training objective was defined as

$$\mathcal{L}_{\text{total}} = w_{\text{los}}\,\mathcal{L}_{\text{los}} + w_{\text{mort}}\,\mathcal{L}_{\text{mort}}.$$

$\mathcal{L}_{\text{total}}$ denotes the total training loss, $\mathcal{L}_{\text{los}}$ the LOS loss, $\mathcal{L}_{\text{mort}}$ the mortality loss, and w_{los} and w_{mort} the corresponding task weights. This follows the usual multitask perspective in clinical time-series modelling, where related endpoints can share representations and thereby improve data efficiency relative to isolated single-task training [21, 30]. In the fixed baseline configuration, both task weights were set to 1.0, such that LOS prediction and mortality prediction contributed equally to the multitask objective. Mortality was optimised using binary cross-entropy, whereas LOS was optimised using mean squared error, consistent with the `H1` reference setting summarised below. ECG recordings were matched to ICU stays using patient identifiers and recording timestamps. Records without a valid ICU-stay match were excluded from the final model datasets. For matched records, LOS was obtained from the ICU stay table, and the mortality target was derived from admission death times occurring within the corresponding ICU stay interval.

Hyperparameter	`H1` baseline setting
Initial learning rate	5.0×10^{-4}
Batch size	64
Weight decay	1.0×10^{-4}
Dropout rate	0.3
Gradient clip norm	1.0
Scheduler patience	5
Early stopping patience	7
Adam betas	$[0.9, 0.999]$
Scheduler factor	0.1
Minimum learning rate	1.0×10^{-6}
Augmentation noise std	0.03
LOS loss	MSE
Mortality loss	binary cross-entropy
Optimiser	Adam

Table 9: Reference baseline hyperparameter profile (`H1`).

The fixed baseline regime also defines the initial optimisation, regularisation, and stopping logic at a high level. Training used adaptive first-order optimisation [33], validation once per epoch, checkpoint selection by minimum validation total loss $\mathcal{L}_{\text{total}}$, and early stopping based on `val_loss`. Dropout, weight decay, and global gradient clipping served as the main regularisation mechanisms, while mixed-precision training and label smoothing were not part of the implemented default pipeline. Training was seeded (typically with 42) for Python, `NumPy`, and PyTorch, but full GPU-level determinism was not enforced because deterministic backend flags remained disabled for performance reasons.

5.7.2 Controlled training protocol and automated tuning

The fixed baseline configuration served as the controlled starting point for model development. Keeping this setting fixed made the early architecture and feature comparisons easier to interpret before additional tuning was introduced.

The later tuning steps had two different scopes. The predefined `H2`–`H4` settings were small, controlled refinements used during cross-architecture screening. The Optuna studies were the final automated hyperparameter optimisation and were restricted to the multimodal hybrid CNN–LSTM configuration selected for the last stage of the workflow.

Training followed the same general loop across experiments: model parameters were updated on the training split, performance was monitored once per epoch on the validation split, and checkpoint selection was based on validation performance under the multitask

objective. Early stopping was applied through the validation loss, so that training ended when additional epochs no longer improved the retained selection criterion. For mortality, threshold-independent metrics and threshold-based decision metrics were tracked separately. Where a binary decision was required, the threshold was chosen on the validation split and then transferred unchanged to the test split.

Automated hyperparameter optimisation was carried out with Optuna [84]. Each Optuna trial launched one training run under the standard validation loop, although trials could also be stopped early by pruning. The study direction was minimisation, and the optimisation target was the validation total loss $\mathcal{L}_{\text{total}}$ defined above for the respective study configuration. At a high level, the search spaces covered optimisation parameters, regularisation settings, augmentation strength, training budgets, and selected architecture-related capacity choices.

5.8 Overview, Metric Definitions and Preanalysis

This subsection introduces the notation and metric definitions used in this work. In particular, it defines the reported measures for LOS prediction, binary in-ICU mortality, the subgroup analysis for observed stays with LOS ≤ 10 days, and a short exploratory preanalysis of augmentation and amplitude handling.

5.8.1 Notation and reported outcomes

Unless noted otherwise, all reported results were computed on the held-out test split. Model training and inference were performed at the level of individual ECG windows, whereas LOS and mortality were defined at the level of the ICU stay. All ECG windows linked to the same stay therefore shared the same target label. For LOS, y_i denotes the observed LOS in days for ICU stay i, and $\hat{y}_i$ the corresponding stay-level prediction. If several ECG windows were available for one stay, the model produced several predictions for the same outcome. For validation and test metrics, these predictions were averaged into one final stay-level prediction so that each ICU stay contributed exactly once and stays with many ECGs did not disproportionately influence the reported metrics. Accordingly, N always denotes the number of evaluated ICU stays. For the LOS metrics below, the prediction error was written as $e_i = \hat{y}_i - y_i$, and $\bar{y}$ denotes the mean observed LOS in the evaluated cohort.

For mortality, $y_i \in \{0, 1\}$ denotes the observed class label and $\hat{p}_i$ the predicted risk. For threshold-based metrics, a threshold τ is chosen on the validation split and then applied unchanged to the test split. This yields the predicted class label $\hat{y}_i$, from which true positives (TP) , false positives (FP) , true negatives (TN) , and false negatives (FN) are derived.

5.8.2 Metric definitions

For LOS, mean absolute error (MAE) [28, 85] is reported as the main regression metric. It is expressed in days and therefore gives a directly interpretable average absolute deviation between prediction and observed LOS [85]:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |e_i|.$$

Here, N denotes the number of evaluated ICU stays and e_i the stay-level prediction error for stay i.

For in-ICU mortality, the main threshold-independent metric is the area under the receiver operating characteristic curve (AUC) . It summarises discrimination across all possible thresholds and can be widely reported for ICU prediction tasks [21, 27]:

$$\mathrm{AUC} = P(\hat{p}^+ > \hat{p}^-).$$

Here, $P(\cdot)$ denotes probability, $\hat{p}^+$ the predicted risk for a randomly chosen positive stay, and $\hat{p}^-$ the predicted risk for a randomly chosen negative stay.

Because AUC does not describe the final operating point of a thresholded classifier, it is complemented by **precision**, **recall**, and **F1** [21, 28]. Precision quantifies which share of predicted positive cases are true events [28]:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}.$$

Recall quantifies which share of the true positive cases is detected at the chosen threshold [28]:

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}.$$

Finally, $F1$ summarises precision and recall in a single threshold-based score and is especially useful when both false positives and false negatives are relevant under class imbalance [21]:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

5.8.3 Subgroup with observed Length of Stay ≤ 10 days

Because LOS was right-skewed as shown in Figure 9, aggregate errors on the full cohort mixed short and long stays and can therefore be dominated by a comparatively small number of very long admissions. To describe model behaviour on a clinically common range more explicitly, LOS results were also reported for the subgroup with observed LOS $y_i \leq 10$ days. The restriction was applied to the *ground-truth* LOS and not to the prediction [85].

The same definition of MAE was used on this subgroup, but only for stays satisfying $y_i \leq 10$ days. The corresponding subgroup size $N_{\leq 10}$ is reported alongside these values. These results are interpreted as descriptive subgroup analyses rather than as separate confirmatory endpoints.

5.8.4 Preanalysis

Before the main cross-architecture comparison, several exploratory CNN-based checks were conducted to probe preprocessing and training choices. One comparison examined whether retaining absolute amplitude information might help the downstream tasks. Both settings used gaussian noise augmentation with random noise at 3 % of the signal standard deviation. The comparison was therefore between `P1`+`AG1`, that is, per-lead z-score normalisation with gaussian noise only, and `P2`+`AG2`, that is, no per-lead z-score normalisation with the same gaussian noise step plus amplitude scaling between 0.85 and 1.15. In this comparison, LOS MAE increased from 2.67 under `P1`+`AG1` to 2.74 under `P2`+`AG2`, so amplitude scaling was not retained and the normalised `P1` pipeline remained the standard preprocessing setup for the later experiments. A separate augmentation check compared one run without augmentation (`AG0`) against one run with gaussian noise only (`AG1`). This reduced LOS MAE from 2.71 to 2.67, corresponding to an improvement of about 2 %. An additional exploratory run then combined random time shuffling with oversampling of the mortality classes in order to test a more aggressive response to class imbalance. This setup corresponds conceptually to the later `AG5` bundle, that is, the `AG4` signal perturbations plus oversampling during training. However, relative to the standard setup, AG5 reduced LOS MAE by only 0.0039 and lowered mortality AUC by 0.0772, so oversampling was not carried forward as a core method. Because these checks were confined to the CNN-from-scratch baseline, they are reported here as exploratory preliminary checks rather than as model-family results. Detailed MAE and AUC plots are provided in Appendix A.2.1 and Appendix A.2.1.

5.9 Model Development Workflow and Progressive Narrowing

Figure 19 provides a compact overview of the overall development process followed in this thesis. It summarises the progression from broad baseline screening and exploratory preanalysis in Phase 1 to focused late-fusion development and final configuration work in Phase 2. The overview thereby serves as the structural reference for the workflow described in the remainder of this section.

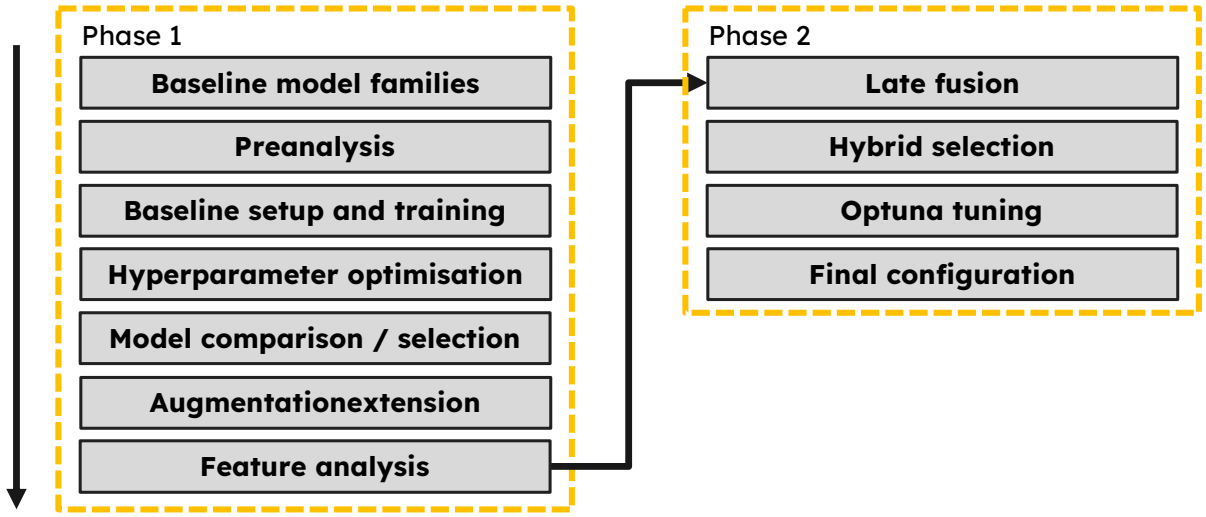

Figure 19: Workflow of model development across Phase 1 and Phase 2.

5.9.1 Phase 1: Broad Screening and Structured Narrowing

Phase 1 of the workflow was designed as a broad screening and structuring stage. The starting point was an intentionally diverse model space that combined a classical ML reference model with several ECG-only DL families, namely XGBoost, a CNN trained from scratch, a unidirectional LSTM, a bidirectional LSTM, a hybrid CNN–LSTM, and two foundation-model lines represented by DeepECG-SL and HuBERT-ECG [41, 69]. The purpose of this stage was not yet to optimise any single architecture in depth, but to establish a controlled comparison framework in which different representation principles could be examined under the same baseline conditions. In practice, this meant beginning with a broad architecture set and delaying more extensive tuning until a clearer picture had emerged regarding stability, feasibility, and general methodological suitability.

Before the shared baseline comparison was carried out, the project also included a general preanalysis stage. This preanalysis consisted of targeted exploratory tests that addressed practical questions relevant to the later workflow. First, a dedicated mortality subset with an approximately balanced label distribution was created in order to test with the CNN-from-scratch baseline whether the model was fundamentally able to learn mortality-related ECG changes under less severe class imbalance. Second, a targeted comparison

across P1 versus P2 and AG1 versus AG2 was used to examine whether retaining absolute amplitude information might carry relevant signal for the downstream tasks. Third, an additional check was performed with AG0, that is, without ECG augmentation, in order to judge whether augmentation contributed useful variation beyond the default training pipeline. These steps did not yet serve as the formal experimental comparison. Rather, they functioned as early feasibility checks that helped determine whether the later stages of the workflow were methodologically worthwhile.

After this exploratory stage, the model comparison was aligned to one shared baseline configuration, namely D1, P1, AG1, and H1. All baseline architectures were then trained under the same general conditions so that differences in validation behaviour could be interpreted primarily as consequences of model family and input design rather than as consequences of different preprocessing or training settings. The same phase also included the implementation of additional augmentation variants and a parallel examination of structured feature groups such as demographics, ICU-unit information, and EHR features. In this way, architectural screening and feature-related preparation progressed together instead of as two completely separate work streams. From the unidirectional LSTM onward, the workflow also included comparisons across multiple hyperparameter settings, since the more complex DL models were sufficiently expressive to justify an initial exploration of training sensitivity before later conclusions were drawn about which families should be pursued more systematically.

As shown in Figure 19, this first phase gradually narrowed the search space. More extensive tuning effort was reserved for architectures that appeared sufficiently promising in terms of validation behaviour, training stability, computational cost, and practical compatibility with the broader multimodal setup. A central part of this narrowing process was the explicit comparison between sequence-oriented and transformer-based DL approaches, since the project aim was not merely to benchmark isolated models, but to determine which family of ECG representations should be carried forward into the later multimodal stage. Within the DL baselines, this narrowing process reduced the initial architecture set to a smaller group of candidates for later-stage work. Within the foundation-model branch, DeepECG-SL and HuBERT-ECG were compared under frozen and unfrozen backbone settings and across three model-size or weight variants. In this comparison, the largest HuBERT-ECG configuration could not be pursued further in practice because training exceeded the available runtime budget and hit the project timeout after approximately 12 hours, which made it too large for the intended workflow. Taken together, these comparisons reduced the search space and defined the model lines that were carried forward into the multimodal stage.

5.9.2 Phase 2: Late Fusion and Final Selection

Phase 2 followed the same logic of progressive focusing, but now within the multimodal setting. After the initial screening phase, the workflow shifted from broad architecture coverage towards structured late-fusion experiments that combined the ECG backbone with demographics, ICU-unit information, and EHR features. In other words, the second phase was no longer concerned with exploring the full original model space, but with developing a smaller number of technically plausible and computationally manageable multimodal configurations in greater depth.

Within this phase, both the hybrid CNN–LSTM architecture and DeepECG-SL became central carriers of the late-fusion workflow. Accordingly, both model lines were trained with the structured feature combinations built from demographics, ICU-unit information, and EHR inputs. This was particularly relevant once the workflow moved beyond ECG-only inputs and into settings in which the architectures had to remain compatible with repeated multimodal runs, feature comparisons, and focused hyperparameter search. Accordingly, the later tuning stage concentrated on these two promising lines together with the selected structured feature combinations. Hyperparameter optimisation in this phase was conducted with Optuna [84], which provided the systematic search framework for the later development steps.

The workflow therefore culminated in a focused final modelling path rather than in an ever-expanding architecture search. As indicated by Figure 19, the project moved from general screening, feasibility checks, and shared baseline comparisons towards a final multimodal configuration stage in which the remaining architectural candidates were studied together with the selected demographic, ICU-unit, and EHR inputs. This section documents that sequence of decisions and the resulting reduction in scope. The quantitative evidence that justifies the resulting focus is presented in Chapter 6.

6 Results

This chapter presents the empirical results of the experimental workflow described in Chapter 5. All reported findings are based on the temporally stratified train, validation, and test splits introduced earlier, with model selection performed on validation and final reporting reserved for the held-out test data. Evaluation notation, metrics, and the exploratory preanalysis are defined in Chapter 5. The chapter first reports LOS and in-ICU mortality results and then turns to multimodal fusion and hyperparameter optimisation.

6.1 Length of Stay Prediction

This subsection reports LOS regression error on the held-out test split. The section begins with cohort-level MAE under a shared baseline configuration (`H1`) and then compares these results with the selected hyperparameter-test settings from `H2`–`H4`, while `D1`/`P1`/`AG1` remained fixed. Agreement between predicted and observed LOS is then illustrated with Bland–Altman plots for one selected example model, the hybrid CNN–LSTM run `A5-R17`. This plot serves only as an example of prediction agreement across the LOS range, whereas the cohort-level MAE comparison is reported across all retained models. Because LOS is heavily skewed with a long tail of very long admissions (Section 4), MAE is also reported on the subgroup with observed LOS $y_i \leq 10$ days (Section 5.8.3), again under baseline and selected `H2`–`H4` hyperparameter-test settings, using the same registry runs as the cohort-level panels.

6.1.1 Full Test Cohort

For the full test cohort, the initial comparison keeps the dataset, preprocessing, and augmentation bundle fixed at `D1`/`P1`/`AG1` and varies only the model family and hyperparameter setting. All runs were trained in the same multitask framework, but the figures in this subsection report only the LOS regression head through test MAE. Across the violin plots in this subsection, each architecture is represented by its own colour.

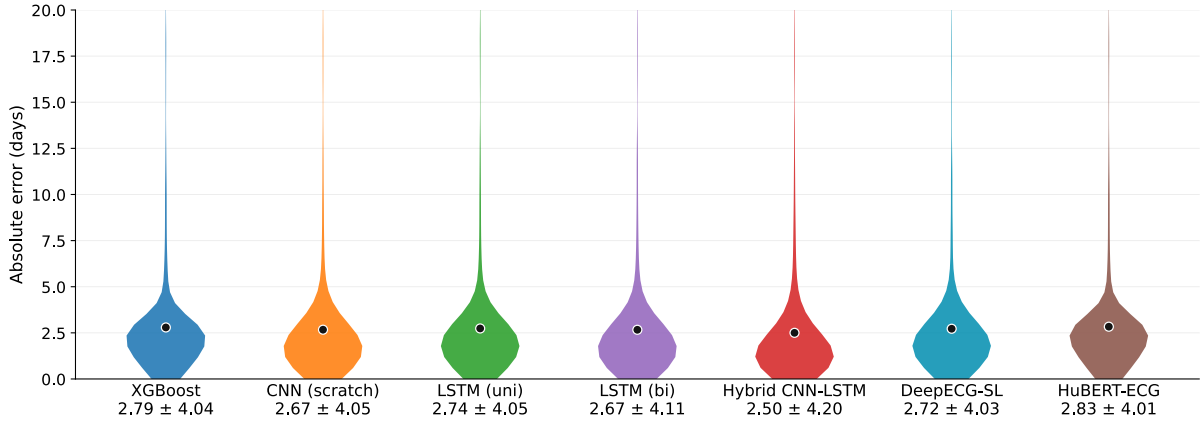

Figure 20: Full test cohort LOS absolute-error violin plots for all baseline architectures trained under the shared setting `D1/P1/AG1/H1`. Black dots indicate test MAE in days, and the x-axis labels report MAE ± standard deviation. Runs: `A1-R1, A2-R2, A3-R6, A4-R10, A5-R14, A6.1-R29, A7.1-R45`.

