FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

MedicAI-driven Generative Modelling for ICU

Vasco Silva

Mestrado em Engenharia Biomédica

Supervisor: Dr. Ricardo Santos

Co-Supervisor: Prof. Carlos Soares

June 21, 2026

Resumo

A crescente disponibilidade de dados de unidades de cuidados intensivos cria oportunidades para o apoio à decisão clínica baseado em inteligência artificial, mas introduz desafios relacionados com heterogeneidade, dados em falta, irregularidade temporal, multimodalidade e generalização entre bases de dados. Esta dissertação aborda estes desafios através da revisão de pipelines de harmonização de dados de cuidados intensivos e de modelos de inteligência artificial clínica, bem como da implementação de uma metodologia de benchmarking para modelação preditiva com dados estruturados em diferentes bases de dados e tarefas clínicas.

A revisão de 32 estudos sobre pipelines de estruturação e harmonização de dados mostrou melhorias na reprodutibilidade e no processamento de diferentes conjuntos de dados, mas também uma frequente falta de integração com modelos generativos e sistemas de apoio ao raciocínio clínico. Por sua vez, a revisão de 23 estudos sobre inteligência artificial generativa revelou que as soluções baseadas em modelos de linguagem estavam frequentemente pouco integradas em pipelines de dados clínicos reutilizáveis. Esta lacuna motivou a avaliação da reutilização dos dados harmonizados em modelos diferentes dos originalmente previstos.

A pipeline METRE foi adaptada e alargada às bases de dados MIMIC-IV e eICU através da reconstrução de componentes de processamento indisponíveis, da geração de representações longitudinais preparadas para modelação e da implementação de um protocolo unificado de treino e avaliação. O benchmark incluiu modelos de referência de aprendizagem automática, arquiteturas de redes neuronais derivadas do METRE e uma adaptação da arquitetura externa EMERGE baseada em dados longitudinais estruturados de registos eletrónicos de saúde. A avaliação combinou validação cruzada estratificada com 10 folds, otimização Bayesiana de hiperparâmetros, conjuntos de teste internos e externos e intervalos de confiança obtidos por bootstrap.

Os resultados demonstraram que a otimização bayesiana de hiperparâmetros melhorou várias configurações modelo-tarefa relativamente às configurações de referência propostas pelos autores dos benchmarks originais, embora os ganhos não tenham sido uniformes. A mortalidade foi a tarefa mais robusta e transferível, enquanto a insuficiência respiratória aguda e o choque se mostraram mais sensíveis às diferenças entre bases de dados e às variáveis relacionadas com tratamentos. O modelo Random Forest manteve um desempenho competitivo face a modelos mais complexos, sugerindo que os dados harmonizados preservaram informação preditiva relevante. A adaptação do EMERGE demonstrou que os dados processados pelo METRE podiam suportar o ramo longitudinal estruturado de uma arquitetura externa. A avaliação externa revelou uma degradação do desempenho em várias tarefas, confirmando que a validação interna, por si só, é insuficiente para avaliar a robustez clínica. Globalmente, esta dissertação apresenta uma revisão crítica e uma metodologia reprodutível de benchmarking aplicada a diferentes bases de dados, tarefas e famílias de modelos.

Palavras-chave: Unidade de Cuidados Intensivos, Inteligência Artificial, Registos Eletrónicos de Saúde, Harmonização de Dados, Modelação Preditiva Clínica, Avaliação entre Bases de Dados, Benchmarking.

Abstract

The increasing availability of intensive care unit data creates opportunities for artificial intelligence-based clinical decision support, but also introduces challenges related to heterogeneity, missingness, temporal irregularity, multimodality, and cross-database generalisation. This dissertation addresses these challenges by reviewing intensive care data harmonisation pipelines and clinical artificial intelligence models, and by developing a benchmark-oriented workflow for structured prediction across databases and multiple clinical tasks.

The review of 32 studies on data structuring and harmonisation pipelines showed improvements in reproducibility and cross-dataset processing, but also a frequent disconnection from downstream generative and reasoning-based models. Conversely, the review of 23 generative artificial intelligence studies found that language-model approaches were often weakly connected to reusable clinical data pipelines. This gap motivated the exploration of whether harmonised pipeline outputs could support models beyond those originally included in the pipeline.

The open-source METRE pipeline was adapted and extended for MIMIC-IV and eICU by reconstructing unavailable processing components, generating model-ready longitudinal representations, and implementing a unified model training and evaluation protocol. The benchmark included baseline machine-learning models, METRE-derived neural-network architectures, and an adapted external EMERGE architecture using structured longitudinal electronic health record data. Evaluation combined stratified 10-fold cross-validation, Bayesian hyperparameter optimisation, internal and external test sets, and bootstrap confidence intervals.

The results showed that Bayesian hyperparameter optimisation improved several model-task configurations relative to the reference configurations proposed by the original benchmark authors, although the gains were not uniform. Mortality was the most robust and transferable task, whereas acute respiratory failure and shock were more sensitive to database shift and treatment-related variables. Random Forest remained competitive with more complex models, indicating that the harmonised representation retained relevant predictive information. The EMERGE adaptation demonstrated that METRE data processing outputs could support the structured longitudinal-data branch of an external model architecture. External evaluation revealed performance degradation across several tasks, confirming that internal validation alone is insufficient to assess clinical robustness. Overall, this dissertation provides a critical review and a reproducible benchmarking workflow spanning databases, tasks, and model families.

Keywords: Intensive Care Unit, Artificial Intelligence, Electronic Health Records, Data Harmonisation, Clinical Predictive Modelling, Cross-database Evaluation, Benchmarking.

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework for promoting social, economic, and environmental development. The framework comprises 17 goals[1] addressing major global challenges, including health, education, inequality, climate change, innovation, peace, and justice.

This dissertation is primarily aligned with SDG 3, Good Health and Well-being, and SDG 9, Industry, Innovation and Infrastructure. Its contribution is methodological and research-oriented rather than directly clinical. Although the developed workflow was not deployed as a clinical system, this work investigates data harmonisation, predictive modelling, benchmarking, and cross-database generalisation in intensive care. These contributions may support the future development of more robust and transferable early-warning and risk-prediction systems for critical care.

The specific Sustainable Development Goals considered are:

SDG 3 Ensure healthy lives and promote well-being for all at all ages.

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialisation and foster innovation.

[1] https://sdgs.un.org/goals

SDG	Target	Contribution	Performance Indicators and Metrics
3	3.d	The dissertation evaluates predictive models for clinically relevant intensive care outcomes, including hospital mortality, acute respiratory failure, and shock. This contributes to research on early-warning and clinical risk-prediction systems.	Predictive performance measured using AUROC, AUPRC, accuracy, precision, sensitivity, specificity, F1-score, and $\min(+P, Se)$; Number and type of clinical prediction tasks evaluated.
	3.d	The study assesses model generalisation across independent Intensive Care Unit (ICU) databases, providing evidence on the robustness and transferability of predictive approaches beyond the database used for model development.	Internal and external test performance; Differences between internal and external evaluation metrics; Identification of tasks and models with reduced cross-database transferability.
9	9.5	The dissertation adapts a multidatabase ICU extraction and harmonisation pipeline and establishes a standardised experimental workflow for training, validating, and comparing multiple predictive models.	Number of databases processed; Number of clinical tasks and model families evaluated; Consistency of cohort construction and preprocessing across databases; Documentation of experimental configurations and results.
	9.5	The work investigates whether harmonised representations generated by Multidatabase ExTRaction PipEline (METRE) can support models beyond those originally included in the pipeline, through the adaptation and evaluation of the EMERGE time-series architecture.	Integration of the adapted architecture with METRE-generated data; Comparative performance of classical machine-learning baselines, neural-network models, and the adapted external architecture under a common evaluation protocol.

Agradecimentos

Por todo o trabalho desenvolvido, gostaria de expressar a minha profunda gratidão a todos aqueles que contribuíram para a sua realização e conclusão.

Agradeço, em primeiro lugar, à FEUP e à Fraunhofer Portugal AICOS pela oportunidade de realizar este estágio. Deixo também um sincero agradecimento à Professora Ana Maria Mendonça, pelo seu compromisso, disponibilidade e prontidão na resposta sempre que foi necessário.

Ao meu orientador na instituição, Doutor Ricardo Santos, deixo um agradecimento muito especial. Ao longo deste ano e meio, a sua orientação foi muito além do conhecimento e da experiência que partilhou comigo. Agradeço-lhe toda a disponibilidade, atenção e paciência, bem como a forma sempre próxima e calorosa com que acompanhou o meu trabalho. Este percurso permitiu-me crescer a nível profissional e guardarei com grande estima tudo aquilo que aprendi consigo.

Agradeço profundamente aos meus pais e à minha irmã, que me apoiaram e, não menos importante, me aturaram ao longo de todo este percurso. Obrigado por acreditarem em mim mesmo nos momentos em que eu próprio duvidei, por celebrarem cada conquista e por serem uma fonte constante de motivação, segurança e força. Esta realização também vos pertence.

Aos meus avós, Teresinha, Zé e Aureliano, deixo um agradecimento muito especial. Mesmo sem compreenderem inteiramente tudo aquilo em que eu estava a trabalhar, nunca deixaram de me apoiar, incentivar e perguntar como tudo estava a correr. Inspiraram-me sempre a trabalhar mais e melhor, quanto mais não fosse para terminar depressa e poder visitá-los mais vezes. À minha avó Lu, que sei que sempre me acompanhou e olhou por mim lá de cima, dedico também esta conquista com muita saudade.

Deixo um agradecimento muito especial à minha namorada, Mariana, sem a qual este caminho certamente não teria sido o mesmo. Ao longo destes muitos anos, esteve sempre ao meu lado, a apoiar-me e a levantar-me nos momentos em que mais duvidei de mim e das minhas capacidades. No final, juntos, como sempre, aqui estamos, a concluir mais uma aventura.

Não poderia também deixar de agradecer a todos os amigos que fiz na FEUP e àqueles cuja amizade trouxe comigo do ISEP. Fizeram com que estes últimos dois anos fossem muito mais do que apenas mais uma etapa académica. Nas pequenas interações do dia a dia e nos momentos académicos mais especiais, estiveram sempre presentes, tornando este percurso mais leve, feliz e memorável.

Por fim, agradeço ao meu fiel amigo Sancho Henriques, que, já pela segunda vez, parece saber exatamente quando preciso dele. Durante as longas horas dedicadas à elaboração desta dissertação, esteve sempre aqui comigo, transmitindo-me a serenidade e a paz que só um verdadeiro amigo de quatro patas consegue oferecer.

Obrigado a todos.

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AKI	Acute Kidney Injury
AMR	Antimicrobial Resistance
APACHE	Acute Physiology and Chronic Health Evaluation
API	Application Programming Interface
ARF	Acute Respiratory Failure
AUROC	Area under the Receiver Operating Characteristic Curve
AUPRC	Area under the Precision–Recall Curve
BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
CARE-AD	Collaborative Analysis & Risk Evaluation for Alzheimer's Disease
CCC	Clinical Care Classification
CDC	Centers for Disease Control and Prevention
CDM	Common Data Model
CDSL	COVID Data for Shared Learning
CHR	County Health Rankings
CLEAR	Clinical Entity Augmented Retrieval
CLIF	Common Longitudinal ICU Format
CNN	Convolutional Neural Network
CoT	Chain of Thought
COVID	COrona VIrus Disease
CPLLM	Clinical Prediction with Large Language Models
CPT	Current Procedural Terminology
CSV	Comma-Separated Values
CT	Computed Tomography
CTGAN	Conditional Tabular GAN
DDW	Dutch Data Warehouse
DFM	Document-Feature Matrix
DL	Deep Learning
DSS	Decision Support System

ECG	Electrocardiogram
ECTS	European Credit Transfer and Accumulation System
EHDEN	European Health Data and Evidence Network
EHR	Electronic Health Record
eICU	electronic Intensive Care Unit
EMERGE	Electronic Medical Records and Genomics Network
EMR	Electronic Medical Record
EPS	Electronic Health Record Performance Score
ETL	Extract, Transform, and Load
FAIR	Findability, Accessibility, Interoperability, and Reusability
FEUP	Faculty of Engineering of the University of Porto
FHIR	Fast Healthcare Interoperability Resources
FIDDLE	Flexible Data-Driven Pipeline
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
HAIL	Hierarchical Attention-based Integrated Learning
HDF5	Hierarchical Data Format, Version 5
HIPAA	Health Insurance Portability and Accountability Act
HiRID	High time-Resolution Intensive Care Database
HIS	Hospital Information System
HL7	Health Level Seven
HPC	High-Performance Computing
ICD	International Classification of Diseases
ICU	Intensive Care Unit
KDIGO	Kidney Disease Improving Global Outcomes
K-MIMIC	Korea Medical Information Mart for Intensive Care
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
LOS	Length of Stay
LR	Logistic Regression
LSTM	Long Short-Term Memory
MB	Megabyte
mCIDE	minimum set of Common ICU Data Elements
MCMC	Markov Chain Monte Carlo
MDT	Multidisciplinary Team

MEB	Master's in Biomedical Engineering
MEREDITH	Medical Evidence Retrieval and Data Integration for Tailored Healthcare
METRE	Multidatabase ExTRaction PipEline
MFSL	Multimodal Few-Shot Learning
MIMIC	Medical Information Mart for Intensive Care
ML	Machine Learning
MRSA	Methicillin-Resistant *Staphylococcus aureus*
MSDM	Minimal Sepsis Data Model
MTB	Molecular Tumour Board
NB	Naive Bayes
NLP	Natural Language Processing
NN	Neural Network
NPY	NumPy Array
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership
PCA	Principal Component Analysis
PDMS	Patient Data Management System
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval Augmented Generation
REST	REpresentational State Transfer
RETAIN	REverse Time AttentIoN
RF	Random Forest
RNN	Recurrent Neural Network
SCRIPT	Successful Clinical Response In Pneumonia Therapy
SDOH	Social Determinants of Health
SHAP	SHapley Additive exPlanations
SNOMED CT	Systematized Nomenclature of Medicine Clinical Terms
SOFA	Sequential Organ Failure Assessment
SPD	Statistical Parity Difference
SVI	Social Vulnerability Index
SptEHR	Soft Prompt Transfer for Electronic Health Records
SQL	Structured Query Language
STARR	STAnford Research Repository
SVM	Support Vector Machine
TCN	Temporal Convolutional Network
TF-IDF	Term Frequency-Inverse Document Frequency
TJH	Tongji Hospital

TS	Time Series
UMC	University Medical Center
US	United States
VaDeSC	Variational deep survival clustering
VHA	Veterans Health Administration
WHO	World Health Organization
XGBoost	Extreme Gradient Boosted
XST-GCNN	eXplainable Spatio-Temporal Graph Convolutional Neural Network
YAIB	Yet Another ICU Benchmark

Chapter 1

Introduction

The project presented in this dissertation was developed with the support of Fraunhofer Portugal Research and carried out at Fraunhofer Portugal AICOS, in Porto. This work was conducted within the scope of the Master's in Biomedical Engineering (MEB) at the Faculty of Engineering of the University of Porto (FEUP), corresponding to the dissertation developed during the second year of the programme and formally completed through the 30-European Credit Transfer and Accumulation System (ECTS) curricular unit Dissertation.

This first chapter presents the background and motivation of the study, as well as the objectives and structure of the dissertation.

1.1 Background

Over the last decades, the widespread adoption of Electronic Health Records (EHRs) has fundamentally transformed healthcare delivery and clinical data availability [1], [2]. EHR systems store large volumes of heterogeneous patient data, including demographic information, diagnoses, laboratory results, medication records, imaging reports, and unstructured clinical notes [2]. Although initially developed to support administrative efficiency and continuity of care, EHRs have become a central resource for secondary data use, enabling large-scale medical informatics research and data-driven clinical analysis [1], [3].

Within this high data volume ecosystem, Decision Support Systems (DSSs) have emerged as a key technological instrument to assist clinicians in managing complex clinical information. Early DSSs originated from symbolic and rule-based approaches aimed at replicating elements of clinical reasoning [4]. Over time, these systems evolved towards data-driven paradigms, increasingly relying on statistical and Machine Learning (ML) methods to uncover latent associations within clinical data and to support, rather than replace, clinical judgement [5], [6]. Formally, DSSs are defined as computer-based systems that combine patient-specific data with medical knowledge to improve clinical decision-making at the point of care [4].

Despite demonstrated benefits in improving diagnostic accuracy, reducing medical errors, and optimising healthcare resource utilisation, the large-scale adoption of clinical DSSs remains limited [7]. Persistent challenges include the lack of standardised medical terminologies, semantic heterogeneity across health information systems, limited interoperability, and difficulties in transferring and maintaining clinical knowledge across institutional boundaries [7]–[9]. These barriers are particularly relevant to the development of data-driven healthcare systems, in which clinical data are continuously generated, analysed, and reintegrated into practice to support evidence-based improvement [10], [11].

These challenges are amplified in Intensive Care Units (ICUs), which represent some of the most data-intensive and time-critical environments in modern healthcare [12]. ICU patients require continuous monitoring and rapid intervention, generating high-frequency, multimodal data streams that include physiological signals, laboratory measurements, ventilator settings, and clinical observations [13]. While this abundance of data creates opportunities for advanced analytics and predictive modelling, it also imposes significant demands on data organisation, real-time interpretation, and clinical reliability [14]. As a result, ICUs exemplify both the potential and the unresolved limitations of data-driven clinical decision support.

In recent years, Artificial Intelligence (AI), particularly Deep Learnings (DLs)–based methods, has emerged as a promising enabler of next-generation clinical DSSs [15]. AI systems are capable of learning complex, non-linear relationships from high-dimensional clinical data and have demonstrated superior performance over traditional statistical approaches in several critical care prediction tasks [12], [16]. These developments highlight the growing role of AI in supporting clinical decision-making, while simultaneously exposing unresolved challenges related to data quality, interpretability, and real-world deployment in safety-critical environments such as the ICU.

1.2 Motivation

The convergence of large-scale EHR data, clinical DSSs, and advanced AI techniques creates a compelling opportunity to improve decision-making in critical care. This opportunity is particularly salient in ICUs, where clinical decisions must often be made under extreme time pressure, high uncertainty, and substantial cognitive load, with direct consequences for patient outcomes [12], [13]. In such settings, even marginal improvements in situational awareness, early warning, or treatment optimisation can have significant clinical impact [14].

At the same time, the increasing availability of rich and heterogeneous ICU data remains largely underexploited in routine practice [12], [16]. Raw data alone does not translate into clinical insights without robust mechanisms for data harmonisation, temporal alignment, and meaningful representation. Many existing AI-driven approaches rely on simplified data abstractions, retrospective validation, or task-specific pipelines that limit generalisability and clinical relevance [7], [16]. As a result, promising predictive performance reported in experimental settings often fails to translate into dependable, real-time decision support in real-world ICU workflows.