Figure 20 presents the baseline comparison. All baseline distributions are right-skewed on the shared 0–20 day scale. Under the common `H1` setting, the annotated MAE ranges from 2.50 days for the hybrid CNN–LSTM (`A5-R14`) to 2.83 days for HuBERT-ECG (`A7.1-R45`). CNN-from-scratch (`A2-R2`) and the bidirectional LSTM (`A4-R10`) both show 2.67 days, followed by DeepECG-SL (`A6.1-R29`) with 2.72 days, the unidirectional LSTM (`A3-R6`) with 2.74 days, and XGBoost (`A1-R1`) with 2.79 days. The reported standard deviations range from 4.01 to 4.20 days.

Starting from this baseline, a limited set of hyperparameter tests was then run within the subset of neural architectures that was carried forward in the workflow described in Section 5.9, while `D1/P1/AG1` remained fixed. This continuation did not follow the baseline MAE alone. Instead, testing effort was concentrated on the recurrent, hybrid, and foundation-model lines because they combined competitive baseline performance with greater model capacity and a clearer expectation of further gains under targeted hyperparameter refinement. Their continued use was also supported by the existing ECG literature, in which LSTM-based sequence models remain a standard reference for physiological time-series [30, 31, 36], hybrid CNN–LSTM architectures are well established for ECG modelling [38, 50, 77], and recent foundation-model work identified DeepECG-SL and HuBERT-ECG as strong large-scale ECG backbones [41, 69]. The next figure therefore no longer includes XGBoost or the CNN-from-scratch baseline and instead compares selected `H2`–`H4` hyperparameter-test runs for the retained neural architectures.

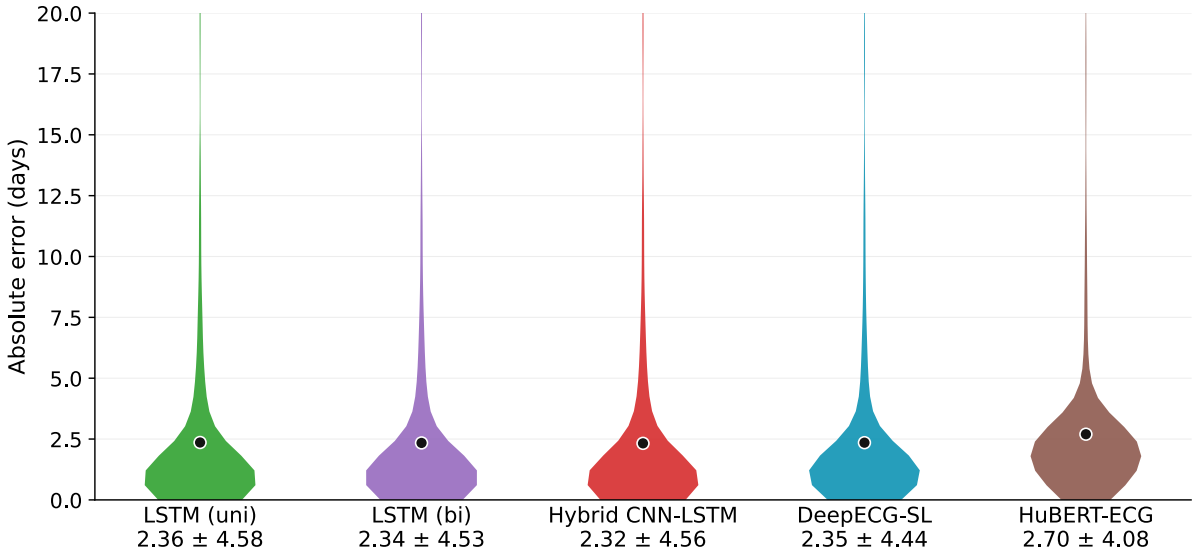

Figure 21: Full test cohort LOS absolute-error violin plots for the retained neural architectures after limited hyperparameter testing under fixed `D1`/`P1`/`AG1`. Black dots indicate test MAE in days, and the x-axis labels report MAE ± standard deviation. Runs: `A3-R9`, `A4-R13`, `A5-R17`, `A6.1-R32`, `A7.2-R47`.

Figure 21 shows that these selected hyperparameter-test runs lowered the reported MAE values relative to the baseline panel while the error distributions remained right-skewed. Within this comparison, the hybrid CNN–LSTM (`A5-R17`) shows the lowest MAE of 2.32 days. The next lowest MAEs are observed for the bidirectional LSTM (`A4-R13`), DeepECG-SL (`A6.1-R32`), and the unidirectional LSTM (`A3-R9`), with 2.34, 2.35, and 2.36 days, respectively. HuBERT-ECG (`A7.2-R47`) shows the largest MAE in this comparison with 2.70 days. The corresponding standard deviations range from 4.08 to 4.58 days.

Taken together, Figures 20 and 21 report the full test cohort absolute-error distributions before and after this limited `H2`–`H4` hyperparameter testing. These runs are not the final tuning stage. The automated Optuna optimisation is reported in Section 6.4. The next step applies the same comparison to the subgroup of stays with observed LOS $y_i \leq 10$ days.

6.1.2 Observed Length of Stay ≤ 10 days

After the full test cohort evaluation, the same selected hyperparameter-test runs were also examined on the subgroup with observed LOS $y_i \leq 10$ days. Figure 22 first illustrates this change with Bland–Altman plots for the selected hybrid CNN–LSTM run `A5-R17`, shown once on the full test cohort and once on the restricted subgroup. The hybrid run is used here only as an example to visualise how the error range changes under the subgroup restriction. Because both panels use the same axis ranges, the difference between the two views can be read directly. On the full test cohort, the mean difference is -1.611 days and

the limits of agreement are −11.405 and 8.183 days. On the subgroup with observed LOS $y_i \leq 10$ days, the mean difference is −0.598 days and the limits of agreement narrow to −4.487 and 3.291 days. The subgroup panel therefore occupies a visibly smaller vertical range while remaining directly comparable to the full test cohort view. Figure 23 reports the corresponding cross-model MAE comparison on the same subgroup.

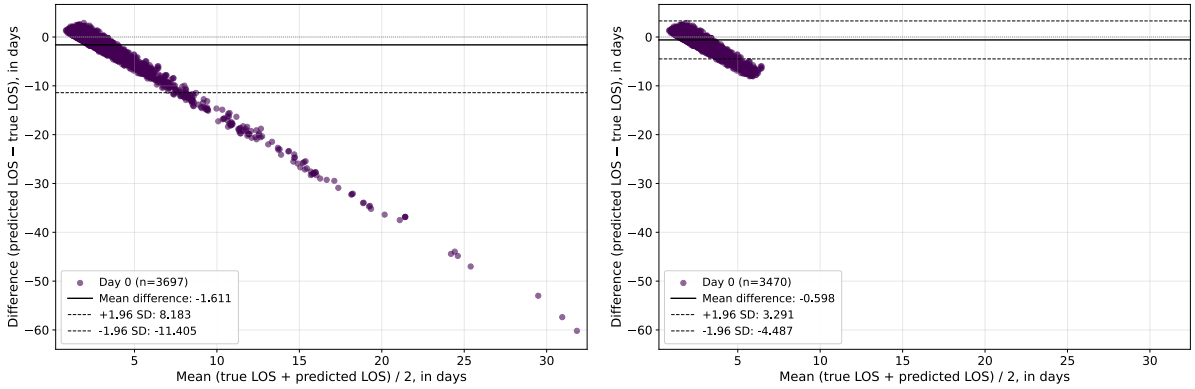

Figure 22: Bland–Altman plots for hybrid CNN–LSTM on the test split. *Left:* full test cohort. *Right:* subgroup with observed LOS $y_i \leq 10$ days. Both panels show `A5-R17`. The solid line marks the mean difference, and the dashed lines mark the limits of agreement defined as mean difference ±1.96 standard deviations.

Figure 23 presents absolute-error distributions and subgroup MAE on this restricted subgroup under the same `D1/P1/AG1` bundle with the selected hyperparameter-test settings from `H2`–`H4`.

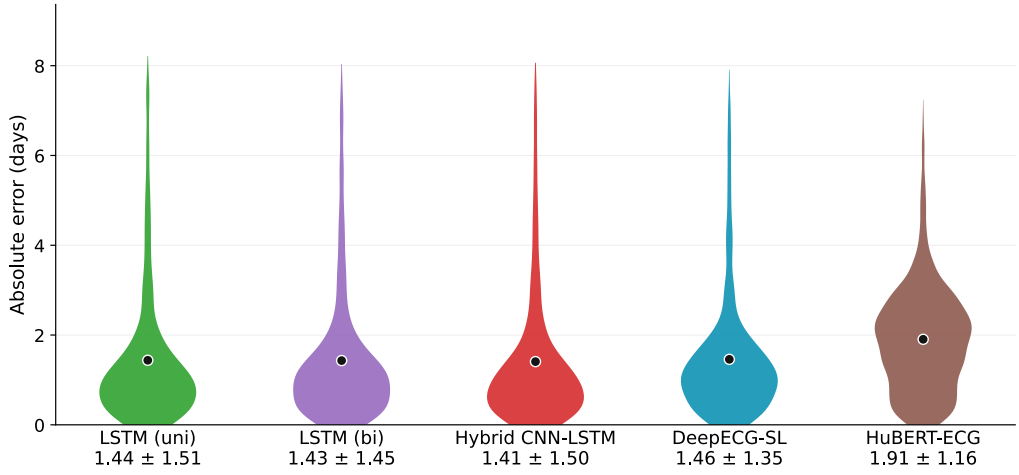

Figure 23: LOS absolute-error violin plots on the test split restricted to stays with observed LOS $y_i \leq 10$ days under `D1/P1/AG1`. Runs shown: `A3-R9`, `A4-R13`, `A5-R17`, `A6.1-R35`, `A7.2-R47`. Violin width indicates the local density of absolute errors. The black marker indicates subgroup MAE in days, and the x-axis labels report MAE ± standard deviation for the subgroup.

Figure 23 shows that the selected hyperparameter-test runs have similar subgroup MAE values on the $\leq$ 10-day subset. The hybrid CNN–LSTM yields the lowest subgroup MAE at 1.41 days, followed by the bidirectional LSTM at 1.43 days, the unidirectional LSTM at 1.44 days, and DeepECG-SL at 1.46 days. HuBERT-ECG is separated from this group at 1.91 days. The violin plots show how the absolute-error is concentrated at lower values for all models except for the HuBERT-ECG. Here, the error distribution is shifted towards higher absolute errors. The reported standard deviations provide supporting context and remain of a similar order for the four lowest-MAE runs, ranging from 1.35 to 1.51 days. HuBERT-ECG shows a standard deviation of 1.16 days together with a subgroup MAE of 1.91 days.

6.2 Mortality Prediction

This subsection reports mortality prediction performance on the held-out test split. The presentation follows two steps. First, a shared baseline setting was used to compare discrimination across model families by AUC. Second, the retained neural models were re-evaluated with the selected `H4` hyperparameter settings. For mortality, the classification threshold was then selected separately on the validation split to maximise the $F1$ score.

6.2.1 Baseline before Threshold Adjustment

Figure 24 shows the baseline ROC curves under the shared `H1` setting. The curves summarise sensitivity and specificity across possible thresholds. The baseline models achieved AUC values between 0.469 and 0.721. XGBoost (`A1-R1`) reached the highest AUC, followed by DeepECG-SL (`A6.1-R29`) and the hybrid CNN–LSTM (`A5-R14`). HuBERT-ECG (`A7.1-R45`) showed the lowest AUC in this baseline comparison. At the fixed default threshold of 0.5, however, the corresponding $F1$ score was effectively zero across all models. In practice, the baseline operating point therefore classified almost all stays as survivors.

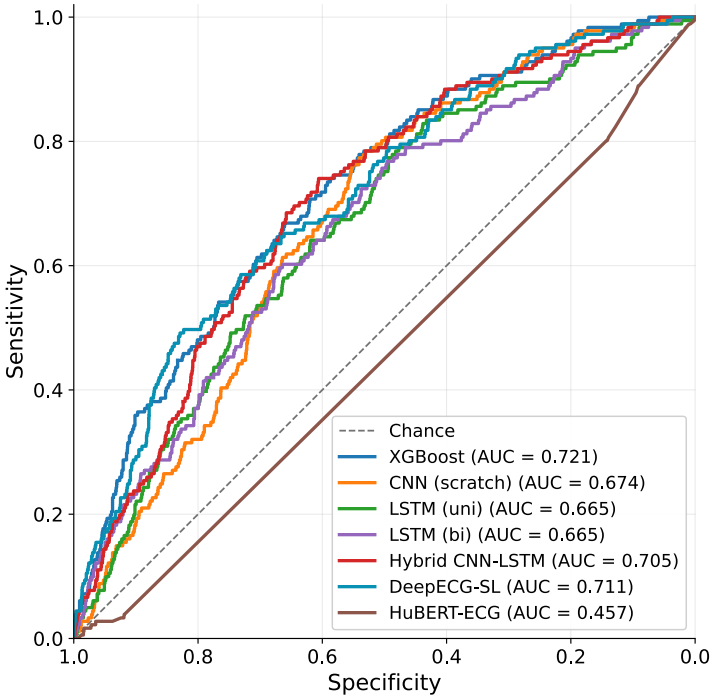

Figure 24: Baseline mortality ROC curves on the test split under `D1`/`P1`/`AG1`/`H1`. The dashed diagonal indicates chance performance. Runs shown, in legend order: XGBoost (`R1`), CNN from scratch (`R2`), Uni-LSTM (`R6`), Bi-LSTM (`R10`), hybrid CNN–LSTM (`R14`), DeepECG-SL (`R29`), and HuBERT-ECG (`R45`).

Because the default threshold produced almost no mortality-positive predictions, an additional CNN-from-scratch run was evaluated on `D7`, a mortality-balanced subset of the same task. This run was not part of the main benchmark, but it tested whether the absence of positive classifications was mainly related to class imbalance. On this balanced subset, the model reached an AUC of 0.715 and an $F1$ score of 0.7099 with an $F1$-optimised threshold. This result motivated the threshold adjustment on the original imbalanced cohort, where the fixed default threshold of 0.5 did not provide a useful operating point.

6.2.2 `H4` after Threshold Adjustment

In the `H4` settings, the classification threshold was adjusted and selected by $F1$ optimisation. The threshold determines from which predicted risk onwards a case is classified as mortality positive. In other words, the model still outputs a risk score, however the threshold defines where the decision boundary between negative and positive predictions is placed. For `H4`, this boundary was moved from the default operating point to the value $t^* = \arg\max_t F1(t)$. This means that some cases were classified as mortality-positive that would still have been counted as negative under the original threshold. Importantly, the model itself did not change its score computation. Only the interpretation of these scores

changed through the new threshold.

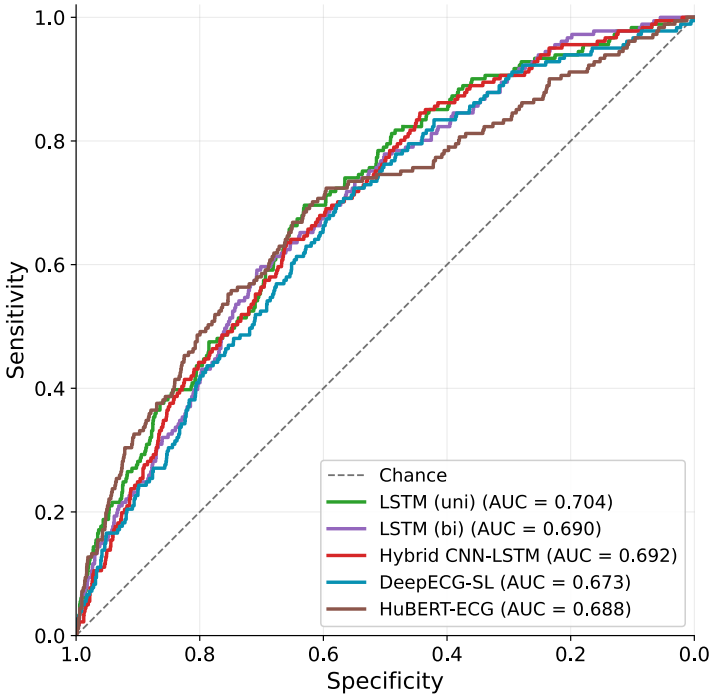

Figure 25: Mortality ROC curves on the test split under `D1`/`P1`/`AG1` with the selected tuned settings. The dashed diagonal indicates chance performance. Runs shown, in legend order: Uni-LSTM (`R9`), Bi-LSTM (`R13`), hybrid CNN–LSTM (`R17`), DeepECG-SL (`R32`), and HuBERT-ECG (`R47`).

Figure 25 shows the ROC curves for the retained neural models after tuning. The unidirectional LSTM (`A3-R9`) reached an AUC of 0.704. The hybrid CNN–LSTM (`A5-R17`), bidirectional LSTM (`A4-R13`), and HuBERT-ECG (`A7.2-R47`) followed with AUC values of 0.692, 0.690, and 0.688, respectively. DeepECG-SL (`A6.1-R32`) reached an AUC of 0.673.