Furthermore, persistent challenges related to data heterogeneity, interoperability, transparency, and ethical deployment continue to hinder the integration of AI-based DSSs into everyday clinical practice [7]–[9]. ICU data combine multiple modalities, variable sampling frequencies, and irregular temporal patterns, complicating both data harmonisation and model robustness [12]. Without principled frameworks that address these challenges holistically, AI systems risk remaining isolated research artefacts rather than clinically trustworthy tools.

This dissertation is motivated by the need to bridge this gap between methodological advances in AI and the practical realities of ICU decision support. This requires moving beyond model development alone and addressing how ICU data are structured, harmonised, and transformed into representations that can support reliable, interpretable, and adaptable clinical decision-making.

1.3 Objectives

The primary objective of this thesis is to investigate how state-of-the-art DL architectures can enhance clinical decision-making by effectively exploring the complexity of ICU data. To achieve this goal, this work will focus on the following practical objectives:

- Conduct a review of the state of the art on ICU medical data, analysing their characteristics, limitations, and existing structuring pipelines to identify requirements for data harmonisation and integration;
- Conduct a review on recent advances on Generative AI models and their adaptation to structured and multimodal medical data;
- Process and harmonise multiple open-access ICU databases into a unified and standardised structure suitable for model training, validation, and cross-dataset comparison;
- Explore, implement, and train novel architectures tailored for ICU data, toward improving the robustness of AI-based clinical DSS when operating on complex ICU data.
- Validate the proposed models across benchmark prediction tasks relevant to critical care using multiple datasets.

The first stage of this work, conducted during the first semester as part of the Dissertation Project curricular unit and corresponding to 18 ECTS, focused on the literature review and theoretical grounding of the dissertation. The remaining objectives were addressed during the second semester, within the 30 ECTS Dissertation curricular unit, through the adaptation of the selected data pipeline, implementation of the modelling workflow, and evaluation of the selected models across internal and external test settings.

1.4 Document Structure

This dissertation is organised into six main chapters, followed by references and appendices.

Chapter 1, Introduction, presents the background and motivation of the study, defines the research objectives, and outlines the structure of the dissertation.

Chapter 2, Theoretical Foundations, introduces the conceptual basis required to understand the work developed in the dissertation. It covers the characteristics of ICU medical data, clinical data representation and standardisation, AI in clinical predictive modelling, and the main modelling and evaluation concepts used throughout the methodology and results chapters.

Chapter 3, Literature Review, presents the state of the art in two research directions relevant to this work. First, it reviews medical ICU data structuring pipelines, with emphasis on data extraction, preprocessing, harmonisation, and cross-dataset reuse. Second, it analyses recent generative AI and Transformer-based models applied to medical data, including structured, longitudinal, and multimodal EHR settings. The chapter concludes with a critical synthesis of the reviewed literature and identifies the methodological gaps that motivate the proposed work.

Chapter 4, Methodology, describes the experimental workflow implemented in this dissertation. It presents the selected data sources, the adaptation of the Multidatabase ExTRaction PipEline (METRE) pipeline, the definition of the prediction tasks, the implementation of baseline and METRE-derived models, the adaptation of the external Electronic Medical Records and Genomics Network (EMERGE) Time Series (TS)-only model, and the unified evaluation protocol used for internal validation and cross-database testing.

Chapter 5, Results and Discussion, presents and analyses the experimental results. It begins with the cohort and task composition, then discusses internal validation, hyperparameter optimisation, final internal and external test performance, model comparison, and the impact of cross-database evaluation. The chapter concludes by summarising the main findings and outlining future research directions.

The final chapter 6, Conclusion, summarises the main contributions of the dissertation, recaps its limitations, and highlights possible extensions of the work. Finally, the appendices provide supplementary material, including auxiliary Structured Query Language (SQL) queries, additional evaluation tables, and hyperparameter information, while the References section lists all cited sources.

Chapter 2

Theoretical Foundations

This chapter introduces the fundamental concepts required to understand the methodological and experimental work developed in this dissertation. It begins by characterising medical data collected in ICUs, with emphasis on their main sources, data types, temporal structure, heterogeneity, and quality limitations. These characteristics are central to the challenges addressed in this work, since they directly affect data harmonisation, model development, and cross-database evaluation.

The chapter then discusses clinical data representation and interoperability, distinguishing between structured, unstructured, and multimodal clinical data, and introducing the main terminologies, exchange standards, and Common Data Models (CDMs) relevant to secondary use of healthcare data. These concepts provide the foundation for understanding the role of harmonisation pipelines and the difficulties involved in reusing ICU data across institutions and databases.

Finally, the chapter presents the AI and predictive modelling concepts used throughout the dissertation. It defines the relationship between AI, ML, DL, Generative AI, and Large Language Models (LLMs), introduces the main learning paradigms, and describes the clinical predictive modelling workflow, including data pre-processing, model training, hyperparameter optimisation, and performance evaluation. The models most relevant to this dissertation are also introduced, including classical ML baselines, Recurrent Neural Networks (RNNs), Temporal Convolutional Networks (TCNs), and Transformer encoders. Together, these concepts establish the theoretical basis for the literature review, methodology, and results presented in the following chapters.

2.1 Characteristics and Limitations of Medical data from the ICU

Largely driven by the widespread adoption of EHR systems, digital data collection during routine clinical practice in hospitals has substantially increased, resulting in exponential data generation within the healthcare sector [17], [18]. For those reasons, the ICU stands out and represents a particularly data-intensive environment, where patients require continuous monitoring and interventions, producing highly complex and multifaceted clinical information [17]. Under these conditions this environment presents both logistical challenges and significant opportunities for the development of DSSs [18].

This section therefore describes the main sources and types of ICU data, followed by the principal limitations that affect their secondary use in predictive modelling and decision support.

2.1.1 Sources and Types of ICU Data

The ICU constitutes a data-rich environment that continuously generates highly detailed and specific patient information, making it an optimal setting for data science applications, such as AI [19], [20]. According to M. Imhoff [21], data acquisition in the ICU primarily arises from four major sources:

- Bedside Medical Devices, that produce high-frequency physiological signals, such as Electrocardiograms (ECGs), arterial blood pressure, oxygen saturation, and respiratory waveforms;
- Hospital Information System (HIS), which maintains administrative records, patient EHRs, diagnostic information, laboratory results, and treatment data, serving as the main repository of patient-related information within the hospital infrastructure;
- ICU Local Area Network, which enables real-time communication, synchronisation, and data transfer among medical devices, monitoring systems, and information servers within the ICU environment;
- Manual input provided by healthcare professionals, encompassing observations, clinical notes, and procedural records that complement automated data streams and capture contextual information not available through electronic systems.

The information generated in the ICU can be also broadly categorised as either quantitative or qualitative data [12]. Quantitative data include numerically measurable parameters, such as physiological signals, laboratory test results, mechanical ventilator settings, medication dosages, and severity of illness scores (for example, Acute Physiology and Chronic Health Evaluation (APACHE) and Sequential Organ Failure Assessment (SOFA) [22], [23]) [12], [13], [24]. In contrast, qualitative data include visual or graphical information, such as ventilator waveforms or radiographic images such as X-rays, and Computed Tomography (CT) scans [12]. These data types require pre-processing and feature extraction prior to computational analysis [25].

Among these sources, bedside medical devices, including monitors, ventilators, and infusion pumps, generate real-time waveforms and numerical measurements that are electronically presented on relational or interval scales [21]. Continuous monitoring facilitates the acquisition of high-resolution data, with sampling frequencies that can reach several hundred hertz, commonly around 400 Hz, for signals such as ECGs, and supports high-granularity datasets that capture minute-by-minute variations in physiological parameters, like heart rate, blood pressure, oxygen saturation, and intracranial pressure [21], [24], [26]. Consequently, even without waveform storage, a single patient may produce 0.5–4 Megabyte (MB) of data per day, increasing to approximately 100 MB when multiple high-frequency channels are recorded [21], [27]. This data volume

and temporal precision support advanced analytics and real-time predictive modelling, but also place significant demands on data storage, processing, and systems synchronisation.

To ensure that these heterogeneous datasets can be effectively utilised, data standardisation and interoperability are essential. This is achieved through established clinical terminologies and protocols [24]. For instance, diagnostic and procedural information commonly adheres to International Classification of Diseases (ICD) codes, while laboratory results are frequently mapped to Logical Observation Identifiers Names and Codes (LOINC), and other clinical terms to the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [13], [24], [26]. High time-resolution data are often stored in portable formats such as Hierarchical Data Format, Version 5 (HDF5) [26], [28]. Moreover, contemporary ICU databases increasingly adopt the Observational Medical Outcomes Partnership (OMOP) CDM, which promotes harmonisation and enhances data sharing [19], [29].

2.1.2 Limitations and Challenges in ICU Data

Despite the richness and diversity of ICU data, its effective utilisation for research and analytical purposes is hindered by several intrinsic limitations that must be carefully addressed during model development [19]. One of the primary challenges is data heterogeneity, defined as the variability in data structure, format, and content across different sources and institutions [30]. This variability arises from substantial differences in patient populations, including comorbidities, age, and disease severity, as well as from variations in clinical practice and information systems across institutions [19], [24]. Consequently, inconsistent data collection patterns emerge. For example, differences in patient length of stay and the frequency of physiological measurements complicate the development of generalisable models suitable for diverse patient cohorts [31].

Closely related to heterogeneity is the issue of data quality, which constitutes another significant limitation for AI applications in this area. The performance and reliability of predictive models are inherently dependent on the quality of their underlying data [24]. Real-world ICU datasets are particularly susceptible to human error, notably in manually input records, which may contain implausible values for parameters such as height or weight [32]. Furthermore, incompleteness and missing values are pervasive challenges [32]. Stringent de-identification procedures mandated by regulations such as the European Union's General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) frequently result in the suppression or generalisation of key variables. For instance, age may be grouped for patients over 89 years, and detailed medical histories may be omitted [19], [32]. Such constraints impede researchers' ability to contextualise patient outcomes and trajectories [19]. Additionally, a considerable portion of clinically relevant information, often conveyed in unstructured formats such as free-text notes or verbal communications, is typically excluded from structured datasets, thereby exacerbating data incompleteness [24].

Beyond heterogeneity and quality issues, the integration of multimodal data, that is, the combination of diverse data types, introduces further technical and methodological challenges [33]. Multimodal data are typically acquired using heterogeneous protocols that highlight different

anatomical, physiological, and biological aspects of the same individual [34]. Since each modality captures distinct physical and biological phenomena, the resulting datasets are inherently non-commensurable, exhibiting heterogeneous units, dimensionalities, and modality-specific artefacts derived from disparate acquisition and processing routines [34], [35]. This intrinsic variability in acquisition methods hinders data standardisation and interoperability across studies, demanding extensive preprocessing and the design of tailored model architectures to enable effective data integration [35]. Additionally, although high-resolution physiological data are collected continuously, only sparse, discrete summaries are typically incorporated into the EHRs, and these are often validated by ICU staff on an hourly basis [13], [26]. This asynchrony impedes precise temporal alignment with contextual data, such as medication administration times, which may be inaccurate by several minutes or even hours [26]. The absence of standardised acquisition methods and comprehensive metadata, including information such as sensor placement or concurrent clinical interventions, further limits the interpretability and collaborative use of multimodal ICU datasets [26].

These limitations constrain real-time data analysis, complicate cross-centre validation, and hinder the broader utilisation of high-resolution ICU datasets in advancing patient-centred, data-driven care like clinical DSSs [19], [24].

2.2 Clinical Data Representation and Interoperability

As discussed in Section 2.1, ICU data are heterogeneous, longitudinal, and generated across multiple clinical and technical systems. Their secondary use therefore depends not only on data availability, but also on consistent representation and interoperability, particularly when EHR and ICU data from different institutions or databases are analysed together [36], [37]. Achieving this requires standardised protocols, terminologies, and CDMs, since equivalent clinical concepts may otherwise be recorded using different structures, vocabularies, measurement units, and local conventions [36], [38], [39].

This section introduces the concepts of structured, unstructured, and multimodal clinical data, and presents the healthcare terminologies, exchange standards, and CDMs most relevant to the work developed in this dissertation.

2.2.1 Structured, Unstructured, and Multimodal Clinical Data

Structured clinical data refers to patient information stored in an organised and computer-readable format, usually according to a predefined schema [40], [41]. Common examples include coded diagnoses, surgical procedures, laboratory test results, medication administration records, physiological vital signs, flowsheets, and other tabular or time-stamped EHR variables, often stored in relational databases or structured tables that make them easier to query and analyse computationally [40], [42]. This structure makes such data particularly suitable for statistical analysis and ML, but it may not fully capture clinical context, uncertainty, causal reasoning, or the nuanced interpretation documented by healthcare professionals [40], [42], [43].

Unstructured clinical data complements structured records by capturing information in free-text narratives and reports that do not follow a fixed data model [40], [41]. Examples include clinical notes, nursing narratives, discharge summaries, radiology reports, and pathology reports [40], [42]. Because these documents often contain detailed descriptions of patient care, symptoms, temporal reasoning, causal relationships, clinical impressions, and diagnostic or therapeutic decisions, they may provide information that is absent or only partially represented in structured EHR fields [40], [43], [44]. Nevertheless, unstructured data also introduces specific challenges, particularly because its information is less directly computable and often requires Natural Language Processing (NLP) or other representation learning methods before it can be integrated into analytical workflows.

For this reason, structured and unstructured clinical data can be viewed as complementary representations of the patient state. This complementarity has motivated increasing interest in multimodal clinical representations, in which different types of patient information are jointly modelled. In this context, multimodal clinical data may include structured time-series variables, clinical text, medical images, physiological waveforms, multi-omics data, and external knowledge sources [38], [45]. Such approaches are increasingly relevant in clinical AI and EHR modelling because each modality can contribute a distinct perspective on the patient trajectory. When appropriately integrated, these complementary data streams may reduce the ambiguity of individual sources and support more robust clinical prediction or decision-support systems [38], [42], [45].

2.2.2 Healthcare Terminologies, Exchange Standards, and Common Data Models

Healthcare interoperability relies on several types of standards that work together. Clinical terminologies and classifications offer standard vocabularies for describing clinical concepts. Exchange standards explain how data moves between systems. CDMs set shared structures and meanings so different datasets can be compared more easily [36], [39], [46]. The following concepts are particularly relevant in this dissertation because they appear throughout the reviewed literature and underpin several approaches to clinical data harmonisation and cross-database research.

These mechanisms operate at distinct but complementary levels of healthcare interoperability. Clinical terminologies and classifications standardise the meaning assigned to diagnoses, laboratory observations, and other clinical concepts [1]. Exchange standards define how clinical and administrative information is structured and transferred between heterogeneous information systems [46]. CDMs address a different requirement by transforming source data into a shared schema and vocabulary suitable for consistent analysis across databases [36], [47]. Their use therefore depends on the interoperability objective: terminologies support semantic coding, exchange standards support communication between systems, and CDMs support harmonised storage and analysis. These mechanisms may coexist within the same data workflow rather than serving as direct alternatives.

1. **International Classification of Diseases**

   The ICD is an international classification system maintained by the World Health Organization (WHO) and used to represent diseases, causes of death, symptoms, and related

health conditions in a standardised way [48]. It is widely used for morbidity and mortality statistics, epidemiological surveillance, administrative reporting, reimbursement, and cohort definition in clinical research. Although ICD codes are frequently reused in clinical research, their structure was primarily developed for archiving and administrative purposes, including billing. Consequently, they may not represent nuanced clinical information such as severity, temporality, laterality, or detailed clinical reasoning [1], [44], [49], [50].

2. **Systematized Nomenclature of Medicine Clinical Terms**

   SNOMED CT is a comprehensive, multilingual clinical terminology designed to facilitate semantic interoperability within EHRs. It provides broad and scientifically validated coverage of clinical concepts, including diagnoses, symptoms, findings, procedures, and other healthcare concepts, and is mapped to several international standards to support semantic interoperability. Its main limitations relate to implementation complexity, the need for careful mapping to local systems or other coding standards, and possible gaps when highly specific administrative or locally defined concepts must be represented [47], [51], [52].

3. **Logical Observation Identifiers Names and Codes**

   LOINC provides a standard terminology for identifying laboratory tests, clinical measurements, documents, and other observations. It is widely used to harmonise laboratory results, vital signs, and clinical assessment instruments, enabling data pooling and comparison across institutions and research datasets. Nevertheless, mapping local observations to LOINC can be challenging when tests, panels, units, question wording, or temporal reference periods differ between institutions [39], [53], [54].

4. **Fast Healthcare Interoperability Resources**

   Fast Healthcare Interoperability Resources (FHIR) is a healthcare data exchange standard created by Health Level Seven (HL7) International and based on modular resources and modern web technologies, including REpresentational State Transfer (REST)ful Application Programming Interfaces (APIs) [55]. It is used to support the exchange of clinical and administrative data across healthcare systems, research applications, mobile applications, wearable devices, and digital health infrastructures through a resource-based representation of healthcare concepts. Its limitations include implementation complexity, variation across local profiles and versions, and the technical requirements associated with deploying and maintaining interoperable FHIR-based services [46].

5. **Observational Medical Outcomes Partnership Common Data Model**

   The OMOP CDM is a standardised CDM for observational health data, designed to transform heterogeneous clinical and administrative datasets into a shared structure. It supports

large-scale observational research, federated analytics, cohort definition, and reproducible cross-database studies, and is closely associated with the Observational Health Data Sciences and Informatics (OHDSI) community and its analytical tooling. However, implementing the OMOP CDM requires substantial Extract, Transform, and Load (ETL) work, careful vocabulary mapping, and, in some specialised domains, additional extensions to represent clinically relevant information not fully covered by the standard model [36], [47], [56].

2.3 Artificial Intelligence and Clinical Predictive Modelling

This section introduces the computational concepts that support the predictive modelling approaches explored in this dissertation. It first defines the relationship between AI, ML, and DL, including the role of Neural Networks (NNs) and representation learning in modern clinical modelling. It then situates Generative AI and LLMs within this broader hierarchy, before outlining the main learning paradigms and the general workflow through which clinical data are transformed, modelled, optimised, and evaluated.

2.3.1 Artificial Intelligence, Machine Learning, and Deep Learning

According to McCarthy, one of the founders of the field, AI can be defined as:

> "the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable." [57], [58].

As a subset of AI, ML focuses specifically on algorithms that improve their performance through experience with data rather than through explicitly programmed rules [59]. In clinical applications, traditional ML methods, such as Logistic Regression (LR) and Random Forest (RF), often rely on structured feature engineering, while DL, a specialised ML approach, uses multi-layered NNs to automatically learn hierarchical data representations [57], [60].

NNs are computational models composed of interconnected units organised in layers, in which input data are transformed through learned weights, biases, and activation functions to produce an output. Through this layered structure, DL models can perform representation learning, deriving informative latent representations from raw or low-level data and reducing the dependence on manual feature engineering [61], [62]. This is particularly relevant in EHR analysis, where DL models have shown strong capability in processing unstructured data, including clinical notes and images, although their reliance on large datasets remains a challenge in healthcare settings where labelled data are often limited [57], [63].

The effectiveness of DL models in processing EHR data relies heavily on optimisation techniques such as backpropagation, a gradient calculation algorithm that enables NNs to learn by

iteratively adjusting weights and biases to minimise a loss function, a quantitative measure of prediction error [64]. During training, outputs are first computed through a feedforward pass and compared with target values; the resulting error is then propagated backward through the network's acyclic directed graph, where the chain rule is used to compute gradients for each parameter with respect to the cost function [65]. These gradients indicate how weights and biases should be updated through gradient descent, and this process is repeated until the model progressively converges towards improved predictions [66].

2.3.2 Generative AI and Large Language Models

Building upon the foundations of AI, ML, and DL, Generative AI refers to a family of models designed not only to recognise patterns or assign labels, but also to produce new data artefacts based on patterns learned from training data. These models are commonly based on deep NNs that learn the statistical structure of large datasets and use this learned representation to generate synthetic outputs with characteristics similar to the original data. As a result, Generative AI can be applied across multiple modalities, including text, images, audio, and molecular structures [67], [68].

Within this broader class of models, LLMs represent one of the most prominent applications of Generative AI to language data [67], [69]. Based on Transformer architectures, LLMs are trained on large collections of linguistic tokens and learn probabilistic relationships between words and contexts, enabling natural language understanding, text generation, zero-shot generalisation, and reasoning-like behaviour [70]–[72]. In healthcare, models such as Generative Pre-trained Transformer (GPT)-4 and MedLM have demonstrated strong performance in processing unstructured clinical data and supporting tasks such as medical documentation, information extraction from EHRs, and diagnostic assistance, although hallucinations and integration barriers remain important limitations [71], [72]. This remains a rapidly evolving field, with new general-purpose and clinically specialised models being continually proposed. Consequently, the examples discussed here represent the state of the field during the period covered by this dissertation rather than an exhaustive or definitive set of models [67], [69].

Taken together, these concepts can be understood as a nested hierarchy: AI is the broadest field, ML comprises data-driven methods within AI, and DL represents a subset of ML based on multi-layered NNs [60], [63]. Generative AI can then be situated within this landscape as a family of mostly deep-learning-based models focused on generating new content, while LLMs correspond to a prominent class of generative models specialised in language and commonly built on Transformer architectures [68]–[70]. This hierarchy is schematically represented in Figure 2.1.

2.3.3 Learning Paradigms

ML approaches can be grouped according to the type of supervision available during training. This distinction is particularly relevant in healthcare, where labelled data may be limited, expensive to obtain, or dependent on expert annotation.

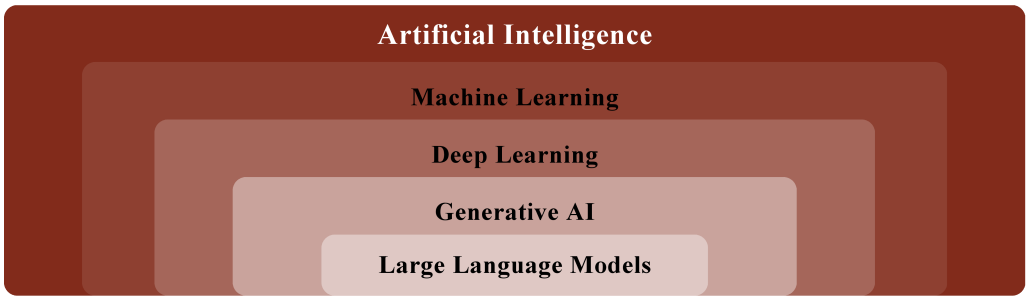


Figure 2.1: Schematic representation of the hierarchical relationship between AI, ML, DL, Generative AI, and LLMs. The figure was created by the author to summarise the conceptual organisation adopted in this dissertation.

- **Supervised learning** uses labelled datasets in which each input sample is associated with a target outcome, such as a diagnosis, prognosis, or clinical event. This is the most common setting for clinical prediction tasks, but it depends on the availability and quality of labelled data, which can be challenging in EHR-based research [60], [73].
- **Unsupervised learning** aims to identify latent structure in unlabelled data, for example through clustering, dimensionality reduction, or pattern discovery. In clinical contexts, these methods may support the identification of patient subgroups, disease phenotypes, or hidden associations without requiring predefined outcome labels [60], [73].
- **Semi-supervised learning** combines a limited amount of labelled data with a larger set of unlabelled samples. This approach is particularly relevant in healthcare, where expert annotation is costly and time-consuming, but large volumes of unlabelled clinical data may be available [60], [73].
- **Self-supervised learning** generates supervisory signals directly from unlabelled data through pretext tasks, such as predicting masked words in text or reconstructing missing parts of an input. By learning reusable representations from large unlabelled datasets, self-supervised methods can support downstream tasks such as classification, prediction, or retrieval, particularly in data-rich domains such as NLP, computer vision, and multimodal clinical modelling [74].

2.3.4 Clinical Predictive Modelling Workflow

After introducing the main families of learning systems, this subsection describes the practical workflow through which clinical predictive models are developed. In EHR-based research, this process begins with the transformation of heterogeneous clinical records into model-ready representations, continues with model training and hyperparameter selection, and concludes with performance evaluation on unseen data. The following sections therefore summarise the main concepts required to understand the experimental pipeline used in this dissertation, from data preprocessing and temporal representation to model training, optimisation, and evaluation metrics.

2.3.4.1 Data Pre-processing and Representation

Clinical and EHR data usually require pre-processing before predictive modelling, since they are often uncurated, high-dimensional, sparse, heterogeneous, and affected by quality issues such as random errors, systematic biases, noise, missing values, duplicate records, and implausible measurements [1], [75]. Data cleaning therefore represents an essential step in clinical modelling workflows, including procedures such as outlier detection, correction or removal of implausible values, duplicate handling, and assessment of clinical plausibility [37], [76]. Missing values may be handled through imputation strategies, sometimes combined with binary indicators that preserve information about whether a value was originally observed or imputed [37], [77]. In addition, harmonisation and scaling procedures are often required when variables originate from different systems, units, or numerical ranges [37], [78].

After cleaning, clinical variables must be transformed into numerical representations that can be processed by predictive models [79]. Categorical variables may require numerical encoding, while discrete medical codes and clinical concepts can also be represented through dense embeddings that project high-dimensional symbolic information into continuous vector spaces [1], [79]. Continuous variables, such as vital signs and laboratory measurements, are commonly normalised or standardised so that features measured on different scales contribute more comparably during model training [75], [78]. One common approach is z-score standardisation, which transforms variables to have approximately zero mean and unit variance, reducing the dominance of features with larger numerical ranges [75], [78].

Temporal representation is particularly important in ICU and longitudinal EHR data, where measurements are time-stamped, irregularly sampled, and collected at different frequencies [62], [75]. Clinical observations may therefore be aligned within predefined observation windows and segmented into meaningful intervals, such as hourly time steps, depending on the task and modelling objective [37], [62], [75]. Irregularly sampled variables can then be summarised within these windows using aggregation strategies, such as minimum, maximum, first, last, or other descriptive statistics, allowing models to capture both absolute values and temporal changes [37].

The result of these steps is a model-ready representation of the patient record. For classical ML models, this representation is often a tabular feature matrix in which each patient or episode is described by a fixed set of variables [76]. For longitudinal and DL approaches, the same clinical information may instead be organised as temporal matrices or tensors, preserving dimensions such as patients, time steps, and clinical variables or events [62], [78]. More advanced models can further learn latent representations from these processed inputs, compressing sparse or high-dimensional patient histories into dense representations for downstream prediction or analysis [1], [62], [78].

2.3.4.2 Model Training

After pre-processing, model training is the stage in which predictive algorithms learn relationships between patient-level representations and clinical outcomes. In supervised clinical prediction, this

process involves fitting model parameters using labelled examples while controlling the model's ability to generalise to previously unseen patients [60].

The available data are commonly divided into training, validation, and test subsets. The training set is used to fit model parameters, and any pre-processing operations that learn from the data, such as imputation or feature scaling, should also be estimated exclusively from this subset to prevent information leakage [37], [78]. The validation set supports model selection and hyperparameter tuning, whereas the test set remains unseen until the final performance assessment [60], [80].

The validation process also helps identify overfitting, which occurs when a model learns patterns or noise specific to the training data and consequently performs well during training but generalises poorly to unseen instances. Cross-validation extends this assessment by repeatedly training and validating the model across different data partitions, providing information about performance stability. In imbalanced clinical prediction tasks, stratified cross-validation is particularly useful because it preserves the outcome distribution across folds [37], [78].

Model training is also influenced by hyperparameters, which are configuration values defined before or during training and affect model behaviour, learning efficiency, and generalisation to unseen data. These settings may include regularisation strength, the number of decision trees in an ensemble, kernel parameters, learning rate, hidden-layer size, number of layers, dropout probability, or batch size [78], [80].

Manual tuning can be inefficient and time-consuming, particularly when several hyperparameters interact within a high-dimensional search space. Systematic strategies such as Bayesian optimisation can instead use information from previous trials to guide the search towards promising regions of that space [78], [81].

Automated optimisation frameworks support this process by managing the definition, execution, and pruning of hyperparameter trials. Optuna is an open-source framework that supports dynamically constructed search spaces through a define-by-run API and includes mechanisms for efficient trial pruning. Its default sampler is based on the Tree-structured Parzen Estimator, a sequential model-based optimisation method that samples promising configurations more efficiently than unguided manual search [81].

2.3.4.3 Classical Machine Learning Models

Classical ML models generally operate on fixed-length feature representations and provide interpretable and computationally efficient baselines for clinical prediction. The models most relevant to this dissertation are introduced below.

1. **Logistic Regression**

   LR is a linear probabilistic classifier that estimates the probability of a target outcome from a weighted combination of input features, commonly using a sigmoidal transformation for binary classification. Owing to its simplicity and interpretability, it is frequently used as a baseline in clinical prediction tasks, although its capacity to model complex nonlinear

interactions is limited unless such relationships are explicitly represented in the input features [60], [78].

2. **Random Forest**

   RF is an ensemble learning method that combines multiple decision trees, typically trained on different samples and feature subsets, to improve predictive robustness and generalisation. This structure allows the model to capture nonlinear relationships and feature interactions in tabular EHR data, although temporal dependencies must usually be represented through aggregated or engineered features rather than modelled directly as sequences [1], [78].

3. **Support Vector Machine**

   Support Vector Machine (SVM) is a maximum-margin classifier that seeks a separating hyperplane between classes in a high-dimensional feature space. Through kernel functions, SVMs can model nonlinear decision boundaries, making them suitable for complex feature spaces, but their training can become computationally expensive when applied to large datasets [78].

4. **Naive Bayes**

   Naive Bayes (NB) is a probabilistic classifier based on Bayes' theorem and the assumption that features are conditionally independent given the class label. This assumption simplifies learning and makes the method computationally efficient, supporting its use as a simple diagnostic or predictive baseline, although the independence assumption is often unrealistic in clinical data [30], [75].

2.3.4.4 Deep Learning Architectures for Longitudinal Clinical Data

Unlike classical models, longitudinal DL architectures can process temporally ordered patient representations and learn dependencies across successive clinical observations. The recurrent and convolutional architectures most relevant to this dissertation are described below.

1. **Recurrent Neural Networks**

   RNNs are neural architectures designed for sequential data, processing observations step by step while maintaining a hidden state that summarises information from previous time points. Gated variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), introduce mechanisms that regulate how information is retained, updated, or discarded, improving the ability to model longer temporal dependencies in patient trajectories [1], [62].

2. **Temporal Convolutional Networks**

TCNs are sequence-modelling architectures based on one-dimensional temporal Convolutional Neural Networks (CNNs). Causal convolutions prevent each output from using future observations, while dilated convolutions expand the receptive field across the sequence without requiring a proportional increase in network depth. Residual connections further support the training of deeper architectures. Together, these mechanisms allow TCNs to capture short- and long-range temporal patterns while enabling greater parallelisation than recurrent models [16], [76], [78].

2.3.4.5 Transformer Encoders and Attention

Transformer encoders are attention-based architectures that transform input sequences into contextual representations for downstream tasks. Their central operation is self-attention, which allows each time step or token to attend to other elements in the same sequence and capture dependencies between distant observations without processing them strictly step by step [82], [83].

In the standard attention formulation, each input is projected into queries (Q), keys (K), and values (V). Queries represent the information being sought, keys represent the elements against which each query is compared, and values contain the information to be aggregated [84]. The 2.1 shows how the equation for scaled dot-product attention is:

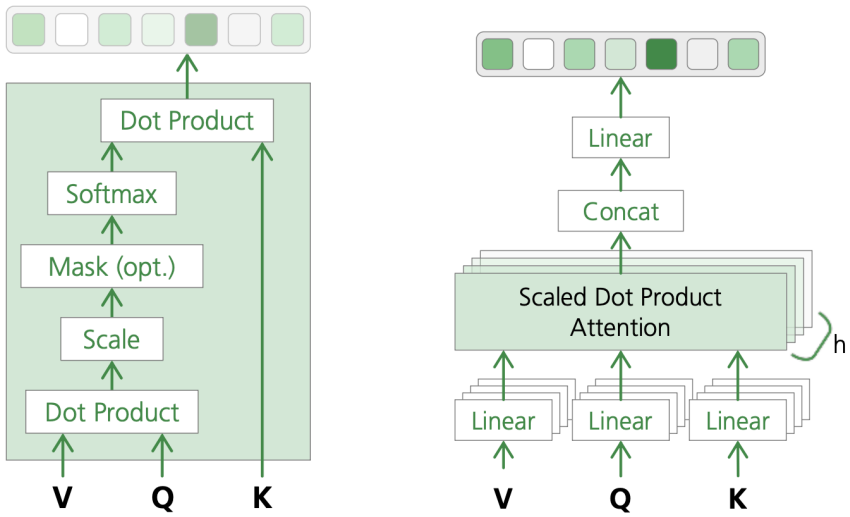
$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.1}$$

where d_k is the dimensionality of the key vectors. The scaling term prevents the dot products from becoming excessively large, while the softmax operation converts the similarity scores into attention weights that determine how strongly each value contributes to the output [82], [85]. When Q, K, and V are derived from the same sequence, the operation is termed self-attention. Multi-head attention applies several attention operations in parallel, allowing different heads to learn complementary relationships within the same input [82].

The Transformer architecture, introduced by Vaswani et al. in *Attention Is All You Need* [82], replaced recurrent sequence processing with stacked attention-based layers. Each encoder layer combines multi-head self-attention with a position-wise feed-forward network, residual connections, and layer normalisation to support stable representation learning. Because self-attention does not inherently encode sequence order, positional encodings are added so that the model can combine content-based relationships with information about the position of observations or tokens [82], [83]. This ability to model long-range dependencies has made Transformer-based architectures central to modern sequence modelling, including language models and clinical NLP applications [83].

2.3.4.6 Performance Evaluation

After model training, performance evaluation estimates how well a trained model generalises to data that were not used to fit its parameters [80]. This evaluation is commonly performed using an independent held-out test set or through k-fold cross-validation, ensuring that the reported



(a) Scaled dot product attention. (b) MHA.

Figure 2.2: Schematic representation of scaled dot-product attention and multi-head attention. Q, K, and V denote queries, keys, and values, respectively, and h denotes the number of attention heads. Retrieved from SANTOS [85], based on Vaswani, Shazeer, Parmar, *et al.* [82].

performance is not obtained from the same data used during training [78]. As discussed above, cross-validation is also useful for assessing model stability across different data partitions [78].

In addition to estimating average performance, it is often necessary to quantify the uncertainty associated with performance metrics. Bootstrapping is a statistical resampling technique in which repeated samples are drawn from the observed dataset with replacement, and the statistic of interest is recalculated for each resampled set. The resulting empirical distribution approximates the variability that would be expected across repeated samples from the underlying population, allowing estimates such as standard errors or confidence intervals to be obtained [86].

For binary classification, most metrics are derived from the confusion matrix, which compares predicted and true class labels [78], [87]. In this setting, true positives (TP) correspond to positive cases correctly identified by the model, and true negatives (TN) correspond to negative cases correctly identified [80]. False positives (FP) occur when negative cases are incorrectly classified as positive, whereas false negatives (FN) occur when positive cases are incorrectly classified as negative [80].

1. **Accuracy**

   Accuracy, defined in Equation 2.2, measures the proportion of correct predictions among all evaluated instances. Although intuitive, it can be misleading in imbalanced clinical datasets, since a model may obtain high accuracy by predominantly predicting the majority class while failing to detect clinically important rare events [80], [87].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.2}$$

2. **Precision**

   Precision, also known as positive predictive value and defined in Equation 2.3, measures the proportion of positive predictions that are truly positive. It is a critical metric in clinical scenarios where false positives are costly or disruptive, such as avoiding unnecessary clinical alarms or invasive follow-up procedures [80], [87].

$$Precision = \frac{TP}{TP+FP} \tag{2.3}$$

3. **Sensitivity**

   Sensitivity, also termed recall or true positive rate and defined in Equation 2.4, measures the proportion of actual positive cases correctly identified by the model. In intensive care prediction tasks, sensitivity is important when false negatives carry high clinical risk, such as failing to identify patient deterioration, mortality risk, or organ dysfunction [80], [87].

$$Sensitivity = Recall = \frac{TP}{TP+FN} \tag{2.4}$$

4. **Specificity**

   Specificity, defined in Equation 2.5, measures the proportion of actual negative cases correctly identified by the classifier. It reflects the model's ability to avoid false positives, which is important in diagnostic and early-warning settings where excessive false alarms may reduce clinical usefulness [87].

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

5. **F1-score**

   The F1-score, defined in Equation 2.6, is the harmonic mean of precision and sensitivity. It provides a single summary measure that balances false positives and false negatives, and is useful when both precision and sensitivity are important, particularly in imbalanced classification settings where accuracy alone may be insufficient [80], [87].

$$F1\text{-}score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity} \tag{2.6}$$

6. **min(+P, Se)**

   The $\min(+P, Se)$ metric corresponds to the minimum between precision, also denoted $+P$, and sensitivity (Se). It provides a conservative summary of their balance because its value

is determined by the weaker of the two quantities. Consequently, a high value can only be achieved when both precision and sensitivity are simultaneously high. Like the F1-score, it depends on the selected classification threshold [88].

$$\min(+P, Se) \quad = \quad \min(Precision, Sensitivity) \tag{2.7}$$

7. **AUROC**

   The Area under the Receiver Operating Characteristic Curve (AUROC) provides a threshold-independent measure of discriminative performance by summarising the receiver operating characteristic curve, which plots the true positive rate against the false positive rate across classification thresholds. An AUROC of 0.5 indicates discrimination no better than random chance, whereas a value of 1.0 indicates perfect discrimination [87].