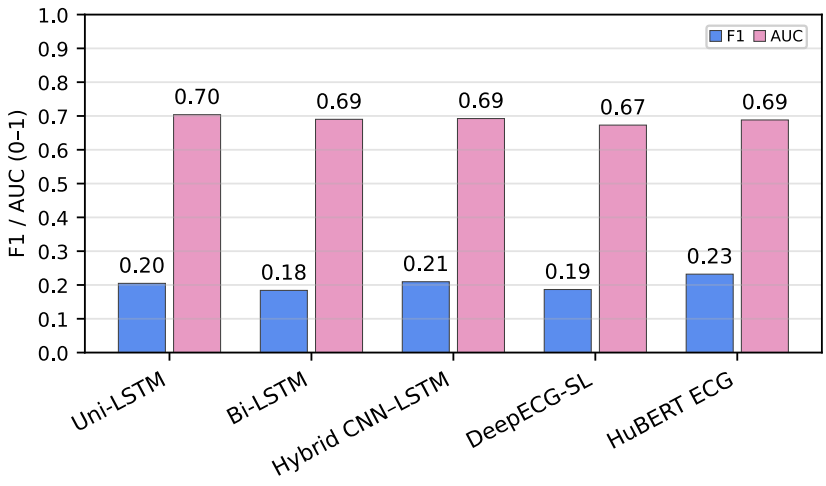

Figure 26: Mortality AUC and $F1$ score on the test split under `D1/P1/AG1` with the selected tuned settings and the $F1$-optimised threshold. Runs shown: `R9, R13, R17, R32, R47`.

Figure 26 reports the corresponding AUC and threshold-dependent $F1$ values. With the optimised thresholds, all retained models produced non-zero $F1$ scores. The reported $F1$ values ranged from 0.18 for the bidirectional LSTM (`A4-R13`) to 0.23 for HuBERT-ECG (`A7.2-R47`). The unidirectional LSTM, hybrid CNN–LSTM, and DeepECG-SL reached $F1$ scores of 0.20, 0.21, and 0.19, respectively.

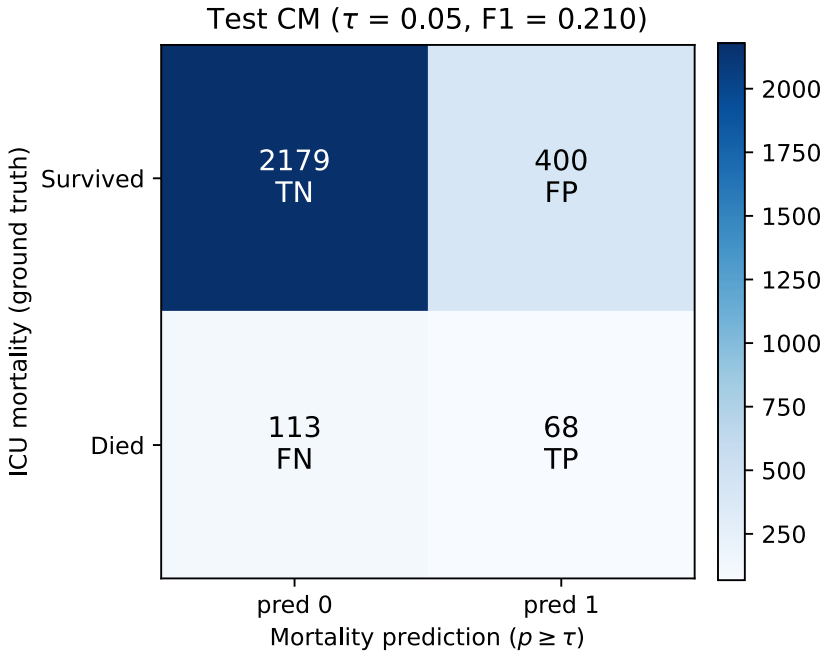

Figure 27: Confusion matrix for the hybrid CNN–LSTM mortality model on the test split. Predictions use the validation-selected mortality threshold $\tau = 0.05$. Run shown: `R17`.

Figure 27 illustrates this for the hybrid CNN–LSTM. The model outputs a mortality probability p. Stays with $p \geq \tau$ are classified as deaths and those with $p < \tau$ as survivors. At $\tau = 0.05$, chosen on the validation set to maximise $F1$, most survivors are classified correctly (2179), while mortality remains harder to detect, with 68 true positives and 113 false negatives. The test-set $F1$ at this threshold is 0.210, and the matrix also shows 400 false positives.

6.3 Multimodal Fusion

This subsection investigates whether adding structured clinical information to the ECG embedding improves prediction compared with ECG-only models. The multimodal comparison was restricted to the hybrid CNN–LSTM and DeepECG lines. This choice followed the preceding model comparison, where both lines remained promising candidates, while the hybrid CNN–LSTM also offered a comparatively efficient parameter profile for repeated feature experiments. This selection was additionally supported by current literature, as hybrid CNN–LSTM architectures remain a well-established design for ECG modelling because they combine local morphology extraction with sequential modelling of temporal dependencies [36, 50, 77]. At the same time, recent ECG foundation-model literature placed DeepECG among the current high-capacity ECG backbones and highlighted the broader promise of pretrained ECG encoders for transfer across tasks and datasets [41, 68, 69].

The dataset identifiers `D1`–`D6` define the late-fusion feature combinations and are listed in Section 5.3. Briefly, `D1` denotes the ECG-only reference, `D2` and `D3` add demographics or ICU unit separately, `D4` combines both context blocks, and `D5` and `D6` additionally include either the full EHR or the reduced EHR-small table. Across this subsection, the experimental setup was kept fixed apart from the dataset ID. In addition, augmentation was extended from `AG1` to `AG4`, adding Gaussian noise, lead dropout, and baseline wander perturbations. This choice is in line with recent ECG augmentation literature, which describes noise-based, scaling-based, lead-level, and baseline-shift transformations as plausible signal-space perturbations for more robust representation learning [74, 86].

6.3.1 Feature Development across `D1`–`D6`

The first step examines how LOS MAE changes when the feature bundle is extended from `D1` to `D6` while the remaining setup is fixed.

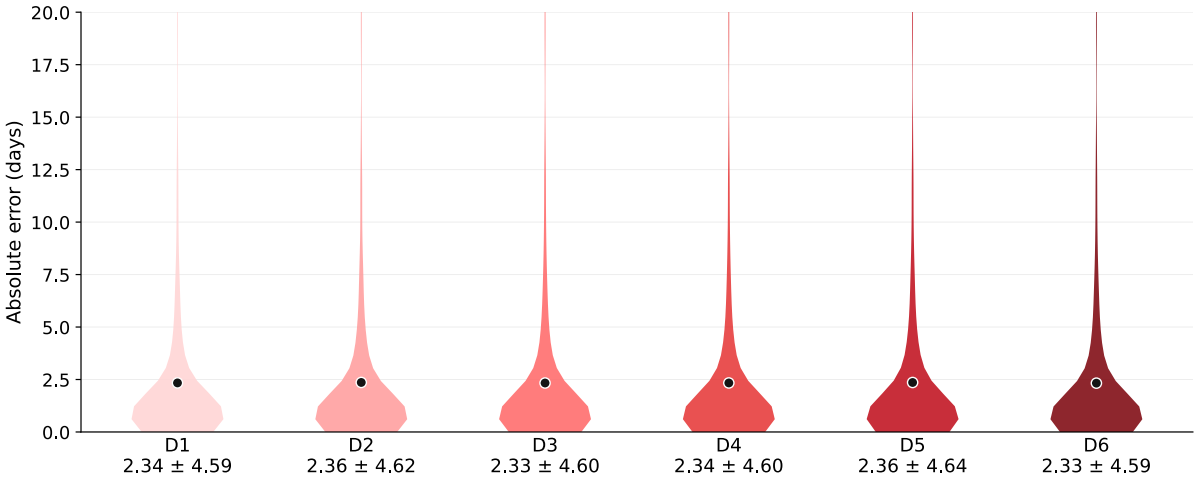

Figure 28: hybrid CNN–LSTM LOS absolute-error violin plots across dataset variants `D1`–`D6`. The black dots indicate MAE in days, and the x-axis labels report MAE ± standard deviation.

Figure 28 shows that the hybrid CNN–LSTM LOS MAE changed only minimally across the six dataset variants. The annotated values range from 2.33 days for `D3` and `D6` to 2.36 days for `D2` and `D5`. The standard deviations shown below the x-axis range from 4.59 to 4.64 days.

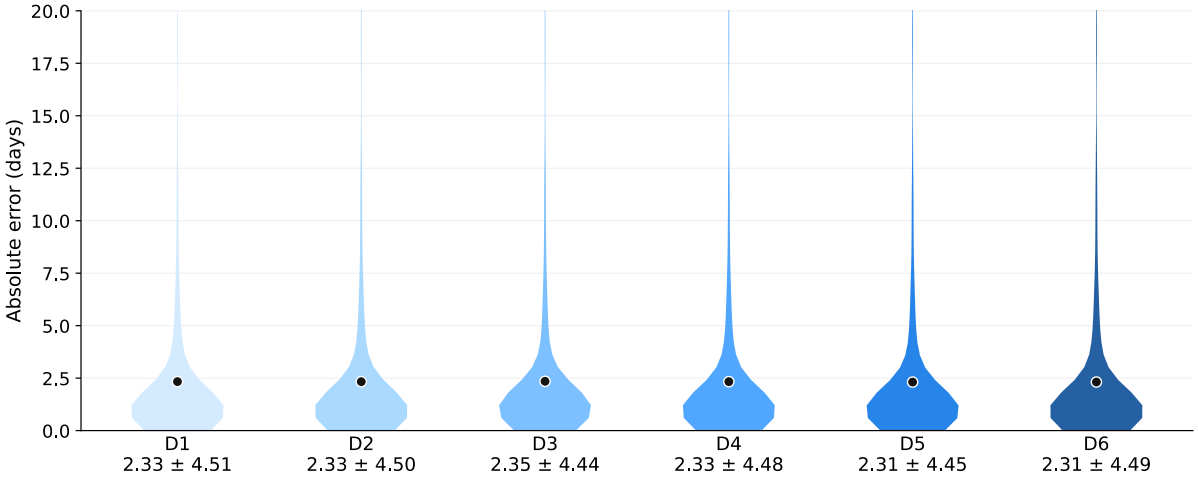

Figure 29: DeepECG LOS absolute-error violin plots across dataset variants `D1`–`D6`. The black dots indicate MAE in days, and the x-axis labels report MAE ± standard deviation.

Figure 29 shows a similarly narrow LOS MAE range for DeepECG. The annotated values range from 2.31 days for `D5` and `D6` to 2.35 days for `D3`. The standard deviations range from 4.44 to 4.51 days.

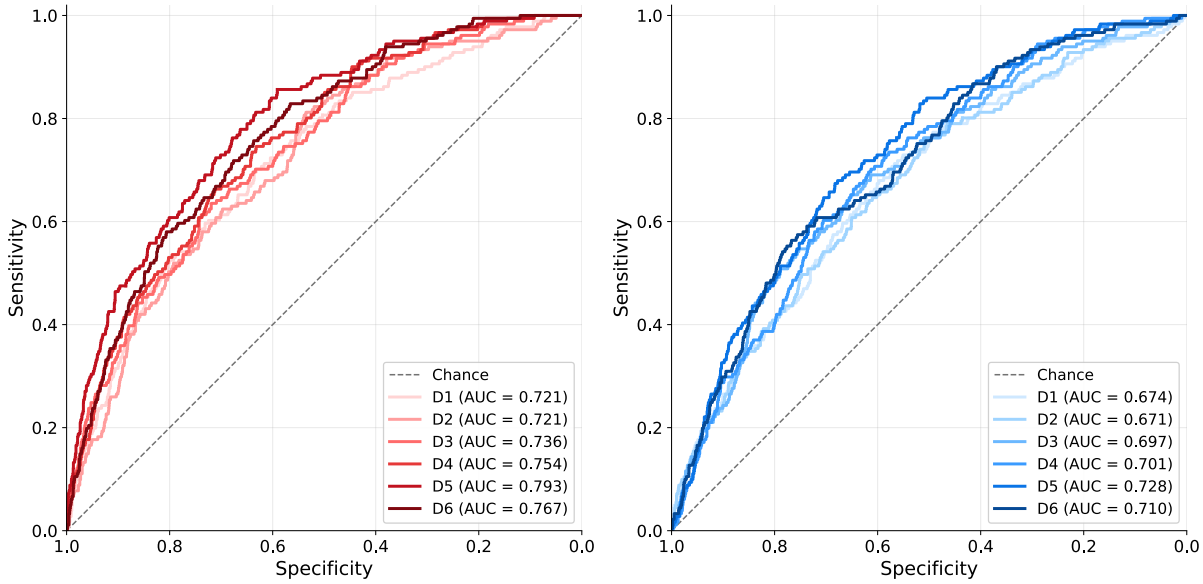

Figure 30: Mortality ROC curves across dataset variants `D1`–`D6`. *Left:* hybrid CNN–LSTM (`R18-R23`). *Right:* DeepECG (`R33-R38`). The dashed diagonal indicates chance performance, and the legend reports the AUC for each dataset variant.

Figure 30 shows the corresponding mortality ROC curves. For the hybrid CNN–LSTM, AUC values range from 0.72 for `D1` and `D2` to 0.79 for `D5`, with `D6` at 0.77. For DeepECG, AUC values range from 0.67 for `D2` to 0.73 for `D5`, with `D6` at 0.71. Across all matched dataset variants, the hybrid CNN–LSTM shows higher mortality AUC values than DeepECG. The $\leq$ 10-day LOS split is reported in the appendix to keep the main text focused on the overall feature-development comparison.

6.3.2 Conclusion of the Multimodal Comparison

Across the multimodal feature-development line, the LOS MAE values remain close to the ECG-only reference in both model families. For the hybrid CNN–LSTM, the range is 2.33–2.36 days across `D1`–`D6`. For DeepECG, the range is 2.31–2.35 days. On mortality, the hybrid CNN–LSTM ranges from 0.72 to 0.79 AUC, whereas DeepECG ranges from 0.67 to 0.73 AUC. DeepECG has slightly lower LOS MAE in most dataset variants, although not in every feature combination. The hybrid CNN–LSTM shows higher mortality AUC across all dataset variants. The subsequent hyperparameter optimisation in Section 6.4 focuses on the multimodal hybrid.

6.4 Final Optuna Hyperparameter Optimisation

After the cross-architecture comparisons in Sections 6.1–6.3, hyperparameter optimisation entered its final automated stage. This stage differs from the earlier predefined `H2`–`H4` refinements, which were used for controlled cross-architecture screening. Here, Optuna was used for a focused search on the retained multimodal hybrid CNN–LSTM line only. Two studies were run under the same model and training protocol, differing only in the late-fusion EHR table. One study used the full EHR block, while the second used the reduced EHR-small block introduced in Section 5.6. In both cases, demographic variables and ICU-unit encoding remained enabled, so the comparison isolates the effect of the broader versus reduced EHR representation within the same multimodal architecture. The ECG-only hybrid configuration from the preceding feature-development analysis remains the reference baseline for interpreting the added value of these multimodal final runs.

6.4.1 Tuning Setup

Both studies were formulated as minimisation problems on `best_val_loss`, that is, the minimum combined multitask validation loss observed across epochs. LOS and in-ICU mortality retained fixed equal loss weights of 1:1. Optuna used a tree-structured Parzen estimator sampler and a median pruner driven by per-epoch validation loss, while the first five trials were exempt from pruning. Early stopping remained part of the training loop and was independent of Optuna pruning. The search itself was restricted to the five hyperparameters in Table 10. This restriction was necessary to keep the tuning efficient under the available server capacity. The final run profiles are listed in the appendix run list as `H5` for the full-EHR configuration and `H6` for the EHR-small configuration, see Table 19. These profiles should not be read as manual extensions of the earlier `H4` screening setup. They document the final Optuna-selected settings for the five searched parameters, while the remaining training choices were fixed to a common final-stage protocol. These fixed choices, including AdamW-based optimisation, validation-based stopping, gradient clipping, and regularisation settings, follow common DL tuning practice in which learning rate, weight decay, dropout, and model capacity are selected on a validation objective under a constrained search budget [31, 33, 84].

For both studies, 100 trial slots were configured and the 48-hour wall time budget was fully used. Both runs terminated by timeout rather than by exhausting the configured trial count. As a result, 83 trials were started in the full-EHR study and 84 in the EHR-small study. Of these, 20 and 27 trials completed, respectively.

Hyperparameter	Sampling	Range / choices
`lr`	log-uniform	10^{-4} – 3×10^{-3}
`weight_decay`	log-uniform	10^{-6} – 10^{-3}
`dropout_rate` (dense/shared)	uniform	0.25 – 0.5
`lstm_dropout`	uniform	0.2 – 0.5
`lstm_hidden_preset`	categorical	`64_64`, `128_128`, `256_256`

Table 10: Optuna search space for the hybrid multimodal line, identical for both EHR variants.

6.4.2 Results and Metric Comparison

Table 11 summarises the hyperparameter combinations selected by Optuna for the two studies. In both cases, the retained configuration was the completed trial with the lowest validation objective `best_val_loss`. Under this criterion, the full-EHR study selected trial 73 with `best_val_loss` = 2.2298, whereas the EHR-small study selected trial 78 with `best_val_loss` = 2.2915.

Hyperparameter	Full EHR selected trial	EHR-small selected trial
`lr`	0.002997	0.001717
`weight_decay`	0.0003868	0.0009777
`dropout_rate`	0.2662	0.4109
`lstm_dropout`	0.4439	0.4838
`lstm_hidden_preset`	`64_64`	`256_256`

Table 11: Hyperparameter configurations selected by Optuna for the completed trial in each study.

The selected hyperparameters differed between the two studies. The full-EHR configuration combined a higher learning rate with lower weight decay, lower shared dropout, and the smallest LSTM preset `64_64`. By contrast, the selected EHR-small configuration used a lower learning rate, stronger regularisation, and the largest LSTM preset `256_256`.

To obtain final test metrics, each Optuna-selected configuration was trained once more as a separate final run under the same protocol. On the held-out test split, the full-EHR final run reached a LOS MAE of 2.28 days, a mortality AUC of 0.82, and a mortality $F1$ score of 0.33. The EHR-small final run reached a LOS MAE of 2.32 days, a mortality AUC of 0.77, and a mortality $F1$ score of 0.25. The corresponding confusion matrices for the final runs are provided in Appendix A.2.2.

Figure 31 shows the LOS absolute-error distributions for the two final runs on the full test cohort and on the subgroup with observed LOS $y_i \leq 10$ days. On the full test cohort, the full-EHR run has an MAE of 2.28 days with a standard deviation of 4.32 days, while the EHR-small run has an MAE of 2.32 days with a standard deviation of 4.41 days. In the 10-day subgroup, the full-EHR run has an MAE of 1.42 days with a standard deviation of 1.32 days, while the EHR-small run has an MAE of 1.44 days with a standard deviation of 1.36 days.