8. **AUPRC**

   The Area under the Precision–Recall Curve (AUPRC) summarises the relationship between precision and sensitivity across different decision thresholds. It is particularly informative in imbalanced clinical datasets because it focuses on performance for the positive class, which is often the clinically relevant minority class [87].

Chapter 3

Literature Review

This chapter presents a comprehensive review of the state of the art relevant to the main research directions addressed in this work. It begins with the Search Strategy section, which describes the systematic methodology adopted for literature identification, screening, and selection.

Two main thematic reviews are subsequently developed. The Medical ICU Structuring Pipelines section examines existing approaches for data structuring, preprocessing, and harmonisation of medical data generated in ICU environments. The Generative AI Models Applied to Medical Data section reviews recent advances in DL and generative AI, with particular emphasis on the adaptation and empirical evaluation of Transformer-based architectures and LLMs for heterogeneous medical data, including EHRs.

The chapter concludes with a Critical Review of the Literature, which synthesises the principal findings from both state of the art analyses, critically examines their limitations, and identifies methodological gaps and open challenges in ICU clinical DSSs. These insights are used to motivate and contextualise the research directions pursued in the present work, highlighting areas where meaningful contributions may be achieved.

3.1 Search Strategy

A search strategy was designed to ensure a comprehensive and reproducible exploration of the literature supporting both state-of-the-art analyses developed in this dissertation. The methodology was informed by the principles of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [89], with the aim of promoting transparency and rigour in the identification, selection, and inclusion of relevant studies. While a full systematic review protocol was not formally adopted, key elements of structured and reproducible evidence synthesis were incorporated. For each of the research directions, a specific research question was formulated, and key concepts were identified to compose the corresponding search strings. The search strings were developed through an iterative process, combining keywords and synonyms using the Boolean operators “AND” and “OR”, to ensure a broad yet precise scope. Searches were conducted across multiple scientific databases, PubMed, ScienceDirect, Scopus, and Web of Science,

selected for their broad coverage of biomedical, clinical, and computational research. All retrieved articles were imported into the screening software Rayyan [90], which supported the selection of relevant studies through the application of predefined inclusion and exclusion criteria, based on a careful analysis of titles and abstracts.

For each state-of-the-art, the applied search strings, the number of retrieved, excluded, and included studies, as well as the results of the selected papers, are presented in the corresponding sections below. This search approach ensures a more robust overview of the literature considered and establishes a solid baseline for the analyses and discussions developed in the subsequent sections.

3.2 Medical ICU Structuring Pipelines

In this section, the literature review will focus on data structuring pipelines for medical ICU settings. It will describe the current approaches to processing, harmonising, and standardising ICU medical data, highlighting methods for integrating heterogeneous datasets, including EHRs and multimodal medical data. Key findings from existing pipelines will be summarised, and some limitations or gaps in current methodologies will be identified.

3.2.1 Search Strategy Results

The literature search for this state of the art was performed covering the last ten years, with the search conducted up to October 10, 2025. The following search string was used: ("Intensive Care Unit") AND ("Electronic Health Record" OR "medical data" OR "database") AND ("pipeline" OR "preprocessing" OR "harmonization" OR "standardization" OR "Benchmark").

A total of 748 articles were retrieved, of which 182 duplicates were removed. The remaining articles were screened. The studies considered for inclusion were required to be written in English and to focus on clinical DSSs for medical professionals, addressing practical methodologies for processing broad medical data, including multimodal sources such as EHRs and other heterogeneous clinical data. Studies were excluded if they were non-article formats, targeted exclusively at non-medical professionals, focused solely on a single data modality without integration with EHRs, or lacked a clearly described and evaluated data processing pipeline. Articles presenting ML pipelines were included only when data processing constituted a substantive methodological contribution rather than a peripheral preprocessing step. Literature reviews were considered solely if they reported original implementations or empirical analyses. This strategy prioritised studies offering methodological depth and evidence-based insights into clinical data pipelines over purely theoretical or conceptual contributions.

Following title and abstract screening in Rayyan [90], 42 articles were included, and 524 articles were excluded. Subsequent full-text assessment led to the additional exclusion of 10 studies that did not meet the predefined selection criteria, resulting in a final set of 32 articles, which constitute the literature review presented in this section.

3.2.2 Considerations for Structuring the Review

The reviewed literature demonstrates considerable diversity in data types, research objectives, and processing strategies, which permits several alternative ways of structuring the review, each with distinct advantages and limitations. Therefore, multiple organisational strategies were considered to determine the most appropriate framework for presenting the findings from the 32 included studies. These included organising studies by the temporal distribution of publications to illustrate the evolution of methodological approaches over the past decade, structuring them according to pipeline stages, from data acquisition and harmonisation to modelling and validation, grouping studies by levels of interoperability, such as institutional or multicentre implementations and standards-based pipelines, and arranging them by clinical or computational objectives to reflect increasing methodological complexity.

However, to ensure a clear and comparable structure, since the number of studies found was so considerable, a consistent pattern was identified regarding the types of data processed by the pipelines. Accordingly, the studies in this section will be presented in three categories: pipelines exclusively based on structured data, pipelines exclusively based on unstructured data, and pipelines based on both structured and unstructured data.

Within each category, the first works discussed will be those that do not propose a fully defined data harmonisation or preprocessing pipeline, but rather rely on partial, auxiliary, or parallel preprocessing strategies. These will then be followed by studies that explicitly present data harmonisation or preprocessing pipelines as a defined and central component of their methodology.

3.2.3 Pipelines Based on Structured Data

As defined in Section 2.2.1, structured data refers to routinely collected clinical variables represented in standardised, tabular formats. In the reviewed ICU literature, this typically includes demographics, vital signs, laboratory measurements, medication records, and coded diagnoses. Several contributions to structured data harmonisation focus not on fully end-to-end pipelines, but rather on isolated methodological components. In this subsection, the methodologies and selected results of 25 such studies are presented.

3.2.3.1 Preliminary and Partial Approaches

The harmonisation of structured EHR data rarely emerges as a single, fully integrated pipeline. Instead, it is typically realised through the progressive formalisation of isolated preprocessing layers that address specific challenges such as temporal alignment, semantic consistency, or feature construction. This subsection discusses eight studies that exemplify this layered evolution, ranging from partial methodologies addressing specific problems to more formalised and reusable harmonisation frameworks.

Concerns related to temporal dataset shift are explicitly addressed by Guo et al. [91], who outline a feature extraction procedure from Medical Information Mart for Intensive Care (MIMIC)-IV [17] for benchmarking domain generalisation algorithms. In this study, raw clinical measurements,

which are inherently irregular, sparse, and affected by missing values, are mapped to patient summary statistic levels computed over predefined temporal windows, including an initial 0–4 h window after ICU admission and broader retrospective intervals, thereby enforcing a consistent representation across training and evaluation periods. The resulting representations were evaluated for in-hospital mortality, long Length of Stay (LOS), sepsis, and invasive ventilation prediction. Although domain generalisation and unsupervised domain adaptation are applied downstream, the central harmonisation mechanism lies in the temporal summarisation of heterogeneous time series into fixed, structured vectors. Even though limited in granularity, this approach highlights the role of preprocessing choices in mitigating temporal drift.

A related approach to temporal abstraction is proposed by Fu et al. [92], who rely exclusively on event timestamps from longitudinal EHR data to classify clinical deterioration. Their methodology assumes that the sequence and density of recorded events reflect clinical workflow and healthcare processes, rather than solely patient pathophysiology. Timestamped events, including vital sign measurements, flowsheet entries, and order records, are segmented into regular time bins, within which event counts are computed per feature, yielding structured representations based purely on temporal occurrence. While this approach demonstrates that temporal abstraction can act as a harmonisation strategy, it discards information regarding the spacing between measurements and, where applicable, quantitative values such as medication dosages, potentially limiting physiological fidelity.

Beyond temporal considerations, several studies illustrate harmonisation through the structured definition of clinical concepts and outcomes. Thomas et al. present a computable phenotype for hospital-acquired venous thromboembolism, constructed to enable consistent reuse across large clinical databases [93]. Their methodology integrates ICD-9/10 diagnosis codes with Present on Admission indicators and Current Procedural Terminology (CPT) codes related to imaging studies, derived from EHR data at the University of Vermont Medical Center. Although narrow in scope, this work exemplifies harmonisation via the integration of established coding systems, enabling consistent outcome definitions across institutional contexts.

Semantic harmonisation also constitutes a further foundational layer for structured data reuse. The work presented by Kim et al. focuses on mapping nursing documentation from the Korea Medical Information Mart for Intensive Care (K-MIMIC) dataset to the SNOMED CT, proposing a preprocessing workflow centred on terminology alignment rather than automation [94]. Their approach involves the expert selection of top-level hierarchies and semantic tags, cross-validation among annotators, and an external expert review. While no scalable ETL pipeline is introduced, this mapping exercise underscores that semantic standardisation constitutes a foundational harmonisation layer, without which downstream AI based DSS development is not feasible.

A complementary perspective on semantic harmonisation is presented by Püttmann et al. [95], who assess the Findability, Accessibility, Interoperability, and Reusability (FAIR)ness of two Dutch ICU research databases previously transformed into the OMOP CDM. Rather than introducing a new transformation pipeline, the study evaluates the outputs of the existing ETL processes, including their data structures, vocabulary mappings, metadata, and publication through

the European Health Data and Evidence Network (EHDEN) portal. Within this framework, clinical data elements are represented using standard terminologies such as SNOMED CT and LOINC, while OHDSI tools support structural validation and data quality assessment. The study therefore demonstrates how CDM and standard vocabularies can facilitate semantic consistency and data reuse, while identifying remaining limitations related to globally unique identifiers and metadata standardisation.

Moving from semantic alignment to clinically informed feature construction, Stein et al. propose a reusable preprocessing workflow for extracting high frequency telemetry data alongside laboratory and demographic variables from paediatric ICU EHRs [96]. Unlike earlier works, this study explicitly documents the preprocessing steps as a standardised sequence, demonstrating how structured harmonisation can be directly integrated into physiological considerations in a population where normal ranges vary substantially with age.

A complementary perspective on preprocessing is provided by Jagesar et al. [97], who compare manually curated versus automatically extracted EHR data for ICU mortality prediction using the Amsterdam University Medical Center (UMC) [98]. Rather than proposing a reusable harmonisation framework, the authors employ an automated SQL-based extraction pipeline and deliberately exclude variables requiring manual curation, such as detailed diagnostic labels, a choice that improves scalability but removes clinically nuanced information and may limit model expressiveness and transferability. Missing physiological and laboratory values are imputed as normal, and feature selection is driven by domain knowledge. Although limited in generalisability, this study demonstrates that automated preprocessing pipelines, combined with clinical knowledge imputation, can reproduce the performance of manually curated reference models, highlighting manual curation itself as an implicit harmonisation strategy.

The final study addresses harmonisation challenges at the level of specific variables or recording practices. Wang et al. develop a standardisation approach for pulse oximetry and supplemental oxygen data within the Veterans Health Administration (VHA) system [99]. Their preprocessing pipeline resolves heterogeneous charting conventions by converting diverse representations into a unified litres-per-minute scale, followed by validation of the derived variables through face, concurrent, and predictive validity analyses. The resulting variables were also examined in relation to in-hospital and one-year post-discharge mortality. Although local in scope, this work illustrates how variable standardisation constitutes a necessary preprocessing layer within broader structured data pipelines.

3.2.3.2 End-to-End Pipelines

In this subsection, 17 works are examined. These end-to-end pipelines differ substantially in both intent and design. To avoid conflating fundamentally distinct methodological contributions, this subsection adopts a taxonomy grounded not in abstract technical characteristics, but in practical reusability and methodological scope. The reviewed works are organised according to the role that harmonisation plays within the overall pipeline and the extent to which the resulting artefacts

can be reused across datasets, tasks, and modelling approaches. Four general categories are distinguished: i) Data harmonisation as an independent and reusable component; ii) Data harmonisation for standardised clinical labels and phenotypes; iii) Data harmonisation within benchmarking and evaluation pipelines; iv) Data harmonisation for specific task and model architectures.

Data harmonisation as an independent and reusable component

The works in this category treat data harmonisation as a standalone methodological contribution, explicitly separating the harmonisation process from any single predictive task or modelling choice. Their primary objective is to reduce structural and semantic heterogeneity across ICU datasets in a way that enables reuse, cross-dataset analysis, and external validation.

A paradigmatic example is provided by Bennett et al. through `ricu` [100], an open-source R package offering a unified interface to multiple large-scale ICU databases, including MIMIC-III [101], MIMIC-IV [17], electronic Intensive Care Unit (eICU) [13], Amsterdam UMC [98], and High time-Resolution Intensive Care Database (HiRID) [102]. The motivation underlying `ricu` is the observation that ICU data is highly heterogeneous and the absence of computational infrastructure for handling multiple datasets simultaneously. Rather than enforcing strict compliance with a single CDM, `ricu` adopts a concept-centred harmonisation strategy, supporting 119 clinical concepts grounded primarily in the OMOP Vocabulary. Harmonisation is achieved through a layered loading process that resolves discrepancies in units, data types, and temporal references, producing aligned longitudinal representations. Importantly, `ricu` is explicitly designed for reuse and extensibility, allowing incremental integration of new datasets and concepts.

A closely related, though narrower, approach is proposed by Liao and Voldman through the METRE [103]. METRE focuses on consistent extraction across MIMIC-IV [17] and eICU [13], transforming raw EHR data into structured, time-aligned data frames suitable for ML. Within a single pipeline, METRE formalises cohort definition, variable extraction across static, vital, and intervention data, unit harmonisation, outlier handling, missingness thresholding, and semantic aggregation. The pipeline was evaluated on hospital mortality, Acute Respiratory Failure (ARF), and shock prediction. Although narrower in scope than `ricu`, METRE exemplifies an end-to-end approach in which harmonisation decisions are made explicit and consistently applied across datasets.

Figure 3.1 presents a schematic overview of the METRE extraction pipeline, illustrating how heterogeneous EHR data are processed into structured, time-aligned tables. The pipeline separates static and time-dependent data streams, applies temporal aggregation and preprocessing, and produces three harmonised outputs: a Static table, demographics and comorbidities, a Vital table, vital signs and blood gas measurements, and an Intervention table, procedures and administered therapies.

While `ricu` and METRE focus primarily on harmonised extraction, other works extend these principles towards the construction of harmonised datasets rather than solely extraction pipelines. Oliver et al. introduced BlendedICU [104], the first international ICU dataset synthesised from Amsterdam UMC [98], eICU [13], HiRID [102], and MIMIC-IV [17]. The authors address the

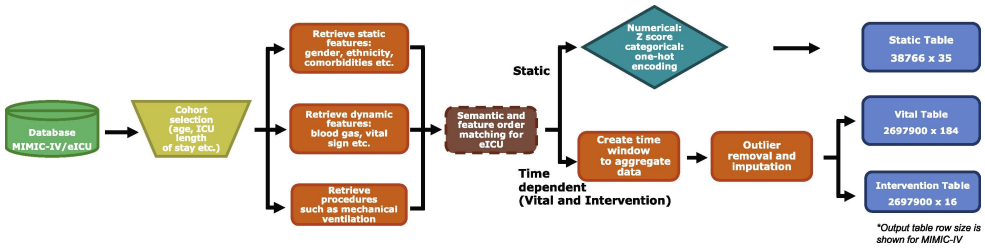


Figure 3.1: METRE schematic retrieved from the original article [103]. (Dashed box is specific for eICU. Output dataframe shape is from MIMIC-IV.)

lack of a harmonised processing pipeline spanning these widely used databases, each characterised by distinct languages, terminologies, units, and clinical practices. Their pipeline extracts variables from each source and maps them to the OMOP CDM, applying temporal resampling, value clipping and normalisation, drug mapping, missing data imputation, and provides both raw and openly available preprocessed versions to the community.

Figure 3.2 presents a schematic overview of the BlendedICU construction pipeline. The figure illustrates the progressive transition from heterogeneous source databases to a harmonised dataset through explicit filtering, variable selection, and clinically validated preprocessing steps, including temporal resampling, normalisation, and imputation, culminating in both raw and preprocessed OMOP-compatible outputs suitable for downstream analysis.

A complementary perspective on large-scale harmonisation is offered by Rojas et al. through the Common Longitudinal ICU Format (CLIF) [105]. CLIF is an open-source data model and ETL pipeline designed to support multicentre ICU research across 39 hospitals and 9 health systems in the United States. It defines a minimum set of Common ICU Data Elements (mCIDE) organised into more than 20 longitudinally structured, clinically meaningful tables, mapping source-specific variables to categorised and standardised representations. CLIF is further implemented as a federated analysis framework, enabling reproducible research across heterogeneous health systems without centralising patient data. The framework was demonstrated through external validation of an in-hospital mortality prediction model and an assessment of 72-hour temperature trajectories associated with mechanical ventilation and in-hospital mortality. Unlike OMOP-based transformations, CLIF prioritises clinically interpretable longitudinal structure over exhaustive vocabulary coverage.

Formal transformation to a CDM is further exemplified by Paris et al. [106], who documented the complete conversion and evaluation of MIMIC-III [101] into the OMOP CDM . Their openly documented SQL-based ETL pipeline involved semantic and structural mapping of source tables, standardising data from 26 MIMIC tables into 19 OMOP tables, and was supported by systematic unit testing and data quality assessment using established OHDSI tooling. Validation was carried out through a large-scale datathon, during which hundreds of participants conducted clinical research projects using the MIMIC-OMOP database, demonstrating the model's interpretability and practical usability. This effort ultimately resulted in a freely accessible MIMIC-OMOP dataset for reproducible intensive care research.

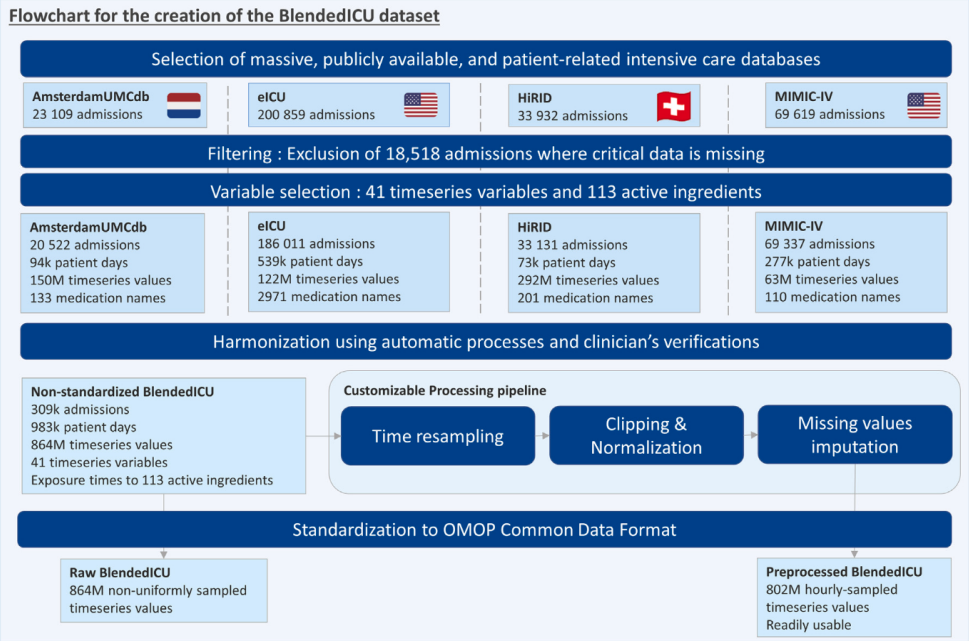


Figure 3.2: Flowchart illustrating the end-to-end pipeline for the creation of the BlendedICU dataset, from multi-database selection and filtering to harmonisation and standardisation into the OMOP CDM. Adapted from Oliver, Allyn, Carencotte, *et al.* [104].

While previous efforts focused on large-scale or multi-database standardisation, Bode et al. propose a data integration pipeline aimed at establishing an interoperable routine for paediatric intensive care medicine at Hannover Medical School [107]. The authors describe a three-level integration framework in which more than 3,000 data items from heterogeneous clinical systems, including the HIS and Patient Data Management System (PDMS), are converted into an openEHR-based representation. The pipeline combines systematic data item identification, an intermediate SQL-based structure, and 31 interoperable ETL processes supported by 15 openEHR templates, resulting in anonymised and standardised longitudinal data for approximately 4,200 paediatric patients. The workflow relies exclusively on open-source software and openly available data models.

Finally, adopting a similarly institutional approach, Fleuren et al. describe the Dutch Data Warehouse (DDW), a multicentre EHR database comprising full-admission data from critically ill COrona VIrus Disease (COVID)-19 patients across the Netherlands [108]. Data from participating centres are processed through an ETL pipeline, manually mapped to a common concept vocabulary defined by the authors, and enriched with derived clinical scores such as SOFA and APACHE II. Although the pipeline draws on OMOP and openEHR-related principles, access to the resulting dataset is governed by data-sharing agreements and ethical constraints, limiting reuse despite the availability of the underlying repository code. Nevertheless, this work illustrates the organisational and procedural challenges inherent to harmonising ICU data at a national scale.

From a comparative perspective, these approaches differ less in their fundamental objectives than in the level at which harmonisation is enforced. `ricu` and METRE prioritise reusable extraction logic across heterogeneous sources, differing mainly in scope and vocabulary coverage

[100], [103]. BlendedICU, CLIF, and OMOP-based transformations shift harmonisation towards dataset-level standardisation, trading flexibility for interoperability and consistency [104]–[106]. Institution-specific pipelines such as those proposed by Bode et al. and Fleuren et al. further embed harmonisation within organisational and governance structures [107], [108]. Across all cases, harmonisation relies on similar mechanisms, concept mapping, unit normalisation, temporal alignment, and missing data handling, with distinctions arising primarily from design choices regarding reuse, scale, and intended scope rather than from fundamentally different methodological paradigms.

Data harmonisation for standardised clinical labels and phenotypes

In contrast to general approaches, the pipelines grouped here employ data harmonisation as enables the consistent derivation of clinically meaningful labels or phenotypes. In these works, the scope of reuse is intentionally limited to a specific outcome or clinical task, but the preprocessing and harmonisation steps are nevertheless fully specified and reproducible.

Hofford et al. introduced OpenSep [109], an open-source pipeline for the automated computation of the SOFA score and Sepsis-3 classification, validated on MIMIC-IV [17]. Addressing the limited transferability and dataset-specific nature of existing Sepsis-3 implementations, OpenSep formalises cohort selection, variable extraction, score computation, and validation within a unified pipeline. Central to this approach is the Minimal Sepsis Data Model (MSDM), which structures essential inputs such as laboratory measurements, vital signs, vasopressor use, and antibiotic administration. Although restricted to sepsis-related tasks and not intended as a general-purpose harmonisation framework, the pipeline prioritises generalisability by requiring substantially less data restructuring than complex data model–based approaches. The public availability of the source code further positions OpenSep as a focused example of transparent, task-driven data harmonisation.

A closely related contribution is provided by Porschen et al. through pyAKI [110], an open-source pipeline for automated Acute Kidney Injury (AKI) classification according to Kidney Disease Improving Global Outcomes (KDIGO) criteria using MIMIC-IV [17] time-series data. The pipeline standardises input formats, baseline serum creatinine definitions, imputation strategies, and temporal logic for AKI staging, and demonstrates through validation against clinician-derived gold standards that preprocessing choices substantially affect outcome labelling. pyAKI further incorporates interpolation to upsample sparse measurements to hourly resolution and derives final AKI stages from both creatinine and urine output. As with OpenSep, its primary contribution lies in the formalisation and standardisation of clinically defined labels, reinforced by a minimal, standardised AKI data model designed to ensure reproducibility across studies.

Data harmonisation within benchmarking and evaluation pipelines

The pipelines discussed in this section used data harmonisation within standardised experimental workflows designed to support systematic evaluation and comparison of AI models. In this con-

text, harmonisation functions as a prerequisite for controlling methodological variability rather than as an end in itself.

Tang et al. [111] developed Flexible Data-Driven Pipeline (FIDDLE), an open source, preprocessing framework evaluated on MIMIC-III [101] and eICU [13]. FIDDLE aims to reduce reliance on manual, specific task feature engineering by systematically standardising the transformation of raw EHR data into representations ready to be applied to models. The pipeline operates in three stages, pre-filter, transform, and post-filter, to address common challenges such as heterogeneous data types, irregular sampling, and high dimensionality. Through largely data-driven strategies, including carry forward imputation and discretisation based on quintile, FIDDLE converts input data into a matrix of invariant time features and a tensor of dependent time features. Although evaluated on in-hospital mortality, ARF, and shock prediction, its primary objective is to generate reusable feature representations that facilitate consistent and reproducible EHR analysis across tasks and datasets. While semantic harmonisation across institutions is not explicitly addressed, FIDDLE provides a general preprocessing pipeline that standardises key design choices and is openly available for reuse.

Building upon `ricu` for initial data harmonisation and concept definition, van de Water et al. [112] introduced Yet Another ICU Benchmark (YAIB), an open-source, modular framework designed to support five major open-access ICU datasets: MIMIC-III [101], MIMIC-IV [17], eICU [13], HiRID [102], and Amsterdam UMC [98]. Unlike `ricu`, which primarily provides the harmonisation and concept extraction layer, YAIB extends this foundation into a benchmarking framework by adding five task definitions, ICU mortality, AKI, sepsis, kidney function, and remaining LOS, together with exclusion criteria, preprocessing configurations, model training, and standardised evaluation. Users can configure imputation, missingness indicators, and multiple AI models, including LR, gradient boosting, GRU, LSTM, TCN, and Transformer-based neural networks. The framework is publicly available and emphasises transparent, reproducible experiment definitions for method development and cross-dataset comparison.

The COVID-19 benchmark by Gao et al. shares a similar objective, providing open-source preprocessing pipelines for cleaning, filtering, missing-value imputation, and cohort construction using the Tongji Hospital (TJH) and COVID Data for Shared Learning (CDSL) COVID-19 EHR datasets [113]. MIMIC-III [101] and MIMIC-IV [17] were additionally considered for external comparison. The complete framework includes standardised processing and the evaluation of 18 predictive models for length-of-stay and early-mortality prediction, enabling controlled and reproducible model comparison.

Data harmonisation for specific task and model architectures

This final group includes pipelines in which preprocessing is designed around a specific clinical task or modelling approach. These workflows provide complete solutions for their intended applications, but their task-specific processing and representations may limit reuse across other tasks, datasets, or models.

Escudero-Arnanz et al. propose the eXplainable Spatio-Temporal Graph Convolutional Neural Network (XST-GCNN) architecture, explicitly designed for processing heterogeneous and irregular multivariate TS data from ICU patient EHRs [114]. The pipeline's core is tightly linked to the prediction model, capturing both temporal and feature dependencies through a unified spatio-temporal graph representation. Using ICU data from a Spanish hospital, heterogeneous clinical variables are embedded into a unified feature space, and pairwise relationships are quantified using a Gower-based distance, a similarity measure suitable for mixed numerical and categorical variables, to accommodate mixed data types. These graphs are then processed within a graph convolutional framework, in which graph construction and temporal modelling are intrinsically linked to the prediction task. As a result, data preprocessing, representation learning, and inference form a model-specific pipeline rather than a reusable harmonisation framework.

Adopting a complementary but less explicitly architectural approach, Xu et al. [115] presented a semi-supervised pipeline for extracting clinical states associated with pneumonia in two cohorts, the Successful Clinical Response In Pneumonia Therapy (SCRIPT) study and MIMIC-IV [17]. The pipeline optimises feature selection, normalisation, temporal aggregation to day resolution, optimisation of preprocessing choices guided by a "common-sense" ground truth, dimensionality reduction, and ensemble clustering. The pipeline process ends by learning a low dimensional embedding space via Principal Component Analysis (PCA) and employing a robust ensemble clustering method to identify and characterise five clinical states.

Rather than being driven primarily by data representation or architecture, Wang et al. [116] presented a physiology phenotyping pipeline motivated by the limitation of based conventional EHR analyses in capturing latent physiological states. Their approach integrates EHR extraction with mechanistic modelling of the glucose–insulin system, using data assimilation and Markov Chain Monte Carlo (MCMC) to estimate unobserved specific patient parameters, followed by unsupervised learning to derive continuous physiological phenotypes that are subsequently validated against ICD-9/10 codes and clinical documentation. In this setting, harmonisation is achieved implicitly through the mechanistic model, which regularises sparse and irregular EHR observations via continuous simulation.

A similarly model dependent interpretation is adopted by Jajcay et al. [117], who propose a pipeline for cardiogenic shock prediction in MIMIC-III [101] centred on systematic management of missing data. After cohort selection based on ICD-9 codes, the pipeline performs detailed data inspection, outlier correction, variable grouping, and multivariate imputation using multiple imputation by chained equations, generating several completed datasets from which a final model cohort is derived and evaluated using a gradient-boosted predictor.

Likewise, Liu et al. [118] address a different clinical objective by developing an unsupervised pipeline for stratifying patients with vital organ failure, in which diagnosis histories encoded as ICD codes are transformed into structured representations through aware ontology embeddings, autoencoder representation learning, and joint optimisation with a clustering objective.

Although these last three pipelines differ substantially in clinical focus and modelling strategy, they share a common characteristic, data harmonisation is inseparable from the modelling

framework itself, with representation construction, inference, and learning jointly defined, yielding end-to-end workflows that are computationally complete but inherently limited in reuse beyond their original tasks and datasets.

3.2.4 Pipelines Based on Unstructured Data

As discussed in Section 2.2.1, unstructured data lacks a fixed tabular schema and includes clinical notes, narrative reports, and textual or semi-structured EHR metadata, such as parameter names and flowsheet labels. This subsection discusses four studies. The first two are presented as preliminary and partial approaches because text processing supports downstream modelling or label generation. The remaining two define end-to-end workflows for concept mapping and harmonisation, although they still require expert validation and are not necessarily fully automated or publicly reproducible.

3.2.4.1 Preliminary and Partial Approaches

Gupta et al. [119] introduced the Hierarchical Attention-based Integrated Learning (HAIL) framework, an attention-based architecture designed to process unstructured clinical notes for predictive tasks, including in-hospital mortality and LOS prediction, using MIMIC-III [101], MIMIC-IV [17], and PUBHEALTH [120]. The framework uses a pre-trained language model, such as Bidirectional Encoder Representations from Transformers (BERT) [121], to generate text embeddings. Hierarchical attention is then applied at the sentence level to aggregate information across text chunks and at the feature level to refine missing value embeddings. The resulting patient embedding is used for prediction through a Graph NN. Missing textual information is handled internally through attention-based representation refinement, rather than through explicit preprocessing or harmonisation before modelling. Therefore, the transformation of unstructured data in HAIL remains dependent on the proposed architecture, and no standalone reusable harmonisation pipeline is defined.

Similarly, Nowak et al. [122] present a Transformer-based approach for structuring radiological free-text reports to create labels for corresponding image data, supporting the development of image-based DSSs. Although the resulting labels are used with radiographic images, the processing workflow itself operates on the associated free-text reports, which justifies its classification within this subsection. The preprocessing task consists of automatically annotating reports to identify five typical radiological findings using text-based Transformer models. The methodology involved training BERT models [121], bidirectional Transformer encoders pre-trained for language representation, on human annotations, referred to as "gold labels", after which the trained models generated automated "silver labels" for the remaining reports. Although this process structures free text, it is designed to support image model training rather than harmonise textual data across systems or studies. The code associated with the report annotation pipeline is referenced through prior work, but is not explicitly released as a standalone reusable preprocessing framework

in this study. Consequently, its utility is primarily defined by its interaction with DSS model training, where performance improvements were observed when using Transformer-generated "silver labels" compared with "gold labels" alone.

3.2.4.2 End-to-End Pipelines

In contrast to the previous approaches, the following studies propose more complete pipelines for harmonising unstructured or semi-structured data elements across heterogeneous EHR systems. Dam et al. [123] proposed an augmented intelligence pipeline to facilitate concept mapping of unstructured parameter names extracted from EHR data originating from multiple Dutch hospital systems. The approach maps parameter names to a locally defined reference set of 1,679 clinically relevant concepts, originally curated through extensive manual effort. The pipeline follows three stages: linguistic preprocessing, including tokenisation, lemmatisation, and synonym handling; encoding through term frequency-inverse document frequency; and multi-class concept prediction using a linear classifier trained via stochastic gradient descent. Predicted labels associated with numeric parameters are then refined using aggregated data distribution statistics. Importantly, the harmonisation process is independent of downstream predictive models and is designed to support ontology alignment across heterogeneous EHR systems. Although manual validation remains necessary, the public pipeline substantially reduces the workload associated with manual concept mapping, initially estimated at 2,000 hours.

Fan et al. [124] present a semi-automated pipeline for aligning EHR data and mapping nursing documentation concepts across multiple healthcare sites. Flowsheet fields, including template and measure names, are reformatted into template-measure pairs, which serve as the units for harmonisation. The pipeline then applies exact matching, lexical matching using string similarity measures, and semantic alignment to the Clinical Care Classification (CCC) terminology. In the semantic stage, LLMs generate embeddings for both template-measure pair names and CCC concepts, with cosine similarity used to rank candidate mappings. The workflow is intended to accelerate manual alignment and concept mapping rather than optimise downstream predictive performance. Although modular and potentially extensible to alternative terminologies, expert validation remains necessary and, due to contractual restrictions, neither the underlying data nor the full implementation code is publicly available. Accordingly, Fan et al. can be considered an end-to-end harmonisation workflow for nursing documentation alignment, but not a fully automated or fully reproducible pipeline, since expert validation remains necessary and the complete implementation is not publicly available.

3.2.5 Pipelines Based on Structured and Unstructured Data

Approaches that integrate structured and unstructured data introduce additional complexity to data harmonisation pipelines, as they must reconcile heterogeneous modalities while preserving semantic and temporal consistency. This subsection discusses three studies combining structured

and unstructured clinical data. Rather than forming a clear progression from partial to end-to-end pipelines, these works illustrate different integration strategies: multimodal dataset linkage for benchmarking and fairness analysis, task-specific feature fusion for predictive modelling, and infrastructure-level integration of structured streams with clinical notes [125]–[127].

Yang et al. [125] introduce the publicly available MIMIC-IV-Social Determinants of Health (SDOH) dataset by linking structured and unstructured EHR data from MIMIC-IV [17] with external community-level SDOH databases, including County Health Rankings (CHR), the Social Vulnerability Index (SVI), and the SDOH itself. The EHR data include structured tabular features, such as demographics, risk scores, and laboratory measurements from the first 24 hours of ICU stay, as well as unstructured discharge notes. Structured numerical features undergo median imputation and standard scaling, while categorical variables are processed through constant imputation and one-hot encoding. Discharge notes are cleaned, tokenised, lemmatised, filtered through corpus-specific stop word removal, and represented using Term Frequency-Inverse Document Frequency (TF-IDF). Community-level SDOH features are integrated by mapping patient zip codes to county-level data using residential ratios. Although the study provides a well-documented multimodal linkage and enables downstream benchmarking across 15 feature sets, its preprocessing choices primarily support comparative modelling and fairness analysis rather than a reusable harmonisation pipeline.

A contrasting approach is presented by Mani et al. [126], who developed a preoperative multimodal architecture integrating structured EHR data, such as demographics, comorbidities, and clinical covariates, with unstructured free-text inputs from past medical and surgical history, medication lists, and problem lists. Structured data were processed through numerical encoding, categorical transformation, standardisation, and one-hot encoding. Free-text reports were stemmed, tokenised, and converted into numerical vectors using a bag-of-words approach to create a Document-Feature Matrix (DFM). The resulting tabular and text-derived representations were combined into a preoperative dataset used by an Extreme Gradient Boosted (XGBoost) model to predict perioperative safety indicators. Although the authors evaluate the full cohort and procedure-specific subcohorts, the pipeline remains specific to a single institutional context, with clinical data institutionally sourced and code available only upon request.

A different multimodal strategy is presented by Noteboom et al., who propose a pipeline for integrating real-time and retrospective ICU data within a cloud environment for structured storage and advanced clinical decision support [127]. The pipeline ingests and stores high-frequency structured physiological data, while incorporating unstructured EHR notes primarily to support cohort selection rather than feature extraction. Structured data streams are continuously transmitted to a cloud data warehouse, where they are stored as a raw full copy and pseudonymised by default. Clinical notes are ingested into the same environment, enabling unified querying across modalities. Integration between structured and unstructured data is achieved through targeted keyword searches in clinical notes to identify events such as cardiac arrest; once these events are located, corresponding high-frequency time-series data can be extracted for temporal analysis. Manual screening of selected keyword cohorts is used to validate the unstructured component. Although

the pipeline combines structured and unstructured data operationally, it does not define extensive text preprocessing or representation learning for unstructured data. It is therefore multimodal at the infrastructure level, but not a comprehensive harmonisation framework for unstructured clinical text. As an institutional implementation at Amsterdam UMC, relying on a cloud service provider and security mechanisms such as private networking and encryption, neither the source code nor the underlying data are publicly available.