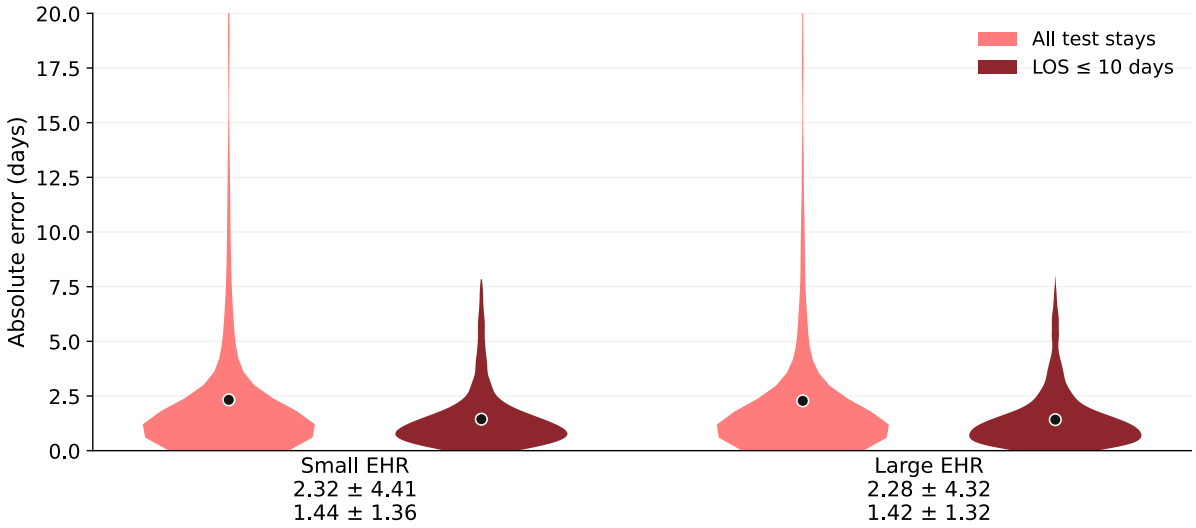

Figure 31: Final hybrid CNN–LSTM LOS absolute-error violin plots for the full-EHR and EHR-small Optuna-selected configurations. For each EHR representation, one violin shows all test stays and one violin shows stays with observed LOS $y_i \leq 10$ days. The black dots indicate MAE in days, and the x-axis labels report MAE ± standard deviation for both groups.

6.4.3 Comparison with SOFA for Mortality

To compare the final hybrid CNN–LSTM with a classical severity score, SOFA was calculated for the same ICU mortality endpoint. The score followed the original SOFA definition as the sum of six organ-system scores from 0 to 4, giving a total range from 0 to 24 [59]. Only measurements documented no later than the ECG timestamp were used. A stay was retained for this comparison only when at least four of the seven locally available SOFA input fields were present. Stays with fewer than 4 non-missing inputs were excluded from the comparison entirely. Among retained stays, any of the seven SOFA input variables that were still missing after aggregation were assigned 0 points for the corresponding organ-system component as a conservative imputation. This reduced the test subset to $n = 1980$ stays, corresponding to 72 % of the original test set. The stay-level SOFA value was obtained by averaging available `sofa_total` values across valid ECGs

for the same stay. The binary SOFA decision threshold was selected on the validation subset by maximising $F1$, resulting in a threshold of 10.

For the model comparison, the best full-EHR hybrid checkpoint (`4685434`) was loaded without retraining and evaluated on the same restricted test subset. The model output was the stay-level mortality probability p, with the validation-selected threshold $p \geq 0.22$ defining a positive mortality prediction. Figure 32 shows the resulting ROC curves. Both methods performed above chance, but the hybrid EHR model reached a higher AUC (0.832 vs. 0.758). At the selected decision thresholds, the hybrid model also achieved a higher $F1$ score (0.331 vs. 0.283) and sensitivity (0.46 vs. 0.35), while specificity remained similar (0.91 vs. 0.92). The corresponding SOFA and hybrid confusion matrices for this reduced test subset are provided in Appendix A.2.2.

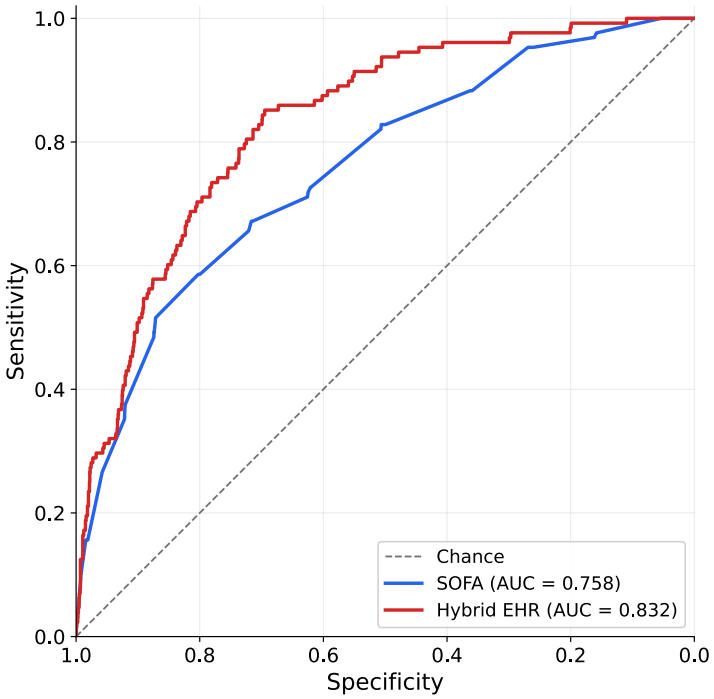

Figure 32: ROC comparison of SOFA and the final full-EHR hybrid CNN–LSTM on the shared mortality test subset.

Overall, the final optimisation stage should be read as the strongest observed results under a fixed compute budget rather than as an exhaustive search. The 48-hour budget limited both Optuna studies to a little more than 80 started trials, and median pruning substantially reduced the number of fully completed runs. Within this setting, the larger EHR table yielded the stronger final hybrid result, and the same full-EHR hybrid model also outperformed SOFA on the shared mortality subset.

7 Discussion

This chapter interprets the empirical findings, relates them to prior ICU and ECG modelling work, and discusses the main limitations and practical implications. Chapter 6 showed that admission ECGs contain prognostic signal, but that LOS is limited by the sparse long-stay tail and mortality by the rare positive class. Multimodal context helped mortality more clearly than LOS, and the hybrid CNN–LSTM emerged as the most practical final tuning candidate.

7.1 Relation to the State of the Art

This thesis is positioned between three related strands of prior work: ICU prediction from structured EHR data, dynamic modelling of ICU time series, and DL-based ECG representation learning. Previous studies have shown that structured admission and early EHR variables can support ICU mortality and LOS prediction [8, 27], while dynamic ICU models based on repeated clinical measurements can update risk estimates as new information becomes available [30, 62]. In parallel, recent ECG research has shown that raw 12-lead ECGs and large ECG foundation models contain prognostic information beyond classical rhythm classification [68–70]. Against this background, this thesis contributes a focused evaluation of the first raw 12-lead ECG after ICU admission for both LOS and in-ICU mortality, including the added value of demographic variables, ICU-unit encoding, and early EHR features. This connects ECG-based prognostic modelling with the ICU outcome-prediction literature, where raw ECG waveforms have so far been used less systematically than structured clinical time series [20].

7.2 Interpretation of the Main Findings

For LOS prediction, the model families already lay in a relatively narrow MAE range under the baseline setting, and this remained the case after predefined hyperparameter refinement. The hybrid CNN–LSTM combined competitive LOS performance with stronger mortality results and a more efficient final tuning path than the larger foundation-model lines. The main LOS challenge was the target distribution itself because most stays were short. Very long stays formed a sparse upper tail. The models were trained on many examples from the common LOS range and comparatively few examples from the range that contributed the largest absolute errors.

This imbalance is visible in the error distributions. The violin plots and Bland–Altman analysis show smaller errors in the short-stay range and larger errors in the long upper tail. In the Bland–Altman plots (Figure 22), underprediction became more pronounced as

true LOS increased. Cohort-level MAE should therefore be read together with the error distribution. The models capture a meaningful LOS signal in the subgroup up to 10 days. Deviations increase once stays move into the sparsely represented long-stay range.

For mortality prediction, the same imbalance problem appeared at the classification threshold. At the default threshold, the baseline models did not form an operationally useful mortality classifier. They mostly reproduced the majority class and classified almost all stays as survivors. The moderate AUC values therefore have to be read as threshold-independent ranking information, not as evidence that the default classifier was clinically usable. Only after threshold optimisation did the models reach non-zero and interpretable $F1$ scores. This distinction is important because discrimination and operational usefulness are not the same. The confusion matrix in Figure 27 illustrates the resulting trade-off: higher recall came with a clear increase in false positives, so threshold selection directly affects the balance between missed cases and additional false alarms.

The baseline mortality comparison also limits the claim that raw ECG input was uniformly superior. Under the shared `H1` setting, feature-based XGBoost reached the highest AUC. This indicates that handcrafted ECG descriptors already captured relevant mortality-ranking signal. The raw-waveform DL models were therefore most useful as a flexible basis for end-to-end representation learning and later multimodal fusion, rather than as a clearly stronger ECG-only baseline.

The multimodal experiments clarify where structured clinical context added value. Additional tabular context had only a small effect on LOS MAE. Mortality AUC improved more visibly. This suggests that demographic, ICU-unit, and early EHR information added complementary signal mainly for risk stratification rather than for exact LOS regression. The final Optuna experiments support this reading: within the fixed compute budget, the multimodal hybrid line with the full EHR block achieved the best reported combined result for both tasks. The gain should still be read cautiously because the feature-development curves remained close together. ECG is therefore a meaningful anchor modality.

The SOFA comparison points in the same direction. On the shared reduced test subset, the full-EHR hybrid model outperformed SOFA for mortality prediction, with a higher AUC (0.832 vs. 0.758) and a higher $F1$ score at the validation-selected threshold (0.331 vs. 0.283) (Section 6.4.3). The model therefore used predictive signal beyond the fixed organ-dysfunction summary captured by SOFA. This comparison is limited to stays with sufficient SOFA inputs.

7.3 Critical Appraisal of the Study

Several aspects of the study strengthen the validity of the comparison. The experiments followed a clear registry of architecture, preprocessing, augmentation, hyperparameter, and dataset identifiers. patient-level splitting reduced leakage from repeated ECGs, and the progression from broad screening to targeted optimisation made the workflow transparent. These are important strengths because they improve reproducibility and make the modelling decisions easier to audit.

At the same time, the thesis has clear limitations that constrain the interpretation of the findings. The study was retrospective and based on a single database, so external validation within this project was not possible. The work used standard 12-lead ECGs and structured clinical variables that are routinely available in many hospitals. The main uncertainty is whether a model trained on this cohort would maintain its performance under different workflows, measurement practices, and patient mixes. In addition, the work relies on diagnostic 10 s ECGs recorded within the first ICU day rather than on continuous bedside waveforms. This fits the practical goal of using routinely available data and limits the amount of physiological information available to the models. The multimodal comparisons also need to be interpreted with care because, within the same overall cohort, early measurements were not equally complete for every patient. The comparison between ECG-only and multimodal variants therefore reflects both the value of added features and differences in coverage and completeness of the structured input.

Further limitations arise from the targets and the available compute budget. The target imbalance discussed above constrains how strongly the results can be generalised: long LOS cases and mortality-positive stays were precisely the cases of greatest clinical interest and the least represented in the data. Mortality was defined as death before ICU discharge, so the endpoint should be interpreted specifically as outcome during intensive care treatment. A central project constraint was the limited server compute time. This shaped the experimental programme, as not all model families could be tuned to the same depth. The final Optuna search was restricted to the multimodal hybrid line, and even there the study ended because of time limits rather than because the search space was exhausted. The reported best configurations should therefore be read as the best observed settings under constrained resources, not as global optima. This limits how strongly the final ranking between model families should be interpreted.

7.4 Clinical and Practical Implications

From a clinical perspective, the results support a cautious decision-support interpretation. The models are not strong enough for autonomous triage or discharge planning. They show that early ECG-based modelling can contribute to risk stratification, especially when combined with demographic variables, ICU-unit information, and early EHR features. For LOS, the subgroup analysis up to 10 days is consistent with prior work that treats short and long ICU stay categories as clinically relevant prediction targets [8]. For mortality, threshold selection remains central because the rare positive class makes the default operating point unsuitable. These findings align with the stakeholder requirements discussed in the introduction, where interpretability, workflow fit, and alarm burden are as important as predictive performance.

This point is especially important because the clinically relevant output is not only a numerical score. ICU staff would need to understand why a patient is assigned to a higher-risk group, which input sources contributed to the estimate, and how stable the estimate is under ordinary data-quality variation. Without such context, even a model with acceptable aggregate performance could be difficult to trust, difficult to document, or poorly aligned with existing clinical responsibility. The results of this thesis therefore support an assistive use case rather than an autonomous decision rule. A practical system should present model outputs together with transparent feature context, calibrated risk information, and validation-based thresholds, so that clinicians can interpret the prediction as one additional signal within the ICU workflow rather than as a replacement for clinical judgement.

8 Outlook

This chapter outlines future work that follows from the limitations discussed in Section 7. The present work established a comparative baseline for ECG-based ICU outcome prediction. The next steps concern targeted model refinement, more systematic handling of imbalanced outcomes, and validation beyond the retrospective benchmark.

8.1 Extending the Experimental Programme

One relevant next step is to extend the experimental programme under a larger compute budget. The restricted server budget and the fixed wall-time limit shaped which architectures could be screened, how deeply they could be tuned, and why the final Optuna study was limited to the multimodal Hybrid line. Future work should therefore revisit targeted tuning for DeepECG-SL and HuBERT-ECG. Because these are large and complex architectures, this should be preceded by a careful analysis of which configurations are most promising, including input features, fusion strategy, learning rates, and regularisation. Such a staged selection would help use the additional compute budget for settings with the highest expected value.

Future work should also address class imbalance more systematically. For mortality, the current results showed that useful classification depended strongly on threshold tuning. A natural next step would therefore be to compare balanced or rebalanced training subsets with methods that act directly on the full cohort, such as oversampling, class-weighted losses, focal loss, and alternative threshold strategies. A related issue also appears in LOS prediction. Extremely long stays were rare in the available data, which limited how well the models could learn this part of the target distribution and contributed to weaker performance on the long-stay tail. Future work should therefore study both class imbalance in mortality prediction and imbalanced target distributions in LOS modelling. The test set should retain the realistic outcome distribution, while the training stage can use imbalance-aware sampling, loss functions, and threshold strategies. In parallel, the experimental grid could be expanded beyond the defined hyperparameter setups by revisiting preprocessing and augmentation choices under a larger compute budget. Further extensions include subgroup analyses where sample size permits and alternative LOS formulations that are less sensitive to the long-stay tail. External validation on an independent ICU-ECG cohort is especially important for the foundation-model experiments, because MIMIC data were included in their development or pretraining context. Testing on a separate cohort would therefore help distinguish genuine generalisation from dataset familiarity.

8.2 Prototype Web Interface

As a secondary outcome of the project, a small web interface was also built as an interactive inference prototype. The prototype is available in the accompanying code repository at `https://github.com/pr014/Master-thesis` in the `app/` directory, where it can be tested locally. It allows a trained model to be loaded, accepts ECG input together with the structured variables used in the multimodal pipeline, including ICU unit, demographic information, and EHR fields, and returns LOS and mortality predictions. Selected views of this prototype are shown in Appendix A.3. Although this interface was not part of the benchmark evaluation, it demonstrates that the modelling pipeline can already be transferred into a simple user-facing workflow.

This prototype also opens a practical direction for future work beyond offline benchmarking. Validation in a real ICU environment would be interesting in order to observe how such a tool is used in practice and which outputs are actually useful for clinicians. In that setting, the software could be refined together with ICU staff so that the interface, input fields, output presentation, and general workflow reflect clinical expertise rather than technical assumptions alone. This would not yet amount to deployment, but it would provide an important intermediate step between retrospective modelling and later clinical translation.

8.3 Summary

Overall, future work should move from retrospective benchmarking towards more robust validation and clinical translation. Methodologically, this requires targeted follow-up experiments and external validation on an independent ICU-ECG cohort. Practically, the web prototype provides a starting point for co-development with ICU staff and, at a later stage, a prospective clinical study. The following chapter closes the thesis by summarising the problem, the approach, and the main contributions.

9 Conclusion

9.1 Problem and Approach

This thesis addressed the early prediction of ICU LOS and in-ICU mortality from data available at or shortly after ICU admission. The central idea was an ECG-first framework that starts from the first diagnostic 12-lead ECG linked to the ICU stay and then extends this signal with demographic variables, ICU-unit context, and structured early EHR features through late fusion. Using MIMIC-IV and MIMIC-IV-ECG, the work implemented a reproducible experimental registry, compared classical and DL architectures, and evaluated both ECG-only and multimodal variants under a fixed multitask setting for LOS regression and mortality classification. Model development followed a staged workflow from preprocessing and augmentation through cross-architecture screening to targeted multimodal optimisation under a restricted compute budget.

9.2 Answers to the Research Questions

Research Question 1

What level of discrimination, calibration, and robustness do DL models achieve when predicting ICU LOS and mortality from the first raw ECG recorded at ICU admission?

The first admission ECG provided measurable but not standalone predictive value. For LOS, the best models reached test MAE values in the low two-day range, but errors increased clearly for very long stays. The signal was therefore most useful for the common short- to medium-stay range. For mortality, the models reached moderate AUC values and learned a useful risk ranking, however positive classification required validation-based threshold selection because the default threshold was not suitable for the imbalanced endpoint. The results are robust as an internal comparison because patient-level splitting, fixed identifiers, and a transparent experiment registry made the rankings traceable. External robustness remains unproven because the study was retrospective, single-centre, and compute-constrained.

Research Question 2

To what extent does integrating ECG with structured EHR and demographic data improve predictive performance compared to ECG-only models?

Integrating structured clinical context improved performance, but the effect differed between the two tasks. For LOS, MAE changed only slightly across the dataset variants `D1`–`D6`, so the richer feature sets were only marginally better than the ECG-only reference.

For mortality, structured clinical context improved the overall picture, especially in the multimodal comparisons. At the same time, the largest practical gain in classification performance came from threshold tuning rather than from feature addition alone. The strongest combined result in the final tuning stage was obtained for the multimodal hybrid CNN–LSTM with the full EHR block. Overall, the results show that structured tabular context adds more clearly to mortality risk stratification than to exact LOS regression.