3.2.6 Summary and Synthesis

Across the reviewed literature, a central conclusion is that data harmonisation in ICU settings is rarely realised as a single procedure, but rather as a spectrum of design choices that balance interoperability, clinical specificity, and computational feasibility. Reproducible pipelines are most clearly articulated in works that explicitly formalise extraction, transformation, and validation stages as reusable artefacts, often supported by open-source implementations or CDMs. Examples include harmonisation frameworks grounded in OMOP or openEHR transformations, as well as universal dataset extraction layers such as `ricu` and METRE, which prioritise transparent preprocessing logic over a specific task optimisation [100], [103], [106], [107]. These approaches consistently emphasise explicit documentation of assumptions, semantic mappings, and temporal alignment as requirements for cross-dataset reuse and external validation.

Beyond these independent and reusable frameworks, several pipelines demonstrate that the scope and design of harmonisation are closely tied to the intended analytical objective. Task focused pipelines for standardised clinical labels or phenotypes, as well as frameworks oriented for some benchmarking, tend to prioritise internal consistency and controlled preprocessing over broad semantic interoperability, yielding workflows that are reproducible but deliberately constrained in reuse [109]–[113]. Conversely, pipelines in which harmonisation is tightly integrated with specific modelling architectures or representation learning strategies achieve coherent end-to-end solutions at the cost of generalisability, while approaches to unstructured or multimodal data show greater transferability when preprocessing remains modular and explicitly separated from downstream models, even when expert validation is still required [114], [116], [118], [123], [124], [127].

Table 3.1 provides a structured summary of the 32 studies included in this part of the review. For each approach, it presents its methodological role, data modalities and datasets, harmonisation focus, clinical tasks, public code availability, primary contribution, and main limitation, enabling direct comparison across the reviewed pipelines. Public code is reported as available when the article provides an accessible implementation repository or project website, and as available upon request when access depends on contacting the authors.

Table 3.1: Structured summary of the 32 included ICU data structuring approaches.

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Bennett et al. (2023) [100] *ricu*	Reusable harmonisation framework	Structured data; MIMIC-III, MIMIC-IV, eICU, AmsterdamUMCdb, and HiRID	Semantic mapping, unit standardisation, and temporal alignment	Lactate–mortality and diabetes–insulin-use analyses	Yes	Provides a unified interface for extracting 119 clinical concepts across five heterogeneous ICU databases.	The concept dictionary is not comprehensive because concepts were added case by case.
Bode et al. (2023) [107]	Institutional integration pipeline	Structured data; Hannover Medical School HIS and PDMS	Semantic mapping, ETL, and longitudinal structuring	No clinical tasks studied	Yes	Converts heterogeneous paediatric ICU data into standardised openEHR representations through interoperable ETL processes.	Source-specific extraction remains resource-intensive and templates may require local adaptation.
Dam et al. (2023) [123] *AutoMap*	Concept-mapping pipeline	Unstructured EHR metadata; Dutch ICU Data Warehouse	Semantic mapping and text structuring	No clinical tasks studied	Yes	Uses NLP and ML to map heterogeneous EHR parameter names to a standardised concept set.	Mapping recall is imperfect and manual validation remains necessary.
Escudero-Arnanz et al. (2025) [114] *XST-GCNN*	Model-specific predictive pipeline	Structured data (time-series); Hospital of Fuenlabrada	Feature construction and temporal representation	Multidrug-resistance prediction	Yes	Models temporal and feature dependencies through a unified spatio-temporal graph representation.	Inference is approximately three times slower than lightweight baseline models.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Fan et al. (2025) [124]	Semi-automated concept-mapping pipeline	Semi-structured EHR metadata; Two institutional Epic sites	Semantic mapping and lexical alignment	No clinical tasks studied	No	Combines exact, lexical, and LLM-based semantic matching for multi-site flowsheet alignment.	Expert verification remains necessary because semantic mappings can be inconsistent.
Fleuren et al. (2021) [108] *DDW*	Multicentre data warehouse	Structured data; Dutch multicentre COVID-19 ICU data	ETL, semantic mapping, and feature derivation	No clinical tasks studied	Yes (upon request)	Establishes a national full-admission ICU data warehouse with harmonised concepts and derived clinical scores.	Data and repository access are governed by data-sharing agreements and ethical restrictions.
Fu et al. (2021) [92]	Task-specific temporal abstraction	Structured data (event timestamps); Institutional EHR data	Feature construction and temporal alignment	Composite clinical deterioration prediction	Yes	Uses only EHR entry timestamps to represent healthcare processes and predict deterioration.	Evaluated on a relatively small dataset from a single healthcare system.
Gao et al. (2024) [113]	Benchmarking framework	Structured data; TJH [113] and CDSL[128], with MIMIC-III and MIMIC-IV for external comparison	Missing-data handling, feature construction, and cohort standardisation	Outcome-specific LOS and early-mortality prediction	Yes	Provides standardised preprocessing and evaluation of 18 models across two COVID-19 ICU datasets.	Mortality and LOS models were constrained to identical architectures for comparison.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Guo et al. (2022) [91]	Temporal-shift evaluation framework	Structured data; MIMIC-IV	Temporal aggregation, feature construction, and missing-data handling	In-hospital mortality, long LOS, sepsis, and invasive ventilation prediction	Yes	Benchmarks domain generalisation and adaptation under temporal dataset shift.	Coarse temporal information prevented detailed estimation of the rate of performance deterioration.
Gupta et al. (2025) [119] *HAIL*	Model-specific text-processing framework	Unstructured data (clinical text); MIMIC-III, MIMIC-IV, and PUBHEALTH	Text representation and missing-data handling	In-hospital mortality, LOS, and medical-claim classification	No	Integrates hierarchical attention and graph neural networks to model long and incomplete clinical text.	Dense patient graphs create substantial computational demands.
Hofford et al. (2022) [109] *OpenSep*	Clinical phenotype pipeline	Structured data; MIMIC-IV	Label derivation, feature construction, and semantic mapping	Sepsis-3 classification and SOFA score calculation	Yes	Provides an open-source pipeline and minimal data model for reproducible Sepsis-3 phenotyping.	Execution speed was traded for code readability and ease of review.
Jagesar et al. (2024) [97]	Extraction comparison study	Structured data; AmsterdamUMCdb	Feature construction and label derivation	In-hospital mortality prediction	No	Compares automatically extracted EHR variables with manually curated mortality-model inputs.	Single-centre evaluation limits the generalisability of the findings.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Jajcay et al. (2023) [117]	Task-specific preprocessing pipeline	Structured data; MIMIC-III	Missing-data handling and feature construction	Cardiogenic shock prediction	Yes (upon request)	Defines a systematic processing and multivariate-imputation workflow for cardiogenic shock prediction.	The pipeline was not evaluated on EHR datasets other than MIMIC-III.
Kim et al. (2025) [94]	Semantic mapping study	Structured nursing data; K-MIMIC	Semantic mapping	No clinical tasks studied	No	Maps ICU nursing data to SNOMED CT to support semantic consistency and reuse.	The mapping process remains manual and requires improved efficiency for broader application.
Liao and Voldman (2023) [103] *METRE*	Multi-database extraction pipeline	Structured data; MIMIC-IV and eICU	Feature construction, missing-data handling, unit standardisation, and temporal alignment	Hospital mortality, ARF, and shock prediction	Yes	Enables harmonised extraction and cross-database model validation between MIMIC-IV and eICU.	Cross-database evaluation is restricted to variables shared by both databases.
Liu et al. (2023) [118]	Model-specific clustering pipeline	Structured data (clinical codes); MIMIC-III	Semantic representation and feature construction	Heart, respiratory, and kidney failure patient clustering	No	Combines ontology-grounded ICD embeddings, graph attention, and deep clustering.	Uses only ICD information without integrating demographics, procedures, or vital signs.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Mani et al. (2025) [126]	Task-specific multimodal pipeline	Structured and unstructured data; Institutional spine-surgery data	Multimodal integration and feature construction	Extended LOS, 90-day reoperation, and ICU admission prediction	Yes (upon request)	Integrates structured EHR variables and free text for perioperative safety prediction.	Some frequent text features lack clinical relevance and may contribute to overfitting.
Noteboom et al. (2025) [127]	Institutional multimodal infrastructure	Structured data (time-series) and clinical text; Amsterdam UMC	Multimodal integration and temporal alignment	Cardiac-arrest cohort identification	No	Integrates high-frequency monitor data and clinical notes within a cloud data environment.	Keyword-based event identification produces excessive false positives.
Nowak et al. (2024) [122]	Text-based label-generation workflow	Unstructured data (radiology reports); University Hospital Bonn	Text structuring and label derivation	Radiological finding label generation	No	Uses text-based Transformers to generate labels for chest radiographs from corresponding reports.	Labels inherit inaccuracies and omissions from the source radiology reports.
Oliver et al. (2023) [104] *BlendedICU*	Harmonised dataset pipeline	Structured data; AmsterdamUMCdb, eICU, HiRID, and MIMIC-IV	Semantic mapping, unit standardisation, and temporal alignment	ICU mortality and LOS comparison	Yes	Produces an international ICU dataset harmonised to the OMOP CDM.	Extraction is less exhaustive than pipelines specialised for a single database.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Paris et al. (2021) [106] *MIMIC-OMOP*	Common-data-model transformation	Structured data; MIMIC-III	Semantic mapping and ETL	No clinical tasks studied	Yes	Transforms MIMIC-III into the OMOP CDM through structural and semantic mapping.	Centralising measured facts in one table increases row volume and reduces query performance.
Porschen et al. (2025) [110] *pyAKI*	Clinical label pipeline	Structured data (time-series); MIMIC-IV demonstration cohort	Label derivation and temporal alignment	AKI classification	Yes	Implements KDIGO criteria in an open-source and standardised time-series pipeline.	Standardised procedures for testing AKI classification software remain unavailable.
Püttmann et al. (2023) [95]	FAIRness assessment workflow	Structured data; DDW	Semantic and structural assessment	No clinical tasks studied	No	Assesses two OMOP-converted ICU databases using standardised FAIRness criteria.	Evaluation was restricted to two overlapping Dutch ICU use cases.
Rojas et al. (2025) [105] *CLIF*	Federated harmonisation framework	Structured data; 39 hospitals across nine United States (US) health systems	Semantic mapping and longitudinal structuring	In-hospital mortality prediction and temperature-trajectory assessment	Yes	Defines an open-source longitudinal ICU format for federated multicentre research.	Federated analyses require labour-intensive manual execution and site-specific debugging.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Stein et al. (2025) [96] *PicEWS*	Feature-engineering pipeline	Structured data (time-series); Great Ormond Street Hospital	Feature construction, unit standardisation, and missing-data handling	Cardiovascular deterioration prediction	Yes	Consolidates high-frequency paediatric ICU data using age-normalised and variability-aware features.	Single-centre evaluation with a relatively small dataset requires external validation.
Tang et al. (2020) [111] *FIDDLE*	General preprocessing framework	Structured data; MIMIC-III and eICU	Feature construction, missing-data handling, and temporal alignment	In-hospital mortality, ARF, and shock prediction	Yes	Systematically transforms structured EHR data into reusable fixed and temporal feature representations.	Porting clinical features across institutions remains unresolved.
Thomas et al. (2020) [93]	Phenotype development workflow	Structured data (ICD and CPT codes); University of Vermont Medical Center	Label derivation and feature construction	Hospital-Acquired Venous Thromboembolism phenotype development	No	Develops and validates a computable phenotype for hospital-acquired venous thrombosis.	Not explicitly reported.
Wang et al. (2021) [99]	Variable-standardisation methodology	Structured data; VHA Corporate Data Warehouse	Unit standardisation and missing-data handling	In-hospital and post-discharge mortality risk analysis	Yes	Standardises pulse oximetry and supplemental oxygen data across more than 130 hospitals.	Developed for non-ICU hospitalisations and excludes advanced respiratory support recordings.

Continued on next page

Work	Approach Role	Data Modalities and Datasets	Harmonisation Focus	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Wang et al. (2023) [116]	Mechanistic phenotyping pipeline	Structured data; Columbia University and UCHealth	Feature construction and physiological representation	Glycaemic phenotyping	No	Uses data assimilation and mechanistic modelling to estimate latent physiological parameters.	Results may be sensitive to preliminary modelling and parameter-estimation choices.
Xu et al. (2024) [115]	Clinical-state extraction pipeline	Structured data (time-series); SCRIPT and MIMIC-IV	Feature construction, missing-data handling, and temporal aggregation	Pneumonia-associated clinical-state extraction	Yes	Uses semi-supervised preprocessing and ensemble clustering to identify five reproducible clinical states.	The MIMIC-IV cohort contains less severe-disease diversity than SCRIPT.
Yang et al. (2023) [125] *MIMIC-IV-SDOH*	Multimodal data-linkage workflow	Structured and unstructured data with external SDOH; MIMIC-IV, CHR, SVI, and SDOH	Multimodal integration and feature construction	In-hospital mortality, 30-day readmission, and one-year mortality prediction	No	Links MIMIC-IV with community-level SDOH data to evaluate prediction and fairness.	Individual-level, high-resolution SDOH data were unavailable.
van de Water et al. (2024) [112] *YAIB*	Multicentre benchmarking framework	Structured data; MIMIC-III, MIMIC-IV, eICU, HiRID, and AmsterdamUMCdb	Feature construction, missing-data handling, and temporal alignment	ICU mortality, AKI, sepsis, kidney function, and remaining LOS prediction	Yes	Provides reproducible cohort definition, preprocessing, training, and evaluation across five ICU databases.	Limited to ICU settings and features shared across the supported datasets.

3.3 Generative AI Models applied to Medical Data

This section presents a structured review of generative AI models applied to medical data. It examines the adaptation and empirical evaluation of architectures such as Transformers and LLMs across tasks involving EHRs, multimodal clinical data, synthetic data generation, representation learning, and clinical DSSs. Particular emphasis is placed on practical implementations and quantitative or qualitative performance assessments, to identify methodological limitations, open challenges, and opportunities that motivate and inform the present work.

3.3.1 Search Strategy Results

The literature search for this review was conducted covering the last five years, with inclusion completed up to November 10, 2025. The following search string was applied: ("Generative AI" OR "Large Language Model" OR "Transformer") AND ("Electronic Health Record" OR "medical data") AND ("clinical decision support" OR "representation learning" OR "Benchmark" OR "Predictive modeling").

A total of 540 articles were retrieved, of which 125 duplicates were removed. The remaining articles were screened. Studies were considered eligible if they were written in English and proposed, implemented, or adapted generative AI models, such as Transformer-based architectures or LLMs, applied to medical data, with a particular focus on EHRs or other structured and multimodal clinical datasets. Eligible studies were required to address tasks including, but not limited to, clinical decision support, predictive modelling, representation learning or synthetic data generation, and to report quantitative experimental evaluation. Studies were excluded if they were non-article formats, focused exclusively on medical imaging, biosignals, or NLP without integration with EHR data, lacked practical implementation or empirical assessment of the proposed generative approach.

Following title and abstract screening in Rayyan [90], 26 articles were included for full-text review, while 389 records were excluded. Subsequent full-text assessment resulted in the exclusion of three additional studies. Two articles were removed for failing to meet the predefined inclusion criteria upon detailed evaluation, and one study was excluded due to the absence of public availability and the inability to obtain full-text access within the required time-frame. Consequently, a final set of 23 articles was retained and forms the basis of the literature review presented in this section.

To facilitate a systematic and conceptually coherent analysis, the selected studies were subsequently organised into four thematic categories, reflecting both the underlying modelling paradigms and their clinical application scope: Transformer and LLM-based Representation Learning for Structured and Longitudinal EHR; Multimodal Representation Learning and Data Fusion in Heterogeneous EHR; LLM for Clinical DSS: From Predictive Models to Integrated Systems; and Generative Models for Synthetic EHR. This categorisation provides the organisational framework for the literature review presented in the following sections.

3.3.2 Transformer and LLM-based Representation Learning for Structured and Longitudinal EHR

This subsection reviews six works that examine how Transformer architectures and LLMs are used to learn representations from structured and longitudinal EHRs. The studies by Dwivedi et al. [129], Wang et al. [130], and Qiu et al. [131] focus primarily on representation learning itself. The remaining works by Wu et al. [132], Shoham et al. [133], and Kang et al. [134] are also included in this subsection because, despite being evaluated on downstream predictive tasks, their main contribution lies in how clinical records are encoded, aligned with language model representations, or enriched with external medical knowledge. Across all studies, the shared objective is to address the limitations of conventional sequence models when faced with the sparsity, heterogeneity, and hierarchical structure of real-world EHR data.

A core representational challenge is explicitly addressed by Dwivedi et al. [129], who show that standard subword tokenisation schemes, fragment the semantics of medical codes and fail to capture clinically meaningful co-occurrence patterns. Their UniStruct architecture introduces an adapted Byte Pair Encoding (BPE) strategy that merges frequently co-occurring medical codes into atomic tokens, thereby preserving relational structure within longitudinal patient histories. Methodologically, UniStruct employs a dual-modality Transformer, a BERT-based encoder [121] that processes unstructured clinical text from the current visit, and a separate Transformer that models structured history using a causal language modelling objective. According to the authors, incorporating structured patient history is the dominant factor driving performance gains, yielding up to a 23% recall improvement on internal hospital data and the public EHRSHOT dataset [135]. However, they note that structured and unstructured modalities remain encoded in separate representation spaces, limiting deeper cross-modal integration. Figure 3.3 provides an overview of the UniStruct architecture.

Where UniStruct focuses on representation quality at the token and sequence level, Wang et al. [130] address task adaptability under limited supervision. The authors proposed a task-independent pipeline designed to improve zero-shot and few-shot learning through soft prompt transfer wich they called Soft Prompt Transfer for Electronic Health Records (SptEHR). Their methodology follows three stages: self-supervised pre-training using masked code modelling on structured EHR codes, where diagnosis, procedure, and medication codes are treated as discrete tokens within a Transformer architecture, using the MIMIC-III dataset [101], supervised multi-task continual learning to acquire task-specific soft prompts, and, for unseen tasks, new prompts are constructed as linear combinations of existing ones based on task similarity. Empirically, SptEHR consistently outperforms conventional fine-tuning and prompt-tuning baselines on MIMIC-III, with particularly large gains in extreme few-shot settings. The study concluded that task scaling serves as an efficient alternative to model scaling, allowing models to rapidly generalise to unseen tasks without extensive computational overhead. Figure 3.4 provides an overview of the SptEHR architecture.

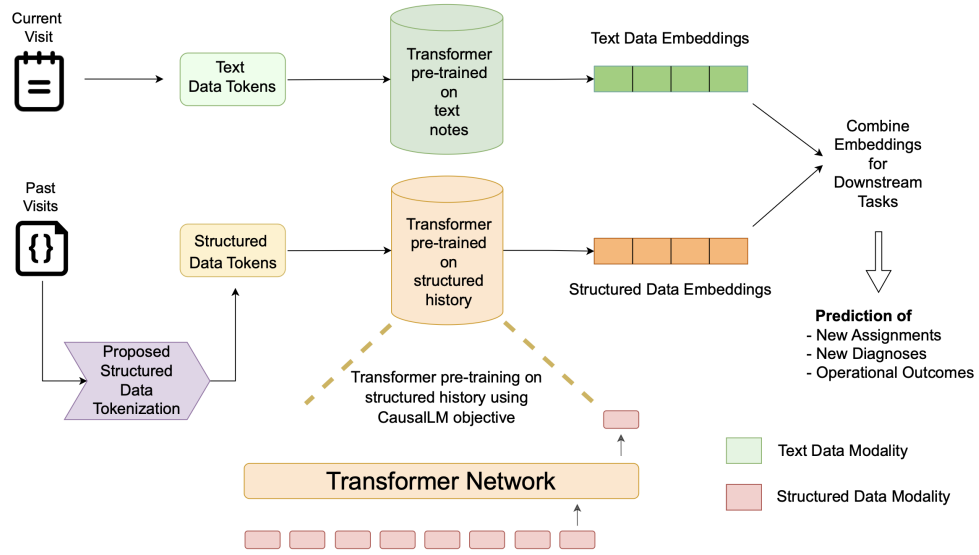


Figure 3.3: Overview of the UniStruct architecture, retrieved from the original article [129]. The figure illustrates the dual-modality Transformer design, where unstructured clinical text from the current visit and longitudinal structured patient history are encoded separately, with the latter modelled using a causal language modelling objective.

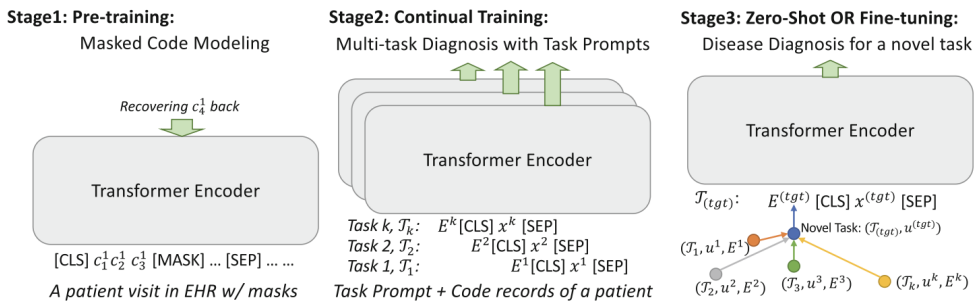


Figure 3.4: Overview of the SptEHR architecture, retrieved from the original article [130]. The figure illustrates the three-stage framework comprising self-supervised pre-training on structured EHR codes, supervised multi-task continual learning with task-specific soft prompts, and prompt transfer for zero-shot and few-shot adaptation to unseen clinical tasks.

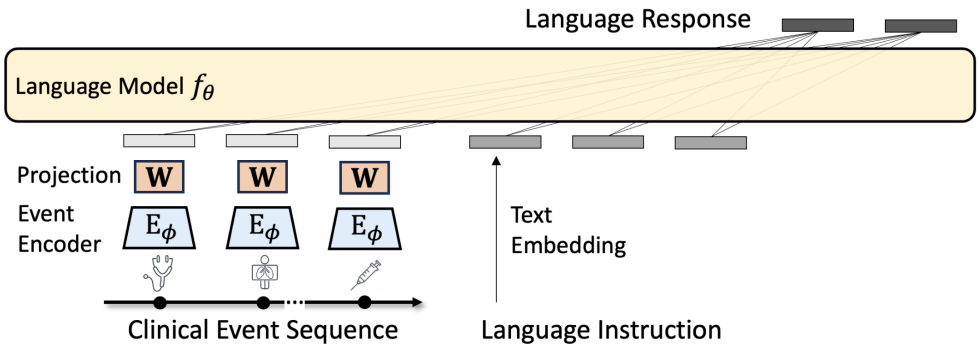


Figure 3.5: Overview of the Llemr model architecture, retrieved from the original article [132]. The figure illustrates how timestamped clinical event sequences are encoded and linearly projected into a LLM to enable instruction-based reasoning and downstream clinical prediction.

Motivated from physician burnout and the difficulty of navigating complex EHR user interfaces, a more radical departure from task-specific training is proposed by Wu et al. [132], who introduced the Llemr framework and the MIMIC-Instr dataset, derived from MIMIC-IV [17]. Operationally, patient records are represented as streams of timestamped clinical events derived from structured EHR data, which are first encoded using ClinicalBERT [136] and then mapped to a backbone LLM, Vicuna-7b-v1.5 [137], via linear projection. Although the original patient records are not free-text, the approach relies on an intermediate text-aligned representation to interface structured clinical events with aLLM. Through two-stage curriculum training focused on schema alignment and clinical reasoning, Llemr achieved an AUROC of 83.9% for mortality prediction, 80.9% for ICU diagnosis classification, and reaches approximately 73% of GPT-4 performance. Although the dataset may contain noise inherent to GPT-assisted generation, this approach effectively bridges the schema gap between structured EHR data and natural language achieving excellent results in answering diverse inquiries about a patient and performs on par with state-of-the-art baselines when further fine-tuned for clinical predictive tasks. Figure 3.5 provides an overview of the Llemr architecture, illustrating the direct alignment between longitudinal clinical event representations and instruction-following LLMs.

In contrast, Shoham et al. [133] investigate whether pre-trained LLMs can be directly repurposed for structured EHR prediction. The Clinical Prediction with Large Language Models (CPLLM) method presented involves the fine-tuning of pre-trained LLMs such as Llama2 [138] and BioMedLM [139] for clinical disease and readmission prediction. CPLLM treats structured EHR data as a chronological document where medical concepts are represented by textual descriptions, utilizing QLoRA [140] and PEFT [141] for efficient training. Experimentally, CPLLM achieves state-of-the-art performance on disease prediction and readmission tasks across MIMIC-IV [17] and eICU [13], including a AUROC of 76.41% for respiratory failure, without requiring extensive clinical pre-training, surpassing Med-BERT [142] and REverse Time AttentIoN (RETAIN) [143] in both AUROC and AUPRC. Although the underlying data originate from structured EHR records, CPLLM operates over an explicitly textualised representation to enable

the reuse of pre-trained LLMs. However, the authors highlight the higher hardware demands associated with large parameter counts.

Extending representation learning beyond supervised prediction, Qiu et al. [131] propose Variational deep survival clustering (VaDeSC)-EHR, a Transformer-based variational autoencoder for clustering longitudinal survival data extracted from EHRs. The architecture combines hierarchical ICD-10 information with patients' diagnosis sequences to learn compact representations suitable for survival-based clustering. Using UK Biobank data [144], VaDeSC-EHR outperformed baseline methods on synthetic and real-world benchmarks and identified four Crohn's disease subgroups with distinct diagnostic trajectories and risk profiles. This work demonstrates the potential of combining survival modelling and unsupervised clustering to identify clinically relevant patient subgroups.

Lastly, Kang et al. [134] present LLM-DG, a framework that enriches disease prediction through inter-patient and intra-patient modelling. Large language models generate semantic representations of diagnostic codes, patients, and discharge summaries, while patient clustering captures shared trajectories and a dynamic graph models disease evolution within individual patients. Experiments on MIMIC-III [101] and MIMIC-IV [17] demonstrated a 12.39% improvement in weighted F1-score for diagnosis prediction compared with state-of-the-art baselines. Despite the computational cost of feature generation, LLM-DG illustrates how external medical knowledge can enrich conventional code-based representations.

3.3.3 Multimodal Representation Learning and Data Fusion in Heterogeneous EHR

This subsection reviews five task-oriented works on multimodal representation learning and data fusion in heterogeneous EHRs. The studies by Engelke et al. [145], Hemamalini et al. [146], and Zhu et al. [88] are included because they explicitly integrate multiple clinical data modalities within unified predictive frameworks, albeit through different mechanisms, ranging from standards-based interoperability to LLMs-mediated semantic integration and knowledge grounding. The remaining works by Yang et al. [147] and Xu et al. [148] could also be viewed as representation or knowledge modelling approaches. However, they are grouped here because their primary contribution lies in fusing heterogeneous clinical information, whether multi-source or knowledge-derived, to support downstream clinical prediction.

Clinical data heterogeneity, reliance on manual feature engineering, and variation in FHIR implementations limit the scalability and generalisability of existing models. To address these challenges, Engelke et al. [145] developed FHIR-Former, an open-source framework combining FHIR with the encoder-based language model GeBERTa [149]. The framework processes structured and unstructured data using sliding and expanding windows and maps predictions back to FHIR risk assessment resources, supporting integration into clinical infrastructures. Evaluated on retrospective inpatient data from Essen University Hospital across four prediction tasks, it achieved 88.1% accuracy for in-hospital mortality without manual feature engineering. Its main contribution is

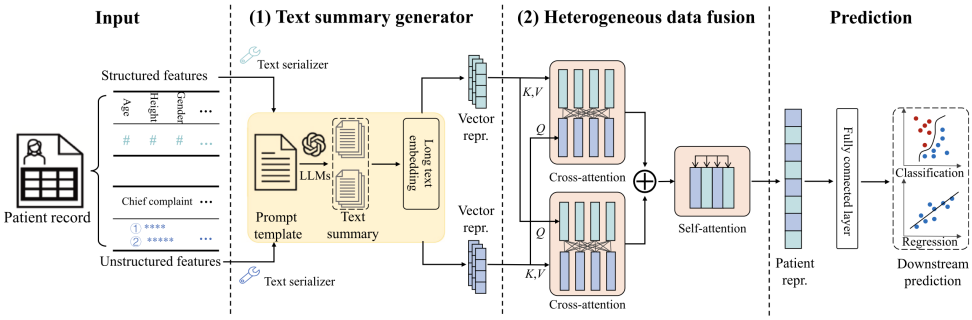


Figure 3.6: Overall network architecture of the TabPF framework, retrieved from the original article [147]. The figure illustrates the LLM-mediated transformation of heterogeneous EHR features into a unified textual modality, followed by attention-based fusion through symmetric cross-attention and self-attention mechanisms for downstream clinical prediction.

the demonstration that multimodal modelling can be integrated with established interoperability standards, rather than a substantial improvement in predictive performance. Nonetheless, they acknowledge limitations in interpretability and restricted generalisability beyond inpatient settings.

While FHIR-Former prioritises interoperability, subsequent works adopt a more model-centric perspective by using LLMs to unify heterogeneous representations. Yang et al. propose TabPF [147], a prompt-learning fusion framework for extracting patient representations from numerical, categorical, and textual EHR features. These features are converted into a unified textual modality through tailored prompts and ChatGLM-4-Flash [150], embedded using BGE M3-Embedding [151], and combined through cross-attention and self-attention. Evaluated on three prediction tasks using eICU [13] and the CECMed elderly cohort [152], TabPF outperformed traditional ML and Transformer-based tabular baselines, achieving an AUROC of 89.8% for mortality prediction. The authors also report that LLM-based summarisation improved regression performance by up to 42%, suggesting that semantic transformation and attention-based fusion can capture relevant patterns across heterogeneous clinical variables. Figure 3.6 provides a high-level overview of the TabPF architecture, highlighting how LLMs are employed as an intermediate semantic layer to enable attention-based fusion of heterogeneous tabular EHR representations.

A related knowledge-based approach is presented by Xu et al. in DearLLM [148], which incorporates LLM-derived feature relationships into personalised clinical prediction. Unlike TabPF, which unifies heterogeneous features through text, DearLLM uses the conditional perplexity of HuatuoGPT-II [153] to estimate how strongly clinical features are related within a patient's visit history. These relationships form a directed prior graph that is integrated with structured EHR representations through graph pooling. Experiments on MIMIC-III [101] and MIMIC-IV [17] showed improvements over state-of-the-art methods, particularly under data scarcity. Although not multimodal, DearLLM performs knowledge-guided feature-level fusion by integrating LLM-derived correlation priors with structured EHR representations. Xu et al. conclude that utilizing LLMs to deduce quantitative correlation priors reduces learning uncertainty and accurately reflects complex medical associations.

Complementing these frameworks, Hemamalini et al. [146] address multimodal representation learning under extreme label scarcity, proposing Multimodal Few-Shot Learning (MFSL), a framework for disease prediction from heterogeneous EHRs. The approach combines structured EHR variables transformed into natural language templates with raw clinical narratives, processed independently by a dual-encoder architecture followed by contrastive learning using InfoNCE loss to align representations in a shared latent space. Few-shot inference is conducted through in-context learning, whereby a pre-trained LLM is conditioned on a small set of labelled patient examples. Evaluated on MIMIC-IV [17], MFSL achieves an accuracy of 86.4% and an F1-score of 84%, significantly outperforming zero-shot, single-modal, and meta-learning baselines. While the results demonstrate that multimodal few-shot learning can remain effective in low-resource and rapidly evolving clinical settings, the dependence on templated text generation and external LLMs may pose challenges for reproducibility across institutions.

While MFSL focuses on aligning heterogeneous modalities through contrastive learning and in-context reasoning, Zhu et al. shift the focus towards grounding multimodal representations in external medical knowledge, introducing the EMERGE framework a Retrieval Augmented Generation (RAG) framework designed to incorporate external medical context into predictive modelling [88], [146]. Addressing the limitations of neural models trained from scratch, EMERGE extracts clinically relevant entities from time-series data and clinical notes using LLM prompting combined with statistical filtering, and aligns them with the PrimeKG knowledge graph [154] to generate task-relevant health status summaries via DeepSeek-V2 Chat [155]. These summaries are then fused with EHR embeddings through an adaptive cross-attention multimodal network. The overall EMERGE architecture is illustrated in Figure 3.7. EMERGE outperforms multimodal and knowledge-augmented baselines on MIMIC-III [101] and MIMIC-IV [17] for mortality and readmission prediction, achieving an AUROC of 89.5% on MIMIC-IV mortality. Ablation studies confirm the critical contribution of both the generated summaries and cross-attention fusion, with robustness maintained even under extreme data scarcity, such as 1% of training samples. The authors argue that grounding generation in curated medical knowledge graphs is shown to mitigate LLM hallucination, even at the cost of increased pipeline complexity and computational overhead.

3.3.4 LLM for Clinical DSS: From Predictive Models to Integrated Systems

This subsection reviews eight works examining the use of LLMs for clinical decision support, illustrating a progression from task-specific predictive augmentation towards systems integrated with clinical practice. Glicksberg et al. [156], Cohen et al. [157], and Li et al. [158] examine how LLMs can contextualise or structure evidence from existing models and clinical data sources, positioning them as decision-support layers rather than standalone predictive models. The remaining works [159]–[163] are grouped here because their main contribution lies in integrating LLM-generated outputs into clinical workflows, with emphasis on interpretability, clinician interaction, and decision support.

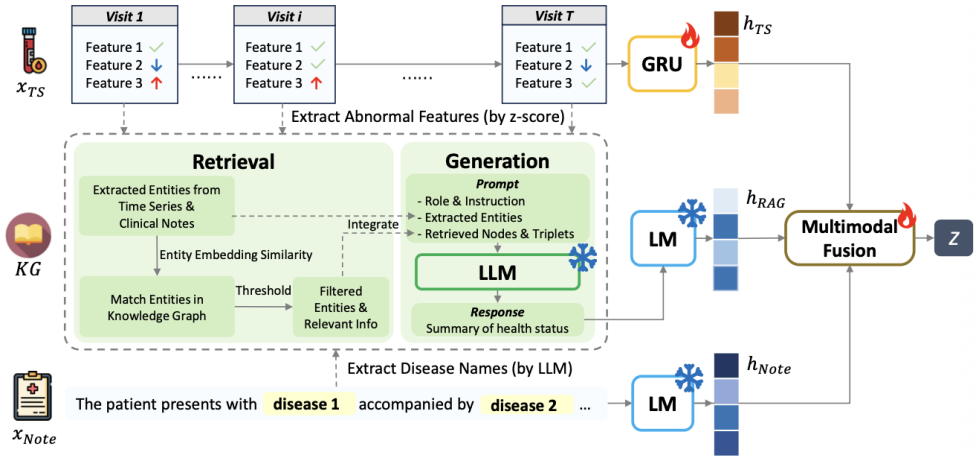


Figure 3.7: Overall architecture of the EMERGE framework, retrieved from the original article [88]. The figure provides a high-level overview of the RAG-enhanced pipeline used to ground multimodal EHR representations in external medical knowledge, with the dashed box highlighting the integration of knowledge retrieval and LLM-based summarisation. "LM" denotes a BERT-based language model [121].

Initial task-oriented studies examine whether LLMs can enhance predictive performance or clinical reasoning in well defined settings. Glicksberg et al. [156] evaluate GPT-4 [164] for predicting hospital admission from emergency department encounters across seven New York City hospitals, directly comparing it with supervised models trained on structured and unstructured EHR data, including an ensemble of Bio-Clinical-BERT [165] and XGBoost [166]. While the conventional ensemble achieved superior standalone performance, AUROC of 88%, naïve zero-shot GPT-4 underperformed. However, augmenting GPT-4 with RAG based on contextually similar historical cases substantially improved performance with further gains observed when numerical admission probabilities from the ML ensemble were provided as input. The authors conclude that LLMs are most effective when used to contextualise and reason over existing predictive signals, rather than as replacements for supervised models in clinical triage.

A complementary perspective is provided by Cohen et al. [157], who compare a clinical LLM, GatorTron [167], with multiple traditional ML models for predicting Antimicrobial Resistance (AMR) and Methicillin-Resistant *Staphylococcus aureus* (MRSA) in hospital-onset sepsis. Structured EHR data from the University of Florida Integrated Data Repository are converted into chronologically ordered textual templates, where structured variables are deterministically mapped to standardised textual descriptions and concatenated into a single paragraph for LLM input, while conventional models rely on manually engineered features. The authors report that GatorTron outperforms traditional approaches for MRSA prediction, achieving an AUROC of 73% compared to 66% for the best-performing ML baseline, alongside higher sensitivity, F1-scores, and a negative predictive value exceeding 90% across most infection syndromes. Although performance for broader AMR prediction remains limited, these findings indicate that clinical LLMs can identify resistant cases using simpler and more readily available feature representations, supporting early clinical decision-making despite persistent microbiological complexity.