Research Question 3

Which ML architectures are best suited for ECG-based ICU outcome prediction when predictive performance, robustness, and practical feasibility are considered together?

No single architecture dominated all settings by a wide margin. After tuning, Uni-LSTM, Bi-LSTM, hybrid CNN–LSTM, and DeepECG-SL formed a relatively close group, especially for LOS. Within this set, the hybrid CNN–LSTM offered the clearest practical balance between predictive quality, consistent performance across both endpoints, and computational cost, which is why it was selected for the final multimodal Optuna study. DeepECG-SL remained competitive and was in some comparisons slightly stronger, especially in the multimodal feature analysis, but at much higher computational cost. The hybrid CNN–LSTM was therefore pursued deliberately in the final tuning stage because it promised the most useful balance between performance, efficiency, and multimodal scalability. Compared with XGBoost, the main gain of DL was not a uniform numerical advantage in every setting, but the ability to learn directly from raw ECG waveforms and to combine waveform representations with structured clinical context in one multitask model. The architectural conclusions should nevertheless be read in light of the fact that tuning depth was not identical across all families.

9.3 Closing Remarks

Taken together, this thesis showed that admission ECGs contain relevant prognostic information for early ICU LOS and mortality prediction, especially when combined with compact structured clinical context. Its main contribution is a systematic and reproducible comparison of modelling strategies under a clearly documented experimental design. The findings support ECG-based ICU prediction as a meaningful direction for decision-support research, but they also show clear limits caused by target difficulty, imbalance in mortality and long LOS, and restricted compute resources. Within these limits, the thesis provides a grounded basis for further work. With external validation, calibration, and clinical workflow testing, this line of work could contribute to earlier risk assessment and, ultimately, to better resource allocation in the ICU.

A Appendix

A.1 Methods

Supplementary material for Chapter 5: DL hyperparameters, multimodal inputs, and preprocessing.

A.1.1 DL Models

Full `H1`–`H6` training profiles for the retained DL models (Section 5.7). XGBoost and CNN-from-scratch omitted (`H1` only).

Unidirectional LSTM

Hyperparameter	`H1`	`H2`	`H3`	`H4`
Learning rate	5×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Batch size	64	128	128	128
Weight decay	1×10^{-4}	1×10^{-4}	2×10^{-4}	2×10^{-4}
LSTM dropout	0.2	0.3	0.4	0.4
Dropout rate	0.3	0.3	0.4	0.4
Gradient clipping (max norm)	1	1	1	1
LR scheduler patience	5	5	5	5
Early stopping patience	7	7	12	12
Hidden dimension	128	128	128	128
Embedding dimension	64	128	128	128
Adam β	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]
LR scheduler factor	0.1	0.1	0.5	0.5
Minimum learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Input noise std. dev.	0.03	0.03	0.03	0.03
Loss	MSE	MSE	MSE	Huber
Optimiser	Adam	Adam	Adam	Adam

Bidirectional LSTM

Hyperparameter	H1	H2	H3	H4
Learning rate	5×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
Batch size	64	128	128	128
Weight decay	1×10^{-4}	2×10^{-4}	2×10^{-4}	2×10^{-4}
LSTM dropout	0.2	0.3	0.4	0.4
Dropout rate	0.3	0.3	0.4	0.4
Gradient clipping (max norm)	1	1	1	1
LR scheduler patience	5	7	10	10
Early stopping patience	7	7	12	12
Hidden dimension	128	256	128	128
Embedding dimension	64	128	128	128
Adam β	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]
LR scheduler factor	0.1	0.1	0.5	0.5
Minimum learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Input noise std. dev.	0.03	0.03	0.03	0.03
Loss	MSE	MSE	MSE	Huber
Optimiser	Adam	Adam	Adam	Adam

Hybrid CNN–LSTM

Hyperparameter	H1	H2	H3	H4	H5	H6
Learning rate	5×10^{-4}	1×10^{-4}	2×10^{-4}	2×10^{-4}	2.997×10^{-3}	1.717×10^{-3}
Batch size	64	64	64	64	64	64
Weight decay	1×10^{-4}	3×10^{-4}	2×10^{-4}	2×10^{-4}	3.868×10^{-4}	9.777×10^{-4}
LSTM dropout	0.2	0.4	0.4	0.4	0.4439	0.4838
Dropout rate	0.3	0.3	0.4	0.4	0.2662	0.4109
Gradient clipping (max norm)	1	1	1	1	2	2
LR scheduler patience	5	7	7	7	6	6
Early stopping patience	7	10	12	12	15	15
Hidden dimension	128	256	128	128	64	256
Embedding dimension	64	64	64	64	64	256
Adam β	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]	[0.9, 0.999]
LR scheduler factor	0.1	0.1	0.5	0.5	0.5	0.5
Minimum learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Input noise std. dev.	0.03	0.03	0.03	0.03	0.03	0.03
Loss	MSE	MSE	MSE	Huber	Huber	Huber
Optimiser	Adam	Adam	Adam	Adam	AdamW	AdamW

DeepECG-SL

Hyperparameter	H1	H2	H3
Learning rate	5×10^{-4}	5×10^{-4}	5×10^{-4}
Batch size	64	32	32
Weight decay	1×10^{-4}	1×10^{-2}	1×10^{-2}
LSTM dropout	—	—	—
Dropout rate	0.3	0.1	0.1
Gradient clipping (max norm)	1	1	1
LR scheduler patience	5	8	8
Early stopping patience	7	15	15
Hidden dimension	768	768	768
Embedding dimension	256	256	256
Adam β	[0.9, 0.999]	[0.9, 0.98]	[0.9, 0.98]
LR scheduler factor	0.1	0.5	0.5
Minimum learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}
Input noise std. dev.	0.03	0.03	0.03
Loss	MSE	MSE	Huber
Optimiser	Adam	Adam	Adam

HuBERT-ECG

Hyperparameter	H1	H2
Learning rate	5×10^{-4}	Head (phase 1): 1×10^{-4}; backbone: 1×10^{-5}; deep backbone: 1×10^{-7}
Batch size	32	32
Weight decay	1×10^{-4}	0.01
LSTM dropout	—	—
Dropout rate	0.1	0.1
Gradient clipping (max norm)	1	1
LR scheduler patience	5	Linear schedule with warmup
Early stopping patience	7	7
Hidden dimension	768	768
Embedding dimension	768	768
Adam β	[0.9, 0.999]	[0.9, 0.98]
LR scheduler factor	0.1	Linear schedule with warmup (8% warmup)
Minimum learning rate	1×10^{-6}	1×10^{-6}
Input noise std. dev.	0.03	0.03
Loss	MSE	MSE
Optimiser	Adam	Adam
Backbone (encoder) LR	5×10^{-4}	1×10^{-5}
Head LR	5×10^{-4}	1×10^{-4}
Scheduler type	ReduceLROnPlateau	Linear schedule with warmup
Warmup steps (% of steps)	—	8%
Layer-wise LR	No	Yes (1×10^{-7}, 1×10^{-5}, 1×10^{-4})
Phase 1	—	Frozen backbone; train heads only (max. 50 epochs)
Phase 2	—	Unfrozen backbone (max. 20 epochs)

A.1.2 Mulitmodal Fusion

EHR feature data

Feature	Clinical role
`respiratory_rate`	Respiratory frequency as a basic indicator of respiratory stress and ventilatory demand
`oxygen_saturation`	Peripheral oxygen saturation as a compact marker of oxygenation status
`map`	Mean arterial pressure as an indicator of overall perfusion pressure
`sbp`	Systolic blood pressure as a marker of macrocirculatory status
`dbp`	Diastolic blood pressure as a complementary pressure measure with vascular relevance
`temperature`	Body temperature as a basic physiological marker of infection, inflammation, or systemic stress
`gcs_eye`	Glasgow Coma Scale eye subscore as a compact descriptor of neurological responsiveness
`gcs_motor`	Glasgow Coma Scale motor subscore as a clinically important neurological function indicator
`platelets`	Platelet count as a laboratory marker relevant to bleeding risk, inflammation, and critical illness severity
`creatinine`	Creatinine as a marker of renal function and organ dysfunction
`urine_output`	Cumulative urine volume as an interval-based indicator of renal perfusion and fluid balance
`weight`	Body weight as supportive context for general physiological status and bedside measurement context

Table 12: Full EHR feature block used in the late-fusion experiments.

EHR feature data small

Feature	Clinical role
`respiratory_rate`	Basic respiratory monitoring variable with high bedside availability
`oxygen_saturation`	Non-invasive oxygenation marker available in routine monitoring
`sbp`	Systolic blood pressure as a standard bedside hemodynamic variable
`dbp`	Diastolic blood pressure as a complementary bedside hemodynamic variable
`temperature`	Core physiological status indicator routinely documented in ICU care
`weight`	Bedside context variable that can be obtained without laboratory testing

Table 13: Reduced EHR-small subset used for the late-fusion comparison.

ICU unit feature

ICU unit	Short description
Medical Intensive Care Unit (MICU)	Internal medicine intensive care unit, not primarily surgical
Medical/Surgical Intensive Care Unit (MICU/SICU)	Mixed medical-surgical intensive care unit
Cardiac Vascular Intensive Care Unit (CVICU)	Cardiovascular intensive care unit, e.g. postoperative cardiac or vascular care
Surgical Intensive Care Unit (SICU)	General surgical intensive care unit
Coronary Care Unit (CCU)	Coronary or cardiac monitoring unit, typically for acute cardiac conditions
Trauma SICU (TSICU)	Trauma-surgical intensive care unit
Neuro Intermediate	Neurological intermediate-care unit with closer monitoring below full ICU intensity
Neuro Surgical Intensive Care Unit (Neuro SICU)	Neurosurgical intensive care unit
Neuro Stepdown	Neurological step-down or transition unit after higher-intensity care
Surgery/Vascular/Intermediate	Surgery or vascular intermediate-care setting

Table 14: ICU-unit categories used for the `first_careunit` encoding

.

Median LOS by ICU unit

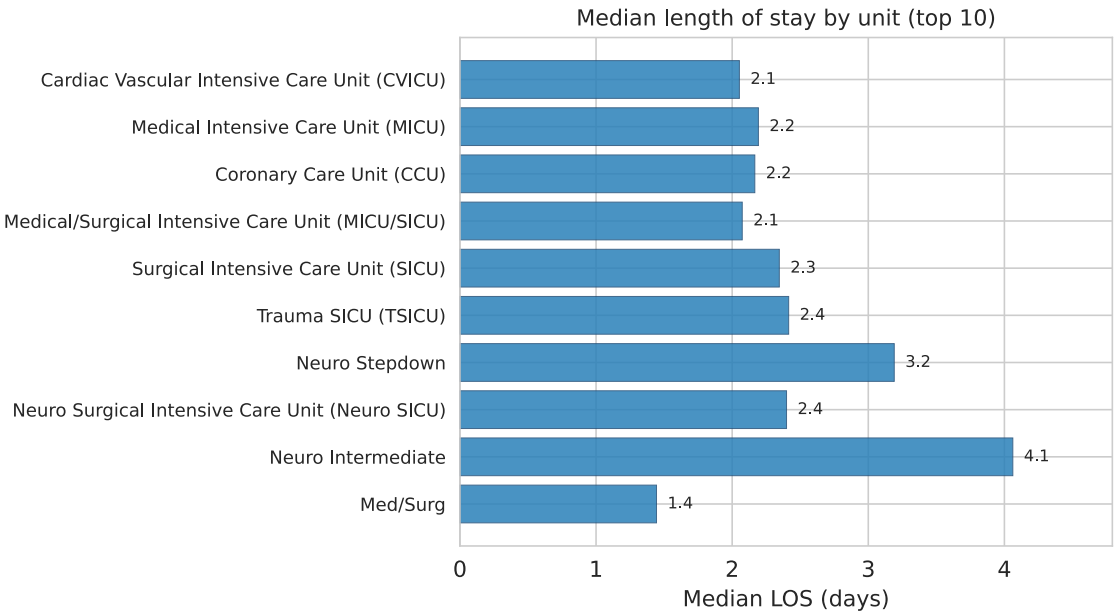

Figure 33: Median LOS by ICU unit, top ten by frequency.

ICU mortality by ICU unit

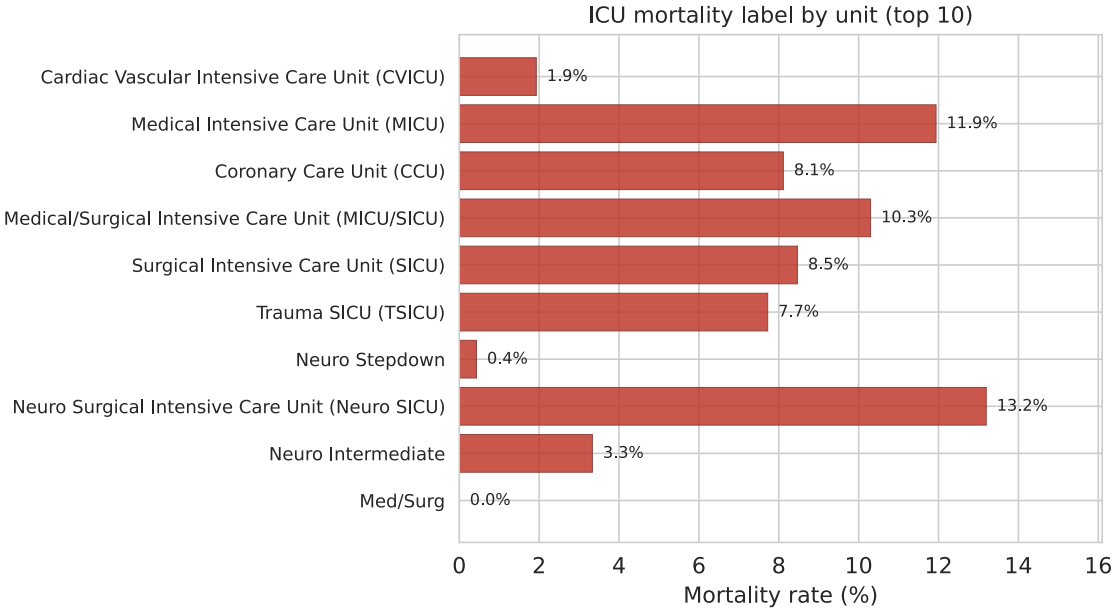

Figure 34: Mortality rate by ICU unit, top ten by frequency.

A.2 Results

A.2.1 Preanalysis

The following figures reproduce the CNN-based sensitivity checks described in Section 5.8.4. They were carried out before the main cross-architecture comparison and are reported here for completeness.

Augmentation Testing

Comparison of augmentation bundles on the CNN-from-scratch baseline under the shared `P1` preprocessing setting. Left: LOS test MAE; right: mortality test AUC.

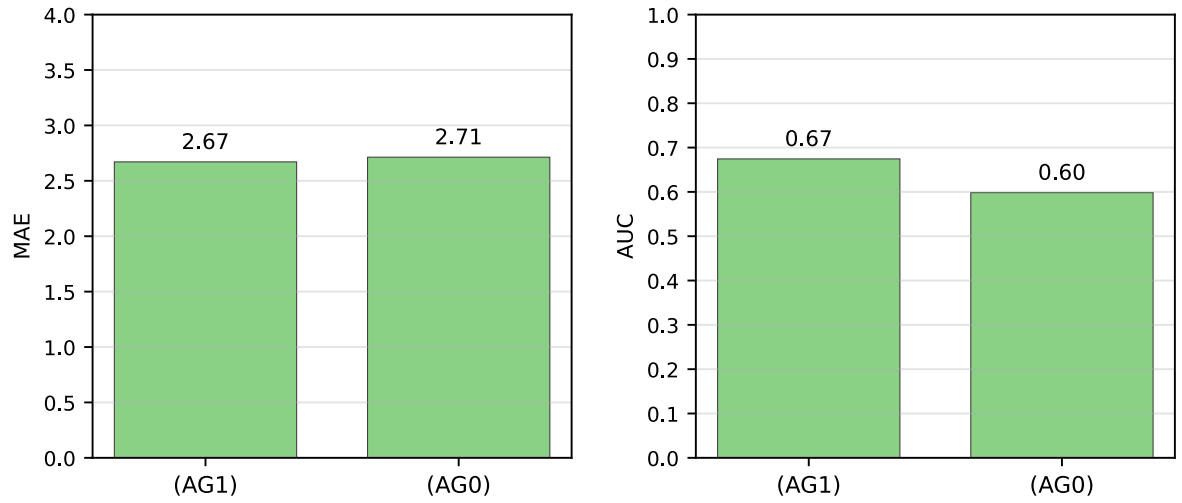

Figure 35: Exploratory CNN augmentation test. *Left:* LOS test MAE. *Right:* mortality test AUC.

Amplitude Testing

Comparison of the default normalised pipeline (`P1`, `AG1`) against a setting without per-lead z-score normalisation combined with amplitude scaling (`P2`, `AG2`). Left: LOS test MAE; right: mortality test AUC.

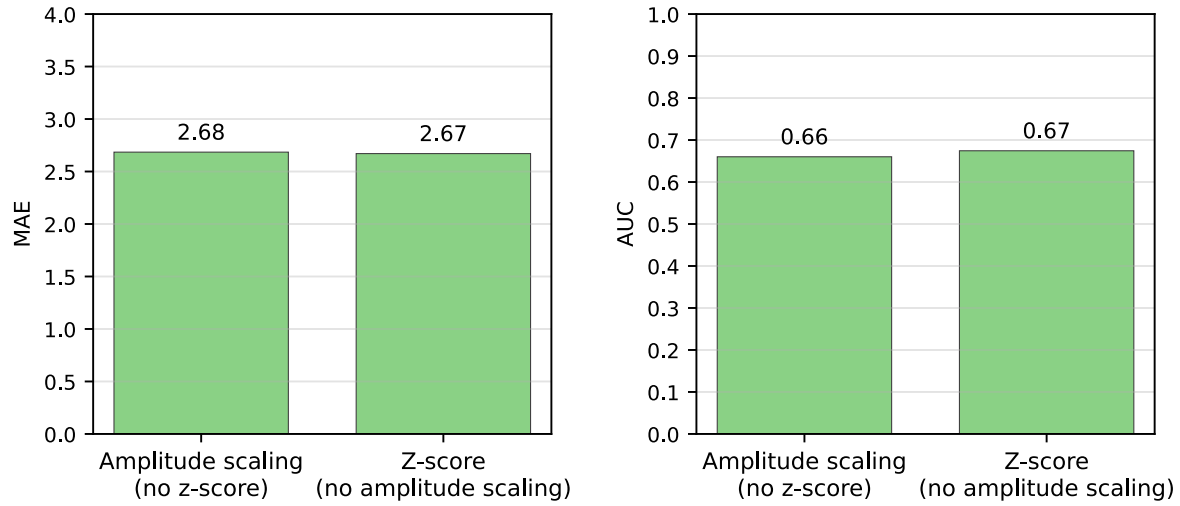

Figure 36: Exploratory CNN amplitude test comparing the default normalised setting against a subset without per-lead z-score normalisation. *Left:* LOS test MAE. *Right:* mortality test AUC.

A.2.2 Overview Results Architectures

The tables below list the final test-split metrics for each retained run. Identifier codes follow Section 5.3; `job_id` is omitted here but retained in the experiment registry. Runs are grouped by model family in the same order as in the main results chapter.

Classical ML: XGBoost

Gradient-boosted trees on tabular features; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality
Run	`A`	`P`	`H`	`AG`	`D`	MAE	R^2	AUC
R1	`A1`	`P1`	`H1`	`AG1`	`D1`	2.8211	−0.0021	0.7193

Table 15: Run list for the XGBoost baseline.

CNN from scratch

Convolutional network trained end-to-end on ECG waveforms; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality
Run	A	P	H	AG	D	MAE	R^2	AUC
R2	A2	P1	H1	AG1	D1	2.6707	0.0228	0.6744
R3	A2	P1	H1	AG0	D1	2.7131	0.0165	0.5983
R4	A2	P1	H1	AG4	D7	N/A	N/A	0.715
R5	A2	P2	H1	AG2	D1	2.743	0.0178	0.6431

Table 16: Run list for the CNN-from-scratch baseline.

Unidirectional LSTM

Unidirectional LSTM on ECG sequences; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality
Run	A	P	H	AG	D	MAE	R^2	AUC
R6	A3	P1	H1	AG1	D1	2.7378	0.008	0.6648
R7	A3	P1	H2	AG1	D1	2.6992	0.0138	0.6621
R8	A3	P1	H3	AG1	D1	2.6585	0.0111	0.6679
R9	A3	P1	H4	AG1	D1	2.3584	−0.0997	0.7038

Table 17: Run list for the unidirectional LSTM baseline.

Bidirectional LSTM

Bidirectional LSTM on ECG sequences; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality
Run	A	P	H	AG	D	MAE	R^2	AUC
R10	A4	P1	H1	AG1	D1	2.6663	0.004	0.6646
R11	A4	P1	H2	AG1	D1	2.8267	0.0023	0.5979
R12	A4	P1	H3	AG1	D1	2.731	0.0083	0.6623
R13	A4	P1	H4	AG1	D1	2.3409	−0.0778	0.6901

Table 18: Run list for the bidirectional LSTM baseline.

Hybrid CNN–LSTM

Hybrid CNN–LSTM on ECG sequences; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality	
Run	A	P	H	AG	D	MAE	R^2	AUC	Notes
R14	A5	P1	H1	AG1	D1	2.5019	0.0091	0.7049	hyperparametertesting
R15	A5	P1	H2	AG1	D1	2.6319	−0.0074	0.6576	hyperparametertesting
R16	A5	P1	H3	AG1	D1	2.7114	0.0035	0.6749	hyperparametertesting
R17	A5	P1	H4	AG1	D1	2.3234	−0.0863	0.6924	hyperparametertesting
R18	A5	P1	H4	AG4	D1	2.3379	−0.1026	0.721	feature testing
R19	A5	P1	H4	AG4	D2	2.3598	−0.1166	0.7076	feature testing
R20	A5	P1	H4	AG4	D3	2.3334	−0.102	0.7361	feature testing
R21	A5	P1	H4	AG4	D4	2.3356	−0.1031	0.7538	feature testing
R22	A5	P1	H4	AG4	D5	2.3577	−0.1237	0.793	feature testing
R23	A5	P1	H4	AG4	D6	2.3265	−0.1001	0.7667	feature testing
R24	A5	P1	H4	AG4	D5	2.3616	−0.0709	0.7158	test: randomtimeshuffle
R25	A5	P1	Tuning	AG4	D5	—	—	—	Optuna-tuning
R26	A5	P1	Tuning	AG4	D6	—	—	—	Optuna-tuning
R27	A5	P1	H5	AG4	D5	2.2777	0.0118	0.8164	
R28	A5	P1	H6	AG4	D6	2.3235	−0.0317	0.7718	

Table 19: Run list for the hybrid CNN–LSTM baseline.

DeepECG-SL

Pretrained DeepECG-SL encoder with task-specific heads; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality	
Run	A	P	H	AG	D	MAE	R^2	AUC	Notes
R29	A6.1	P1	H1	AG1	D1	2.7222	0.0191	0.7111	unfrozen
R30	A6.1	P1	H2	AG1	D1	2.6113	0.0172	0.3273	unfrozen
R31	A6.1	P1	H2	AG1	D1	2.7186	0.0171	0.2674	frozen
R32	A6.1	P1	H3	AG1	D1	2.3505	−0.0467	0.673	unfrozen
R33	A6.1	P1	H3	AG4	D1	2.3344	−0.0687	0.6736	unfrozen
R34	A6.1	P1	H3	AG4	D2	2.3295	−0.0639	0.6712	unfrozen
R35	A6.1	P1	H3	AG4	D3	2.3472	−0.0472	0.6966	unfrozen
R36	A6.1	P1	H3	AG4	D4	2.3307	−0.0589	0.7011	unfrozen
R37	A6.1	P1	H3	AG4	D5	2.3114	−0.0416	0.7285	unfrozen
R38	A6.1	P1	H3	AG4	D6	2.314	−0.0577	0.7095	unfrozen
R39	A6.2	P1	H3	AG4	D1	2.3271	−0.0755	0.6604	unfrozen / weight analysis
R40	A6.1	P1	H3	AG4	D1	2.328	−0.065	0.6823	unfrozen / weight analysis
R41	A6.3	P1	H3	AG4	D1	2.3393	−0.0705	0.6037	unfrozen / weight analysis
R42	A6.4	P1	H3	AG4	D1	2.3488	−0.0811	0.6261	unfrozen / weight analysis
R43	A6.1	P1	H3	AG4	D1	2.3397	0.0083	0.8332	unfrozen

Table 20: Run list for the DeepECG-SL baseline.

HuBERT-ECG

HuBERT-ECG pretrained encoder runs; each row links preprocessing, augmentation, and tuning IDs to LOS (MAE, R^2) and in-ICU mortality (AUC). Missing entries use — or `N/A`.

						LOS		Mortality	
Run	A	P	H	AG	D	MAE	R^2	AUC	Notes
R44	A7.1	P1	H1	AG1	D1	3.563	−0.5296	0.5266	frozen
R45	A7.1	P1	H1	AG1	D1	2.8337	−0.0019	0.4702	unfrozen
R46	A7.1	P1	H2	AG1	D1	2.6988	0.0069	0.6882	unfrozen
R47	A7.2	P1	H2	AG4	D1	2.7126	0.0114	0.7096	unfrozen
R48	A7.3	P1	H2	AG4	D1	N/A	N/A	N/A	failed → timeout

Table 21: Run list for the HuBERT-ECG baseline.

Feature Development on the 10-Day LOS Subset

LOS absolute-error distributions for the ≤ 10-day subgroup across the multimodal feature-development variants D1–D6. Each dataset variant is represented by its own colour.

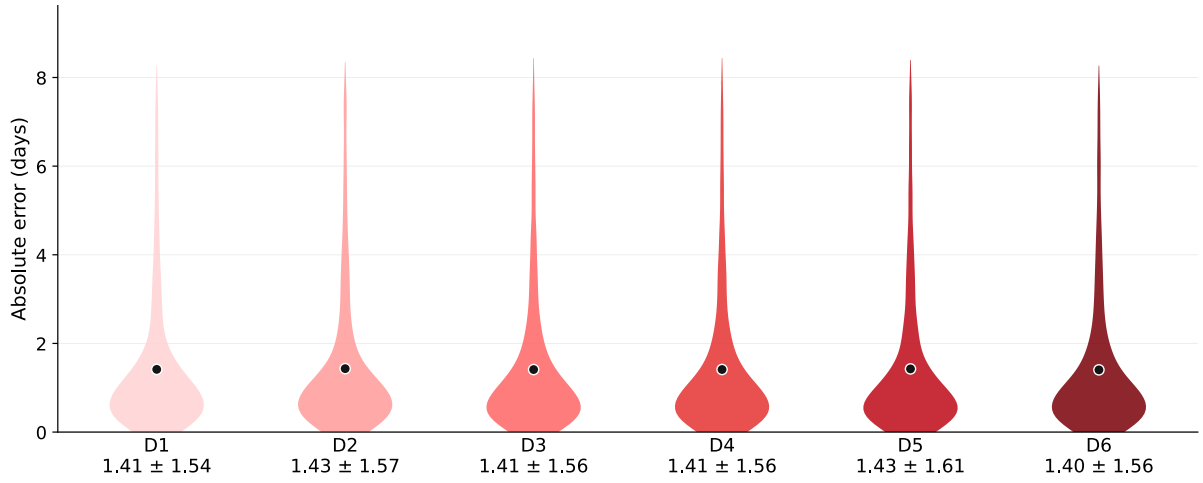

Figure 37: Hybrid CNN–LSTM feature-development LOS absolute-error violin plot on the ≤ 10-day subgroup across D1–D6. The black dots indicate MAE in days, and the x-axis labels report MAE ± standard deviation.

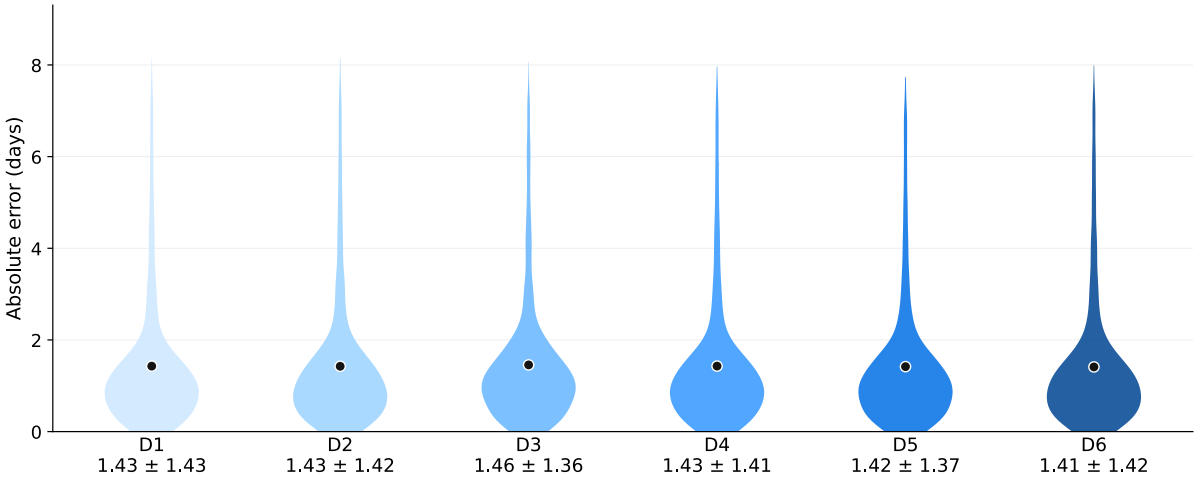

Figure 38: DeepECG feature-development LOS absolute-error violin plot on the $\leq$ 10-day subgroup across `D1`–`D6`. The black dots indicate MAE in days, and the x-axis labels report MAE $\pm$ standard deviation.

Confusion Matrices of final tuning run

Confusion matrices for the final Optuna-selected multimodal Hybrid CNN–LSTM configurations on the held-out test split: `R27` with the full EHR block (left) and `R28` with the reduced EHR-small block (right).

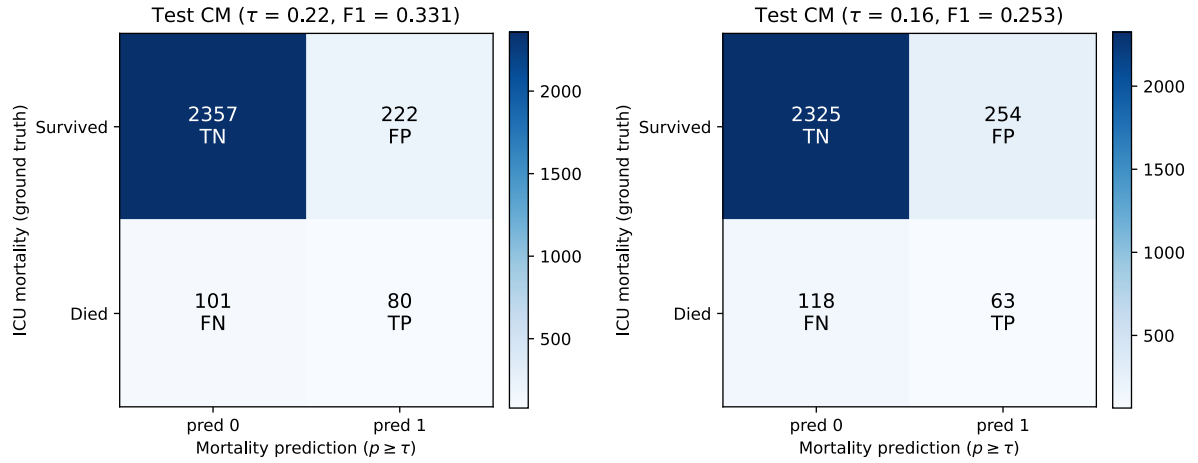

Figure 39: Confusion matrices for the final tuning runs. *Left:* `R27` with the full EHR feature block. *Right:* `R28` with the EHR-small feature block.

Confusion Matrices on the Reduced Test Set

SOFA and full-EHR Hybrid CNN–LSTM confusion matrices on the shared reduced mortality test subset used in Section 6.4.3.

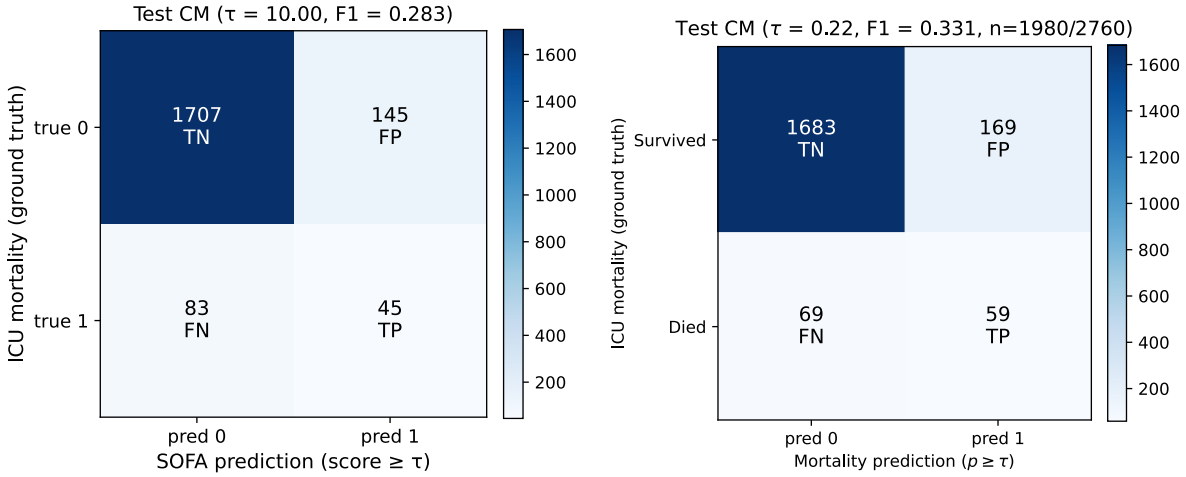

Figure 40: Confusion matrices on the reduced mortality test subset. *Left:* SOFA. *Right:* full-EHR hybrid CNN–LSTM.

A.3 Web app

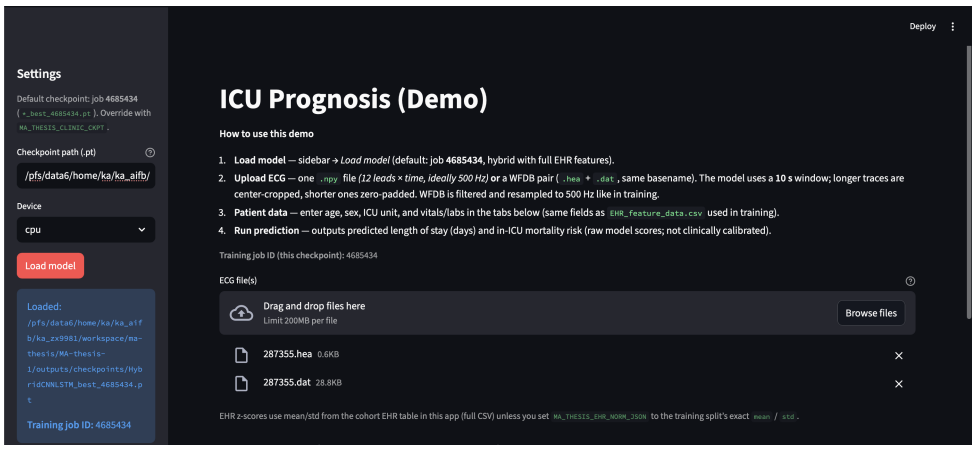

Figure 41: Web app prototype: model loading and ECG upload.

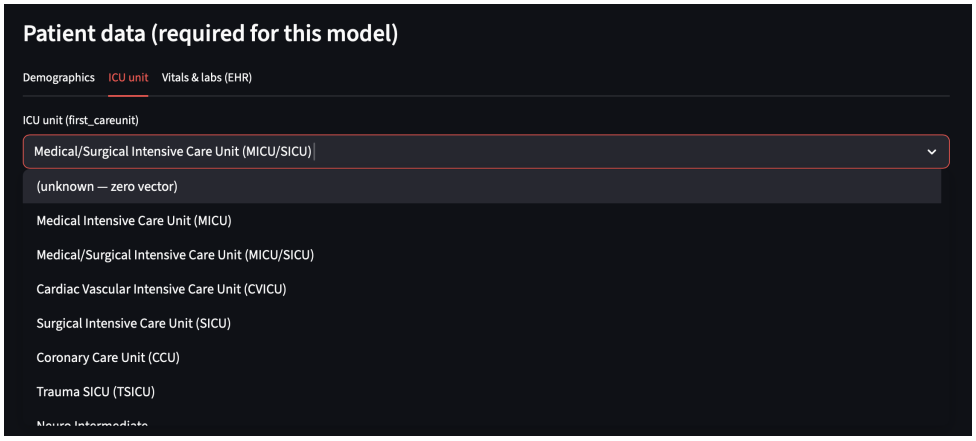

Figure 42: Web app prototype: ICU Units.

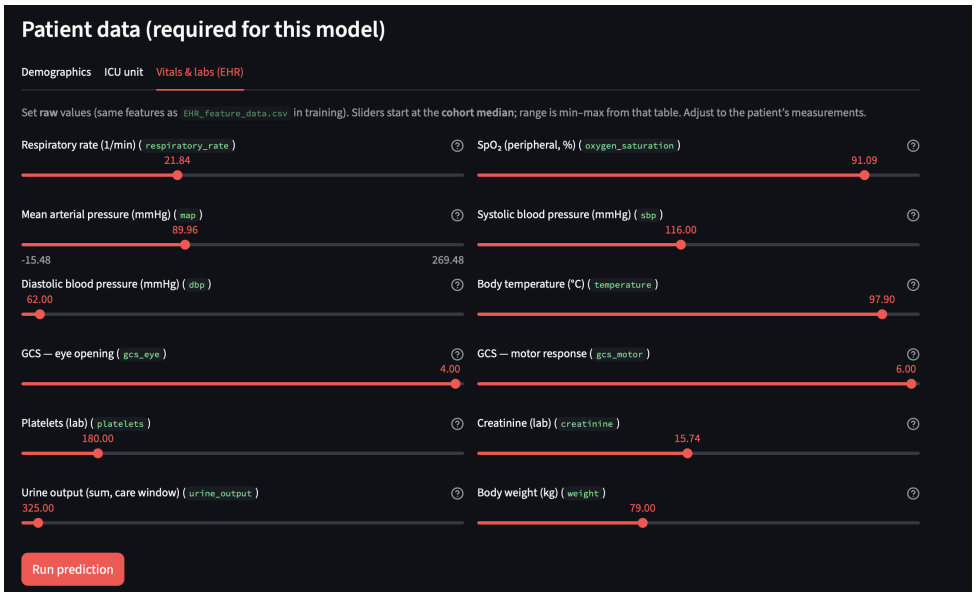

Figure 43: Web app prototype: EHR data selection.

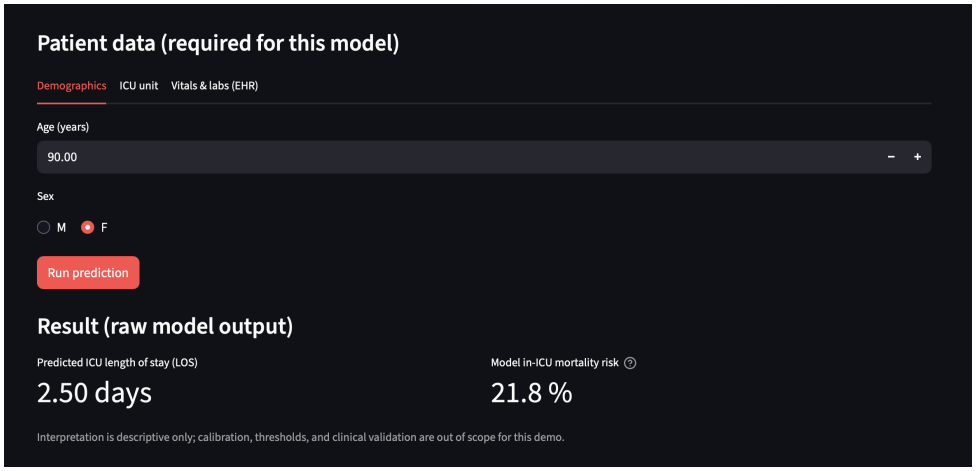

Figure 44: Web app prototype: Age, sex and output.

Declaration of Generative AI Tool Usage

The following table documents the use of generative AI systems during the preparation of this thesis.

Work step	**Used AI system(s)**	**Description of usage**
Literature research	Perplexity ChatGPT	Used as a search engine, to explore literature and test initial questions, approaches, and ideas.
Coding	Cursor	Supported implementation, debugging, and refactoring of project code.
Analysis and evaluation	Cursor	Supported exploratory analysis, evaluation scripts, and review of experimental outputs.
Linguistic formulations	ChatGPT	Supported phrasing, grammar checks, and stylistic refinement of thesis text.

References

[1] B. Hermann, S. Benghanem, Y. Jouan, A. Lafarge, A. Beurton, and the ICU French FOXES (Federation Of eXtremely Enthusiastic Scientists) Study Group, "The positive impact of COVID-19 on critical care: From unprecedented challenges to transformative changes, from the perspective of young intensivists", en, *Annals of Intensive Care*, vol. 13, no. 1, p. 28, Apr. 2023, ISSN: 2110-5820. DOI: 10.1186/s13613-023-01118-9. Accessed: Aug. 28, 2025. [Online]. Available: https://annalsofintensivecare.springeropen.com/articles/10.1186/s13613-023-01118-9.

[2] C. Dziegielewski et al., "Characteristics and resource utilization of high-cost users in the intensive care unit: A population-based cohort study", en, *BMC Health Services Research*, vol. 21, no. 1, p. 1312, Dec. 2021, ISSN: 1472-6963. DOI: 10.1186/s12913-021-07318-y. Accessed: Sep. 12, 2025. [Online]. Available: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-021-07318-y.

[3] K. H. Polderman, A. R. J. Girbes, L. G. Thijs, and R. J. M. Strack Van Schijndel, "Accuracy and reliability of APACHE II scoring in two intensive care units: Problems and pitfalls in the use of APACHE II and suggestions for improvement", en, *Anaesthesia*, vol. 56, no. 1, pp. 47–50, Jan. 2001, ISSN: 0003-2409, 1365-2044. DOI: 10.1046/j.1365-2044.2001.01763.x. Accessed: Mar. 13, 2026. [Online]. Available: https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1046/j.1365-2044.2001.01763.x.

[4] K. Zangmo and B. Khwannimit, "Validating the APACHE IV score in predicting length of stay in the intensive care unit among patients with sepsis", en, *Scientific Reports*, vol. 13, no. 1, p. 5899, Apr. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-33173-4. Accessed: Aug. 28, 2025. [Online]. Available: https://www.nature.com/articles/s41598-023-33173-4.

[5] R. P. Moreno et al., "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission", en, *Intensive Care Medicine*, vol. 31, no. 10, pp. 1345–1355, Oct. 2005, ISSN: 0342-4642, 1432-1238. DOI: 10.1007/s00134-005-2763-5. Accessed: Sep. 10, 2025. [Online]. Available: https://link.springer.com/10.1007/s00134-005-2763-5.

[6] P. G. H. Metnitz et al., "SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description", en, *Intensive Care Medicine*, vol. 31, no. 10, pp. 1336–1344, Oct. 2005, ISSN: 0342-4642,

1432-1238. DOI: 10.1007/s00134-005-2762-6. Accessed: Dec. 5, 2025. [Online]. Available: https://link.springer.com/10.1007/s00134-005-2762-6.

[7] Y. Ruan, D. J. Tan, S.-K. Ng, L. Huang, and M. Feng, "Towards Accurate and Reliable ICU Outcome Prediction: A Multimodal Learning Framework Based on Belief Function Theory using Structured EHRs and Free-Text Notes", en, *Journal of Healthcare Informatics Research*, Nov. 2025, ISSN: 2509-4971, 2509-498X. DOI: 10.1007/s41666-025-00219-3. Accessed: Mar. 13, 2026. [Online]. Available: https://link.springer.com/10.1007/s41666-025-00219-3.

[8] S. Iwase et al., "Prediction algorithm for ICU mortality and length of stay using machine learning", en, *Scientific Reports*, vol. 12, no. 1, p. 12 912, Jul. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-17091-5. Accessed: Sep. 10, 2025. [Online]. Available: https://www.nature.com/articles/s41598-022-17091-5.

[9] M. Bodini, M. W. Rivolta, and R. Sassi, "Opening the black box: Interpretability of machine learning algorithms in electrocardiography", en, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2212, p. 20 200 253, Dec. 2021, ISSN: 1364-503X, 1471-2962. DOI: 10.1098/rsta.2020.0253. Accessed: Feb. 10, 2026. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0253.

[10] E. D'Hondt, T. J. Ashby, I. Chakroun, T. Koninckx, and R. Wuyts, "Identifying and evaluating barriers for the implementation of machine learning in the intensive care unit", en, *Communications Medicine*, vol. 2, no. 1, p. 162, Dec. 2022, ISSN: 2730-664X. DOI: 10.1038/s43856-022-00225-1. Accessed: Feb. 10, 2026. [Online]. Available: https://www.nature.com/articles/s43856-022-00225-1.

[11] K. E. Sandau et al., "Update to Practice Standards for Electrocardiographic Monitoring in Hospital Settings: A Scientific Statement From the American Heart Association", en, *Circulation*, vol. 136, no. 19, Nov. 2017, ISSN: 0009-7322, 1524-4539. DOI: 10.1161/CIR.0000000000000527. Accessed: Mar. 19, 2026. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000527.

[12] J. M. L. Alcaraz, H. Bouma, and N. Strodthoff, "Enhancing clinical decision support with physiological waveforms – a multimodal benchmark in emergency care", 2024, Version Number: 4. DOI: 10.48550/ARXIV.2407.17856. Accessed: Nov. 19, 2025. [Online]. Available: https://arxiv.org/abs/2407.17856.

[13] O. J. Monfredi et al., "Continuous ECG monitoring should be the heart of bedside AI-based predictive analytics monitoring for early detection of clinical deterioration", en, *Journal of Electrocardiology*, vol. 76, pp. 35–38, Jan. 2023, ISSN: 00220736. DOI: 10.1016/j.jelectrocard.2022.10.011. Accessed: Mar. 13,

2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022073622002047.

[14] P. Hadweh et al., "Machine Learning and Artificial Intelligence in Intensive Care Medicine: Critical Recalibrations from Rule-Based Systems to Frontier Models", en, *Journal of Clinical Medicine*, vol. 14, no. 12, p. 4026, Jun. 2025, ISSN: 2077-0383. DOI: 10.3390/jcm14124026. Accessed: Mar. 13, 2026. [Online]. Available: https://www.mdpi.com/2077-0383/14/12/4026.

[15] I. Kose, C. Zincircioglu, Y. K. Ozturk, N. Senoglu, and R. H. Erbay, "Characteristics and Outcomes of Patients withProlonged Stays in an Intensive Care Unit", *Electronic Journal of General Medicine*, vol. 13, no. 2, Apr. 2016, ISSN: 25163507. DOI: 10.15197/ejgm.1546. Accessed: Feb. 12, 2026. [Online]. Available: http://www.ejgm.co.uk/article/characteristics-and-outcomes-of-patients-withprolonged-stays-in-an-intensive-care-unit-7323.

[16] J. I. Salluh and M. Soares, "ICU severity of illness scores: APACHE, SAPS and MPM", en, *Current Opinion in Critical Care*, vol. 20, no. 5, pp. 557–565, Oct. 2014, ISSN: 1070-5295. DOI: 10.1097/MCC.0000000000000135. Accessed: Mar. 13, 2026. [Online]. Available: http://journals.lww.com/00075198-201410000-00015.

[17] S. Balaji, M. Ellenby, J. McNames, and B. Goldstein, "Update on Intensive Care ECG and Cardiac Event Monitoring", en, *Cardiac Electrophysiology Review*, vol. 6, no. 3, pp. 190–195, Sep. 2002, ISSN: 1385-2264, 1573-725X. DOI: 10.1023/A:1016300202560. Accessed: Mar. 13, 2026. [Online]. Available: https://link.springer.com/10.1023/A:1016300202560.

[18] D. E. Becker, "Fundamentals of Electrocardiography Interpretation", en, *Anesthesia Progress*, vol. 53, no. 2, pp. 53–64, Jun. 2006, ISSN: 0003-3006. DOI: 10.2344/0003-3006(2006)53[53:FOEI]2.0.CO;2. Accessed: Aug. 28, 2025. [Online]. Available: https://anesthesiaprogress.kglmeridian.com/view/journals/anpr/53/2/article-p53.xml.

[19] Y. Han, V. Murino, X. Liu, X. Zhang, and C. Ding, *A Systematic Review on Foundation Models for Electrocardiogram Analysis: Initial Strides and Expansive Horizons*, Version Number: 3, 2024. DOI: 10.48550/ARXIV.2410.19877. Accessed: Nov. 20, 2025. [Online]. Available: https://arxiv.org/abs/2410.19877.

[20] Q. Wang et al., "Interpretable Machine Learning Model Integrating Electrocardiographic and Acute Physiology Metrics for Mortality Prediction in Critical Ill Patients", en, *Journal of Clinical Medicine*, vol. 14, no. 20, p. 7163, Oct. 2025, ISSN: 2077-0383. DOI: 10.3390/jcm14207163. Accessed: Mar. 13, 2026. [Online]. Available: https://www.mdpi.com/2077-0383/14/20/7163.

[21] B. P. Bednarski, A. D. Singh, W. Zhang, W. M. Jones, A. Naeim, and R. Ramezani, "Temporal convolutional networks and data rebalancing for clinical length of stay and mortality prediction", en, *Scientific Reports*, vol. 12, no. 1, p. 21 247, Dec. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-25472-z. Accessed: Nov. 26, 2025. [Online]. Available: https://www.nature.com/articles/s41598-022-25472-z.

[22] S. Baker, W. Xiang, and I. Atkinson, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: A hybrid neural network approach", en, *Scientific Reports*, vol. 10, no. 1, p. 21 282, Dec. 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-78184-7. Accessed: Nov. 5, 2025. [Online]. Available: https://www.nature.com/articles/s41598-020-78184-7.

[23] H. D. J. Hogg et al., "Stakeholder Perspectives of Clinical Artificial Intelligence Implementation: Systematic Review of Qualitative Evidence", en, *Journal of Medical Internet Research*, vol. 25, e39742, Jan. 2023, ISSN: 1438-8871. DOI: 10.2196/39742. Accessed: Feb. 12, 2026. [Online]. Available: https://www.jmir.org/2023/1/e39742.

[24] W. G. Anderson et al., "A Multicenter Study of Key Stakeholders' Perspectives on Communicating with Surrogates about Prognosis in Intensive Care Units", en, *Annals of the American Thoracic Society*, vol. 12, no. 2, pp. 142–152, Feb. 2015, ISSN: 2329-6933, 2325-6621. DOI: 10.1513/AnnalsATS.201407-325OC. Accessed: Feb. 12, 2026. [Online]. Available: https://academic.oup.com/annalsats/article/12/2/142/8429422.

[25] S. Jana, T. Dasgupta, and L. Dey, "Predicting medical events and ICU requirements using a multimodal multiobjective transformer network", en, *Experimental Biology and Medicine*, vol. 247, no. 22, pp. 1988–2002, Nov. 2022, ISSN: 1535-3702, 1535-3699. DOI: 10.1177/15353702221126559. Accessed: Mar. 12, 2026. [Online]. Available: https://www.ebm-journal.org/journals/experimental-biology-and-medicine/articles/10.1177/15353702221126559.

[26] B. Gow et al., *MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset*. DOI: 10.13026/B95V-FF39. Accessed: Jan. 3, 2026. [Online]. Available: https://physionet.org/content/mimic-iv-ecg/.

[27] D. Shillan, J. A. C. Sterne, A. Champneys, and B. Gibbison, "Use of machine learning to analyse routinely collected intensive care unit data: A systematic review", en, *Critical Care*, vol. 23, no. 1, p. 284, Dec. 2019, ISSN: 1364-8535. DOI: 10.1186/s13054-019-2564-9. Accessed: Feb. 11, 2026. [Online]. Available: https://ccforum.biomedcentral.com/articles/10.1186/s13054-019-2564-9.

[28] T. Jiang, J. L. Gradus, and A. J. Rosellini, "Supervised Machine Learning: A Brief Primer", en, *Behavior Therapy*, vol. 51, no. 5, pp. 675–687, Sep. 2020, ISSN: 00057894. DOI: 10.1016/j.beth.2020.05.002. Accessed: Mar. 19, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0005789420300678.

[29] S. Tiwari, "Supervised Machine Learning: A Brief Introduction", in *Proceedings of the International Conference on Virtual Learning - VIRTUAL LEARNING - VIRTUAL REALITY (17th edition)*, The National Institute for Research & Development in Informatics - ICI Bucharest (ICI Publishing House), Dec. 2022, pp. 219–230. DOI: 10.58503/icvl-v17y202218. Accessed: Mar. 19, 2026. [Online]. Available: https://icvl.eu/vol-17-2022/supervised-machine-learning-a-brief-introduction/.

[30] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data", en, *Scientific Data*, vol. 6, no. 1, p. 96, Jun. 2019, ISSN: 2052-4463. DOI: 10.1038/s41597-019-0103-9. Accessed: Sep. 5, 2025. [Online]. Available: https://www.nature.com/articles/s41597-019-0103-9.

[31] N. Strodthoff, P. Wagner, T. Schaeffter, and W. Samek, "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL", *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1519–1528, May 2021, ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2020.3022989. Accessed: Jan. 27, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/9190034/.

[32] M. Wu, Y. Lu, W. Yang, and S. Y. Wong, "A Study on Arrhythmia via ECG Signal Classification Using the Convolutional Neural Network", *Frontiers in Computational Neuroscience*, vol. 14, p. 564015, Jan. 2021, ISSN: 1662-5188. DOI: 10.3389/fncom.2020.564015. Accessed: Dec. 3, 2025. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fncom.2020.564015/full.

[33] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, Version Number: 9, 2014. DOI: 10.48550/ARXIV.1412.6980. Accessed: Mar. 20, 2026. [Online]. Available: https://arxiv.org/abs/1412.6980.

[34] Y. Bengio, A. Courville, and P. Vincent, "Representation Learning: A Review and New Perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2013.50. Accessed: Mar. 19, 2026. [Online]. Available: http://ieeexplore.ieee.org/document/6472238/.

[35] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network", en, *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019, ISSN: 1078-8956, 1546-170X. DOI: 10.1038/

s41591-018-0268-3. Accessed: Sep. 10, 2025. [Online]. Available: https://www.nature.com/articles/s41591-018-0268-3.

[36] G. Petmezas et al., "State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review", en, *JMIR Medical Informatics*, vol. 10, no. 8, e38454, Aug. 2022, ISSN: 2291-9694. DOI: 10.2196/38454. Accessed: Dec. 3, 2025. [Online]. Available: https://medinform.jmir.org/2022/8/e38454.

[37] F. Khan, X. Yu, Z. Yuan, and A. U. Rehman, "ECG classification using 1-D convolutional deep residual neural network", en, *PLOS ONE*, vol. 18, no. 4, M. Hammad, Ed., e0284791, Apr. 2023, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0284791. Accessed: Nov. 5, 2025. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0284791.

[38] Y. Ye, K. Chipusu, M. A. Ashraf, B. Ding, Y. Huang, and J. Huang, "Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals", en, *Scientific Reports*, vol. 15, no. 1, p. 34 510, Oct. 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-17671-1. Accessed: Mar. 13, 2026. [Online]. Available: https://www.nature.com/articles/s41598-025-17671-1.

[39] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG Heartbeat Classification: A Deep Transferable Representation", in *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, New York, NY: IEEE, Jun. 2018, pp. 443–444, ISBN: 978-1-5386-5377-7. DOI: 10.1109/ICHI.2018.00092. Accessed: Jan. 3, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/8419425/.

[40] Q. Zhang, L. Fu, and L. Gu, "A Cascaded Convolutional Neural Network for Assessing Signal Quality of Dynamic ECG", en, *Computational and Mathematical Methods in Medicine*, vol. 2019, pp. 1–12, Oct. 2019, ISSN: 1748-670X, 1748-6718. DOI: 10.1155/2019/7095137. Accessed: Dec. 3, 2025. [Online]. Available: https://www.hindawi.com/journals/cmmm/2019/7095137/.

[41] E. Coppola, M. Savardi, M. Massussi, M. Adamo, M. Metra, and A. Signoroni, *HuBERT-ECG as a self-supervised foundation model for broad and scalable cardiac applications*, en, Nov. 2024. DOI: 10.1101/2024.11.14.24317328. Accessed: Nov. 19, 2025. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2024.11.14.24317328.

[42] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review", *Journal of Neural Engineering*, vol. 16, no. 3, p. 031 001, Jun. 2019, ISSN: 1741-2560, 1741-2552. DOI: 10.1088/1741-2552/ab0ab5. Accessed: Dec. 3, 2025. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/ab0ab5.

[43] R. Kamaleswaran, R. Mahajan, and O. Akbilgic, "A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length", *Physiological Measurement*, vol. 39, no. 3, p. 035 006, Mar. 2018, ISSN: 1361-6579. DOI: 10.1088/1361-6579/aaaa9d. Accessed: Dec. 3, 2025. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-6579/aaaa9d.

[44] K. Yu, M. Zhang, T. Cui, and M. Hauskrecht, "Monitoring ICU Mortality Risk with A Long Short-Term Memory Recurrent Neural Network", en, in *Biocomputing 2020*, Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, Dec. 2019, pp. 103–114, ISBN: 978-981-12-1562-9 978-981-12-1563-6. DOI: 10.1142/9789811215636_0010. Accessed: Mar. 13, 2026. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/9789811215636_0010.

[45] S. Bai, J. Z. Kolter, and V. Koltun, *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*, Version Number: 2, 2018. DOI: 10.48550/ARXIV.1803.01271. Accessed: Nov. 26, 2025. [Online]. Available: https://arxiv.org/abs/1803.01271.

[46] Y. Jiang, Q. Li, and W. Zhang, "A Real-Time Signal-Based Wavelet Long Short-Term Memory Method for Length-of-Stay Prediction for the Intensive Care Unit: Development and Evaluation Study", en, *JMIR AI*, vol. 4, e71247–e71247, Aug. 2025, ISSN: 2817-1705. DOI: 10.2196/71247. Accessed: Sep. 5, 2025. [Online]. Available: https://ai.jmir.org/2025/1/e71247.

[47] A. Vaswani et al., *Attention Is All You Need*, arXiv:1706.03762 [cs], Aug. 2023. DOI: 10.48550/arXiv.1706.03762. Accessed: Dec. 3, 2025. [Online]. Available: http://arxiv.org/abs/1706.03762.

[48] A. Natarajan et al., "A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification", Dec. 2020. DOI: 10.22489/CinC.2020.107. Accessed: Dec. 3, 2025. [Online]. Available: http://www.cinc.org/archives/2020/pdf/CinC2020-107.pdf.

[49] M. Y. Ansari et al., "A survey of transformers and large language models for ECG diagnosis: Advances, challenges, and future directions", en, *Artificial Intelligence Review*, vol. 58, no. 9, p. 261, Jun. 2025, ISSN: 1573-7462. DOI: 10.1007/s10462-025-11259-x. Accessed: Mar. 13, 2026. [Online]. Available: https://link.springer.com/10.1007/s10462-025-11259-x.

[50] S. Alyounis, A. Khandoker, C. Stefanini, and L. J. Hadjileontiadis, "A Hybrid CNN-LSTM Model for Heart Failure Detection Using Raw ECG Signals", Dec. 2024. DOI: 10.22489/CinC.2024.197. Accessed: Jan. 29, 2026. [Online]. Available: https://www.cinc.org/archives/2024/pdf/CinC2024-197.pdf.

[51] M. Najia and B. Faouzi, "An Enhanced Hybrid Model Combining CNN, BiLSTM, and Attention Mechanism for ECG Segment Classification", en, *Biomedical Engineering and Computational Biology*, vol. 16, p. 11 795 972 251 341 051, Jun. 2025, ISSN: 1179-5972, 1179-5972. DOI: 10.1177/11795972251341051. Accessed: Mar. 13, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.1177/11795972251341051.

[52] A. J. Weinhaus and K. P. Roberts, "Anatomy of the Human Heart", en, in *Handbook of Cardiac Anatomy, Physiology, and Devices*, P. A. Iaizzo, Ed., Totowa, NJ: Humana Press, 2009, pp. 59–85, ISBN: 978-1-60327-371-8 978-1-60327-372-5. DOI: 10.1007/978-1-60327-372-5_5. Accessed: Mar. 24, 2026. [Online]. Available: http://link.springer.com/10.1007/978-1-60327-372-5_5.

[53] S. Vieau and P. A. Iaizzo, "Basic ECG Theory, 12-Lead Recordings, and Their Interpretation", en, in *Handbook of Cardiac Anatomy, Physiology, and Devices*, P. A. Iaizzo, Ed., Cham: Springer International Publishing, 2015, pp. 321–334, ISBN: 978-3-319-19463-9 978-3-319-19464-6. DOI: 10.1007/978-3-319-19464-6_19. Accessed: Mar. 13, 2026. [Online]. Available: https://link.springer.com/10.1007/978-3-319-19464-6_19.

[54] T. G. Laske, M. Shrivastav, and P. A. Iaizzo, "The Cardiac Conduction System", en, in *Handbook of Cardiac Anatomy, Physiology, and Devices*, P. A. Iaizzo, Ed., Cham: Springer International Publishing, 2015, pp. 215–233, ISBN: 978-3-319-19463-9 978-3-319-19464-6. DOI: 10.1007/978-3-319-19464-6_13. Accessed: Mar. 24, 2026. [Online]. Available: https://link.springer.com/10.1007/978-3-319-19464-6_13.

[55] P. C. Franzone, L. F. Pavarino, and S. Scacchi, *Mathematical cardiac electrophysiology* (MS & A volume 13), eng. Cham: Springer, 2014, ISBN: 978-3-319-04801-7.

[56] *Basiswissen Physiologie* (Springer-Lehrbuch), de. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ISBN: 978-3-540-71401-9. DOI: 10.1007/978-3-540-71402-6. Accessed: Mar. 19, 2026. [Online]. Available: http://link.springer.com/10.1007/978-3-540-71402-6.

[57] V. Ariyarajah and D. H. Spodick, "The Bachmann Bundle and Interatrial Conduction", en, *Cardiology in Review*, vol. 14, no. 4, pp. 194–199, Jul. 2006, ISSN: 1061-5377. DOI: 10.1097/01.crd.0000195221.26979.2b. Accessed: Mar. 24, 2026. [Online]. Available: https://journals.lww.com/00045415-200607000-00010.

[58] G. O. Hofmann, *Medizin für Ingenieure: Grundwissen für Entwickler, Techniker und Wissenschaftler*, de. München: Carl Hanser Verlag GmbH & Co. KG, Apr. 2021, ISBN: 978-3-446-46148-2 978-3-446-46423-0. DOI: 10.3139/9783446464230.

Accessed: Mar. 19, 2026. [Online]. Available: https://www.hanser-elibrary.com/doi/book/10.3139/9783446464230.

[59] J. .-. Vincent et al., "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure: On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine (see contributors to the project in the appendix)", en, *Intensive Care Medicine*, vol. 22, no. 7, pp. 707–710, Jul. 1996, ISSN: 0342-4642, 1432-1238. DOI: 10.1007/BF01709751. Accessed: Mar. 13, 2026. [Online]. Available: http://link.springer.com/10.1007/BF01709751.

[60] T. P. Pellathy, M. R. Pinsky, and M. Hravnak, "Intensive Care Unit Scoring Systems", en, *Critical Care Nurse*, vol. 41, no. 4, pp. 54–64, Aug. 2021, ISSN: 0279-5442, 1940-8250. DOI: 10.4037/ccn2021613. Accessed: Mar. 13, 2026. [Online]. Available: https://aacnjournals.org/ccnonline/article/41/4/54/31512/Intensive-Care-Unit-Scoring-Systems.

[61] J. E. Zimmerman, A. A. Kramer, D. S. McNair, F. M. Malila, and V. L. Shaffer, "Intensive care unit length of stay: Benchmarking based on Acute Physiology and Chronic Health Evaluation (APACHE) IV*:" en, *Critical Care Medicine*, vol. 34, no. 10, pp. 2517–2529, Oct. 2006, ISSN: 0090-3493. DOI: 10.1097/01.CCM.0000240233.01711.D9. Accessed: Mar. 13, 2026. [Online]. Available: http://journals.lww.com/00003246-200610000-00001.

[62] H.-C. Thorsen-Meyer et al., "Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records", en, *The Lancet Digital Health*, vol. 2, no. 4, e179–e191, Apr. 2020, ISSN: 25897500. DOI: 10.1016/S2589-7500(20)30018-2. Accessed: Feb. 2, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589750020300182.

[63] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and P. Rashidi, "DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning", en, *Scientific Reports*, vol. 9, no. 1, p. 1879, Feb. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-38491-0. Accessed: Aug. 28, 2025. [Online]. Available: https://www.nature.com/articles/s41598-019-38491-0.

[64] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. Clifton, and G. D. Clifford, "Machine Learning and Decision Support in Critical Care", *Proceedings of the IEEE*, vol. 104, no. 2, pp. 444–466, Feb. 2016, ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2015.2501978. Accessed: Sep. 10, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/7390351/.

[65] J. Wu, Y. Lin, P. Li, Y. Hu, L. Zhang, and G. Kong, "Predicting Prolonged Length of ICU Stay through Machine Learning", en, *Diagnostics*, vol. 11, no. 12, p. 2242, Nov. 2021, ISSN: 2075-4418. DOI: 10.3390/diagnostics11122242. Accessed: Nov. 19, 2025. [Online]. Available: https://www.mdpi.com/2075-4418/11/12/2242.

[66] E.-T. Jeon et al., "Machine learning-based prediction of in-ICU mortality in pneumonia patients", en, *Scientific Reports*, vol. 13, no. 1, p. 11 527, Jul. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-38765-8. Accessed: Aug. 28, 2025. [Online]. Available: https://www.nature.com/articles/s41598-023-38765-8.

[67] D. Castiñeira et al., "Adding Continuous Vital Sign Information to Static Clinical Data Improves the Prediction of Length of Stay After Intubation: A Data-Driven Machine Learning Approach", en, *Respiratory Care*, vol. 65, no. 9, pp. 1367–1377, Sep. 2020, ISSN: 0020-1324. DOI: 10.4187/respcare.07561. Accessed: Mar. 13, 2026. [Online]. Available: https://journals.sagepub.com/doi/10.4187/respcare.07561.

[68] K. McKeen, S. Masood, A. Toma, B. Rubin, and B. Wang, "ECG-FM: An open electrocardiogram foundation model", en, *JAMIA Open*, vol. 8, no. 5, ooaf122, Sep. 2025, ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooaf122. Accessed: Mar. 13, 2026. [Online]. Available: https://academic.oup.com/jamiaopen/article/doi/10.1093/jamiaopen/ooaf122/8287827.

[69] A. Nolin-Lapalme et al., "Foundation models for electrocardiogram interpretation: Clinical implications", en, *European Heart Journal*, ehaf1119, Jan. 2026, ISSN: 0195-668X, 1522-9645. DOI: 10.1093/eurheartj/ehaf1119. Accessed: Mar. 13, 2026. [Online]. Available: https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaf1119/8436833.

[70] W. Sun et al., "Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms", en, *npj Digital Medicine*, vol. 6, no. 1, p. 21, Feb. 2023, ISSN: 2398-6352. DOI: 10.1038/s41746-023-00765-3. Accessed: Dec. 3, 2025. [Online]. Available: https://www.nature.com/articles/s41746-023-00765-3.

[71] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset", en, *Scientific Data*, vol. 10, no. 1, p. 1, Jan. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01899-x. Accessed: Sep. 10, 2025. [Online]. Available: https://www.nature.com/articles/s41597-022-01899-x.

[72] A. Salimi, S. V. Kalmady, A. Hindle, O. Zaiane, and P. Kaul, *Exploring Best Practices for ECG Pre-Processing in Machine Learning*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2311.04229. Accessed: Nov. 28, 2025. [Online]. Available: https://arxiv.org/abs/2311.04229.

[73] D. Ricciardi et al., "Impact of the high-frequency cutoff of bandpass filtering on ECG quality and clinical interpretation: A comparison between 40 Hz and 150 Hz cutoff in a surgical preoperative adult outpatient population", en, *Journal of Electrocardiology*, vol. 49, no. 5, pp. 691–695, Sep. 2016, ISSN: 00220736. DOI: 10.1016/j.jelectrocard.2016.07.002. Accessed: Mar. 13, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022073616300802.

[74] M. M. Rahman, M. W. Rivolta, F. Badilini, and R. Sassi, "A Systematic Survey of Data Augmentation of ECG Signals for AI Applications", en, *Sensors*, vol. 23, no. 11, p. 5237, May 2023, ISSN: 1424-8220. DOI: 10.3390/s23115237. Accessed: Mar. 30, 2026. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5237.

[75] N. Nonaka and J. Seita, *Data Augmentation for Electrocardiogram Classification with Deep Neural Network*, Version Number: 1, 2020. DOI: 10.48550/ARXIV.2009.04398. Accessed: Mar. 30, 2026. [Online]. Available: https://arxiv.org/abs/2009.04398.

[76] Y. Ansari, O. Mourad, K. Qaraqe, and E. Serpedin, "Deep learning for ECG Arrhythmia detection and classification: An overview of progress for period 2017–2023", *Frontiers in Physiology*, vol. 14, p. 1 246 746, Sep. 2023, ISSN: 1664-042X. DOI: 10.3389/fphys.2023.1246746. Accessed: Nov. 26, 2025. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fphys.2023.1246746/full.

[77] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, "A lightweight hybrid CNN-LSTM explainable model for ECG-based arrhythmia detection", en, *Biomedical Signal Processing and Control*, vol. 90, p. 105 884, Apr. 2024, ISSN: 17468094. DOI: 10.1016/j.bspc.2023.105884. Accessed: Mar. 13, 2026. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809423013174.

[78] C. Sudlow et al., "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age", en, *PLOS Medicine*, vol. 12, no. 3, e1001779, Mar. 2015, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1001779. Accessed: May 16, 2026. [Online]. Available: https://dx.plos.org/10.1371/journal.pmed.1001779.

[79] M. A. Al-Masud, J. M. L. Alcaraz, and N. Strodthoff, *Benchmarking ECG Foundational Models: A Reality Check Across Clinical Tasks*, Version Number: 1, 2025. DOI: 10.48550/ARXIV.2509.25095. Accessed: Nov. 20, 2025. [Online]. Available: https://arxiv.org/abs/2509.25095.

[80] S. R. Stahlschmidt, B. Ulfenborg, and J. Synnergren, "Multimodal deep learning for biomedical data fusion: A review", en, *Briefings in Bioinformatics*, vol. 23, no. 2, bbab569, Mar. 2022, ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbab569.

Accessed: Apr. 23, 2026. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab569/6516346.

[81] C. Cui et al., "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review", *Progress in Biomedical Engineering*, vol. 5, no. 2, p. 022 001, Apr. 2023, ISSN: 2516-1091. DOI: 10.1088/2516-1091/acc2fe. Accessed: Apr. 23, 2026. [Online]. Available: https://iopscience.iop.org/article/10.1088/2516-1091/acc2fe.

[82] S. Kumar, S. Rani, S. Sharma, and H. Min, "Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review", en, *Bioengineering*, vol. 11, no. 12, p. 1233, Dec. 2024, ISSN: 2306-5354. DOI: 10.3390/bioengineering11121233. Accessed: Apr. 23, 2026. [Online]. Available: https://www.mdpi.com/2306-5354/11/12/1233.

[83] J. Takala et al., "Variation in severity-adjusted resource use and outcome in intensive care units", en, *Intensive Care Medicine*, vol. 48, no. 1, pp. 67–77, Jan. 2022, ISSN: 0342-4642, 1432-1238. DOI: 10.1007/s00134-021-06546-4. Accessed: Feb. 12, 2026. [Online]. Available: https://link.springer.com/10.1007/s00134-021-06546-4.

[84] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A Next-generation Hyperparameter Optimization Framework*, Version Number: 1, 2019. DOI: 10.48550/ARXIV.1907.10902. Accessed: Apr. 8, 2026. [Online]. Available: https://arxiv.org/abs/1907.10902.

[85] Y. Hu, L. Zheng, and J. Wang, "Predicting ICU Length of Stay for Patients with Diabetes Using Machine Learning Techniques", in *2022 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, Nanjing, China: IEEE, Nov. 2022, pp. 417–422, ISBN: 978-1-6654-9835-7. DOI: 10.1109/ICCSI55536.2022.9970666. Accessed: Mar. 13, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/9970666/.

[86] D. Reddy Dade, J. Bergquist, R. MacLeod, R. Ranjan, B. A. Steinberg, and T. Tasdizen, "A Survey of Augmentation Techniques for Enhancing ECG Representation through Self-Supervised Contrastive Learning", Dec. 2024. DOI: 10.22489/CinC.2024.223. Accessed: Mar. 30, 2026. [Online]. Available: https://www.cinc.org/archives/2024/pdf/CinC2024-223.pdf.