Li et al. [158] extend this paradigm with Collaborative Analysis & Risk Evaluation for Alzheimer's Disease (CARE-AD), a multi-agent LLM framework for forecasting Alzheimer's disease onset from longitudinal clinical notes. The framework emulates a multidisciplinary diagnostic process: a fine-tuned LLaMA-based agent [138] extracts relevant clinical signals, specialised agents assess distinct clinical dimensions, and a central Alzheimer's disease specialist agent synthesises their conclusions. Evaluated retrospectively on US VHA data, CARE-AD outperformed single-model baselines at long prediction horizons, achieving 53% accuracy up to ten years before formal diagnosis, compared with 26%–45% for comparator methods. These results demonstrate the potential of multi-agent systems for early risk stratification and more transparent clinical assessment.

Later works shift explicitly towards workflow integration, interpretability, and clinician alignment. Nateghi Haredasht et al. [159] show that LLM-derived features extracted from unstructured clinical notes can improve prediction of six-month treatment retention in medication for opioid use disorder. Their Clinical Entity Augmented Retrieval (CLEAR) pipeline uses Flan-T5 [168] and GPT-4 [164] to extract 13 psychosocial features, which are combined with structured EHR variables in classification and survival models. Across datasets from STAnford Research Repository (STARR) [169] and NeuroBlu [170], the inclusion of LLM-derived features yields consistent performance gains, with the largest relative improvements observed in simpler models such as LR, although overall discrimination remains moderate, achieving an AUROC of 65%. Beyond predictive performance, the authors highlight that LLMs capture clinically salient contextual risks not reliably represented in structured data, supporting more interpretable, clinician-facing risk assessment.

Two related studies by Laaouar and Yanes introduce CLARA [160], [161], a drug recommendation framework combining LLMs with graph NNs. In both variants, LLMs extract symptoms, allergies, and chronic conditions from clinical notes, enrich patient representations within a heterogeneous graph, and generate readable explanations. The first study emphasises contextual feature extraction, semantic representation, and post-hoc explanation, whereas the second focuses more directly on workflow integration, clinician-facing interpretability, and physician validation. On MIMIC-III [101], CLARA outperformed graph-based baselines using only structured data, achieving AUROC values up to 93% and F1-scores above 87%. Ablation studies showed that removing narrative context caused the largest performance degradation, while physician assessments found the generated explanations generally clear and clinically appropriate.

Purely data-driven clinical models often provide limited interpretability for clinical decision-making. Wang et al. propose ColaCare [162], an Multidisciplinary Team (MDT)-inspired framework that combines expert EHR models, multi-agent collaboration, and retrieval-augmented generation. The framework separates quantitative evidence extraction from clinical reasoning. Expert time-series models first derive predictive logits and SHapley Additive exPlanations (SHAP)-based [171] feature attributions from structured EHR data. These outputs are then contextualised through retrieval-augmented alignment with authoritative medical guidelines from the Manual of Diagnosis and Therapy. Clinical decision support is produced through a multi-agent consultation process

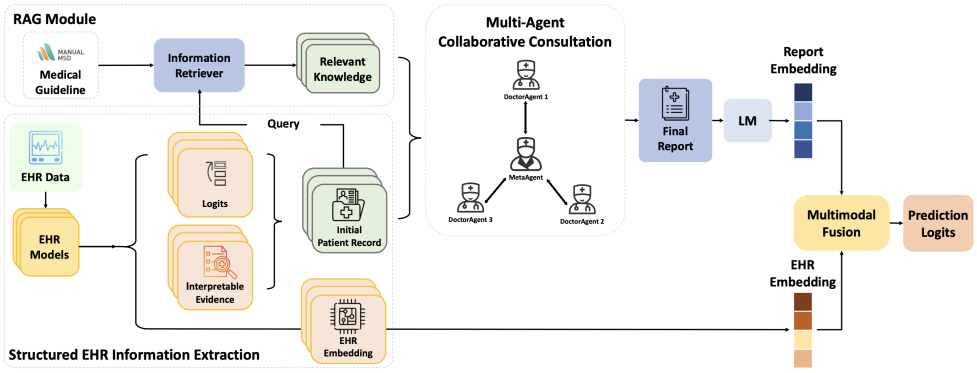


Figure 3.8: Overall architecture of the ColaCare framework, retrieved from the original article [162]. The figure provides a high-level overview of an MDT-inspired, LLM-driven collaborative decision pipeline, where structured EHR evidence extracted by expert models is contextualised via retrieval-augmented medical knowledge and synthesised through multi-agent deliberation to support interpretable clinical prediction.

in which several LLM-based DoctorAgents independently assess patient risk using the model evidence and retrieved knowledge. A MetaAgent subsequently synthesises their assessments into a consensus report. Although ColaCare combines multiple data and knowledge sources, its principal contribution lies in structuring LLM-mediated reasoning within a clinical decision-support workflow rather than in multimodal representation learning. Evaluated on MIMIC-IV [17], CDSL [128], and End-Stage Renal Disease datasets, ColaCare consistently improved AUPRC across prediction tasks. Ablation studies showed marked performance degradation when the expert models were removed, indicating that the LLM agents alone were insufficient without structured predictive evidence. This collaborative design mirrors multidisciplinary clinical decision-making by combining quantitative predictions, feature-level explanations, retrieved medical knowledge, and agent-based deliberation. Figure 3.8 illustrates the separation between evidence extraction and collaborative, LLM-mediated clinical reasoning.

Finally, Lammert et al. introduce Medical Evidence Retrieval and Data Integration for Tailored Healthcare (MEREDITH) [163], an expert-guided LLM system for precision oncology, evaluated through simulated Molecular Tumour Board (MTB) case reviews . Built on Gemini Pro, the pipeline utilises RAG and Chain of Thought (CoT) to integrate PubMed studies, trial databases, drug approval lists, and oncologic guidelines. Benchmarked on expert fictional case vignettes, MEREDITH achieves a high concordance rate with human expert recommendations, 94.7%, while often identifying a broader set of clinically relevant treatment options. The study concluded that replicating the precision oncologist's evaluation process significantly enhances LLM utility, although reliance on proprietary institutional access remains a limitation for wider adoption.

3.3.5 Generative Models for Synthetic EHR

This subsection reviews four model-centric and task-independent studies on synthetic EHR generation, including the works by Miletic and Sariyar [172], Pamisetty et al. [173], Zhong et al. [174],

and Huang et al. [175]. These studies address synthetic data generation as a response to data scarcity and privacy constraints, reflecting a clear methodological progression from benchmarking generators based on transformer against classical Generative Adversarial Networks (GANs), to autoregressive and diffusion-based architectures capable of modelling longitudinal and multimodal EHRs, and finally to systematic evaluations of bias in generated data.

Privacy regulations and the scarcity of public medical datasets hinder research, yet traditional generative models such as GANs often suffer from mode collapse and unstable training. Miletic and Sariyar conducted a benchmark study comparing transformer-based LLMs from the Pythia Suite [176] against Conditional Tabular GAN (CTGAN) for synthetic tabular health data generation [172]. Using datasets such as the Centers for Disease Control and Prevention (CDC) Diabetes Health Indicators, the researchers trained models on varying sample sizes and assessed synthetic quality via random forest classification on real test sets. The results indicated that LLMs tend to perform better at capturing complex joint probability distributions as parameter size increases, with Pythia 1B achieving the highest accuracy. The study concluded that while LLMs are more flexible than GANs, their demand for specialised hardware and the influence of pre-training data are key considerations for real-world usage.

Moving beyond the benchmarking of static tabular generators, Pamisetty et al. propose EHRGPT [173], an autoregressive transformer for generating full longitudinal patient records from structured EHR data. The model adopts a task-independent modelling approach based on the GPT-3 [177] architecture to generate full-trajectory synthetic health records, relying on a unified "master record" representation per admission and a specialised tabular tokenizer to preserve structural integrity during chronological sequence training. Evaluated on MIMIC-IV [17], EHRGPT demonstrates high statistical fidelity, achieving an inverse KL-divergence similarity score of 97.2%, substantially outperforming GAN baselines such as CorGAN [178] and MedBGAN [179] while maintaining low membership and attribute inference risks. The authors conclude that autoregressive language modelling can capture realistic longitudinal EHR trajectories while maintaining acceptable privacy guarantees, positioning EHRGPT as a high-fidelity alternative for synthetic data sharing and downstream model development. Figure 3.9 provides a high-level overview of the EHRGPT workflow, clarifying how structured EHR data are transformed into a unified master record and subsequently modelled autoregressively to generate full longitudinal patient trajectories.

Where EHRGPT adopts an autoregressive language modelling paradigm to generate longitudinal records [173], Zhong et al. introduce MedDiTPro [174], a diffusion transformer specifically designed for multimodal and longitudinal EHR synthesis. Departing from U-Net–based diffusion baselines [180], MedDiTPro architecture employs self-attention for intra-modality code-level interactions and parallel prompt learners to capture global data context and local inter-modality dependencies. Evaluated on MIMIC-III [101], MIMIC-IV [17], and eICU [13], the framework achieves the lowest longitudinal imputation perplexity and the highest synthesis alignment scores across all baselines. Downstream utility assessments further show that predictive models trained on MedDiTPro-generated data yield consistent improvements in mortality prediction, with gains

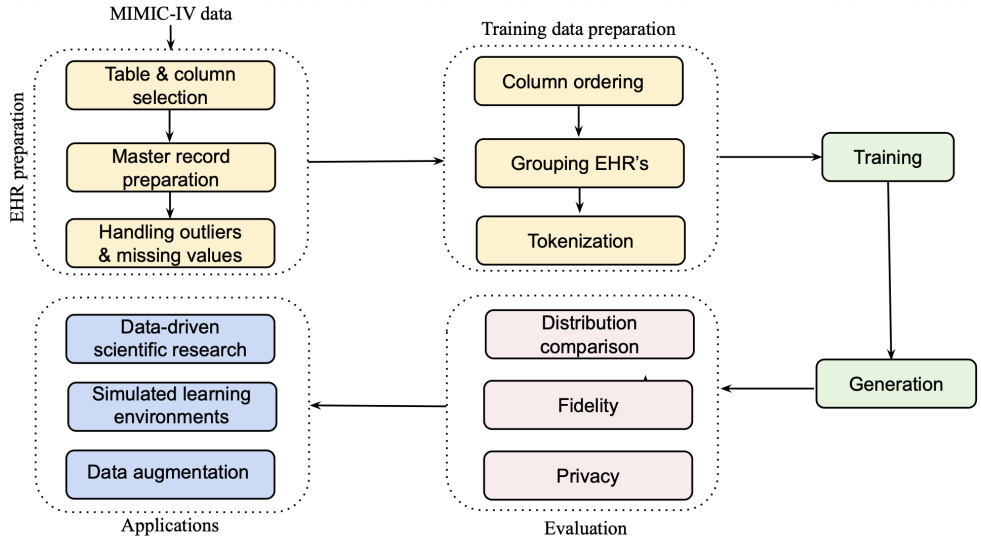


Figure 3.9: High-level workflow of the EHRGPT framework, retrieved from the original article [173]. The diagram illustrates the construction of a unified "master record", its transformation into a sequential tokenised representation, and the autoregressive generation of longitudinal synthetic EHR data, together with downstream evaluation and application stages.

in both AUPRC and F1 Score when synthetic data are combined with real samples. The authors concluded that self-attention mechanisms are pivotal for generating realistic, diverse synthetic records that maintain clinical meaningfulness and high utility in downstream tasks.

While the preceding works primarily prioritise statistical fidelity and downstream predictive utility, Huang et al. shift the focus towards a critical evaluation of demographic bias and representational fairness in LLM-generated synthetic EHRs [175]. The study evaluates approximately 140,000 synthetic records generated across seven open-source LLMs and 20 diseases, introducing the Electronic Health Record Performance Score (EPS) to assess completeness and Statistical Parity Difference (SPD) to quantify bias. Larger models consistently achieve higher completeness, with Yi-34B [181] reaching an EPS of 96.8. However, this performance gain is accompanied by amplified demographic distortions. For female-dominated diseases, Qwen-14B [182] generates 97.3% female records compared with 75.8% in real data, while Hispanic populations are systematically under-represented across models. The authors identify a clear performance–bias trade-off, demonstrating that high-fidelity synthetic generation does not ensure demographic representativeness and may risk reinforcing existing healthcare inequities without explicit bias aware evaluation and mitigation strategies.

3.3.6 Summary and Synthesis

Across the reviewed literature, a clear trajectory emerges in the use of LLMs for clinical data, progressing from task-bound predictive enhancements towards more generalisable representational and system-level roles. Early studies primarily explore how Transformer architectures and LLMs can address long-standing limitations of structured and longitudinal EHR data, by improving how

medical codes are represented, enabling faster adaptation to new clinical tasks with limited labelled data, and making structured records compatible with pre-trained language models [129], [130], [133]. Subsequent works extend this paradigm beyond prediction-centric approaches, demonstrating that instruction tuning, unsupervised clustering, and knowledge-guided representation learning allow LLMs to operate over heterogeneous and longitudinal clinical data without being tied to a specific predictive task [131], [132], [134], [148]. In parallel, multimodal frameworks increasingly position LLMs as semantic mediators between disparate data sources, whether through standards-compliant pipelines, prompt-based unification of heterogeneous tabular features, or retrieval and contrastive-based integration of external medical knowledge [88], [145]–[147].

Complementing these representational advances, a second body of work illustrates how LLMs are being embedded into broader clinical DSSs, where predictive accuracy is contextualised within interpretability, workflow integration, and clinician-aligned reasoning. Across diverse clinical settings, LLMs consistently demonstrate their greatest utility when augmenting, contextualising, or synthesising evidence derived from conventional predictive models rather than replacing them outright [156]–[158]. Workflow-integrated systems further emphasise transparency and clinical plausibility by grounding LLM outputs in curated guidelines, knowledge graphs, or multidisciplinary consultation paradigms, as seen in CLEAR [159], CLARA [160], [161], ColaCare [162], and MEREDITH [163]. Finally, studies on synthetic EHR generation underscore that gains in how realistically synthetic records resemble real data and how useful they are for training clinical prediction models must be considered alongside privacy preservation and demographic bias, reinforcing that technical realism alone is insufficient for clinical applicability [172]–[175]. Together, these findings establish a common foundation for critically examining how representational choices, evaluation practices, and system-level design decisions shape the clinical reliability and generalisability of LLM-enabled healthcare systems.

Table 3.2 provides a compact overview of the reviewed studies, consolidating each article's methodological role, data modalities and datasets, clinical tasks, public code availability, primary contribution, and main limitation. Public code is reported as available when the article provides an accessible implementation repository or project website, and as available upon request when access depends on contacting the authors. In this sense, this table functions as a reference scaffold for the review, anchoring the deconstructed discussion of individual works in a unified comparative summary.

Table 3.2: Structured summary of the 23 included generative AI approaches applied to medical data.

Article	Approach Role	Data Modalities and Datasets	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Cohen et al. (2025) [157] *GatorTron*	Clinical LLM evaluation	Structured EHR; UF Integrated Data Repository	AMR and MRSA prediction	No	GatorTron achieves higher sensitivity than traditional ML using simplified, readily available features.	Performance for broad AMR prediction remains limited; Evaluated within a single healthcare system.
Dwivedi et al. (2024) [129] *UniStruct*	Representation learner	Structured EHR + clinical text; Cheng Hsin General Hospital and EHRSHOT	Medical code and clinical outcome prediction	No	Adapted tokenisation preserves medical code semantics and improves generalisation across tasks.	Structured and unstructured modalities remain encoded in separate representation spaces.
Engelke et al. (2025) [145] *FHIR-Former*	Predictive model within interoperable framework	Structured EHR + clinical text (FHIR); Essen University Hospital	Readmission, mortality, imaging, and ICD-10-GM prediction	Yes	Eliminates manual feature engineering and supports deployment within FHIR-compliant infrastructures.	Interpretability is limited and generalisability beyond inpatient settings is unclear.
Glicksberg et al. (2024) [156]	Reasoning augmentor	Structured EHR + clinical text; Mount Sinai Health System	Emergency department admission prediction	No	RAG enables LLM reasoning to contextualise traditional ML predictions effectively.	Standalone LLM performance is inferior to supervised ML; Evaluated in urban hospital systems.
Hemamalini et al. (2025) [146] *MFSL*	Multimodal representation learner	Structured EHR + clinical text; MIMIC-IV	Few-shot disease prediction	No	Contrastive multimodal alignment enables robust learning with very limited labelled data.	Dependence on templated text generation and external LLMs may hinder reproducibility.

Continued on next page

Article	Approach Role	Data Modalities and Datasets	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Huang et al. (2025) [175]	Synthetic data evaluation and bias analysis	Synthetic multimodal EHR; Synthetic EHR dataset	Synthetic EHR generation and bias evaluation	No	Evaluation framework reveals performance-bias trade-offs in LLM-generated EHRs.	Generation quality improvements amplify demographic biases without mitigation strategies.
Kang et al. (2025) [134] *LLM-DG*	Knowledge-enhanced clinical predictor	Structured EHR (longitudinal) + clinical text; MIMIC-III and MIMIC-IV	Diagnosis and heart failure prediction	Yes (upon request)	LLM-enriched representations capture inter- and intra-patient relationships effectively.	High computational cost associated with LLM-based feature extraction.
Laaouar & Yanes (2025a) [160] *CLARA*	DSS	Structured EHR + clinical text; MIMIC-III subset	Medication recommendation and explanation generation	Yes	Multi-level LLM integration improves predictive accuracy and human-readable explanations.	Temporal dynamics across longitudinal visits are not explicitly modelled.
Laaouar & Yanes (2025b) [161] *CLARA*	DSS	Structured EHR + clinical text; MIMIC-III	Medication recommendation and explanation generation	No	Semantic integration of narrative context substantially improves recommendation performance.	Latency and operational cost of LLM components limit real-time deployment.
Lammert et al. (2024) [163] *MEREDITH*	DSS	Multimodal EHR + genomic data + medical guidelines; Fictional oncology case vignettes	Precision oncology treatment recommendation	Yes	Expert-guided RAG closely mirrors molecular tumour board reasoning and reduces hallucinations.	Evaluated only on fictional case vignettes and at limited scale.

Continued on next page

Article	Approach Role	Data Modalities and Datasets	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Li et al. (2025) [158] *CARE-AD*	Multi-agent clinical decision-support framework	Structured EHR + clinical text; VHA Corporate Data Warehouse	Alzheimer's disease onset prediction and symptom extraction	Yes (upon request)	Multi-agent design captures complementary specialist perspectives and improves long-horizon prediction.	Selection bias and reliance on retrospective notes limit generalisability.
Miletic & Sariyar (2024) [172]	Benchmark and evaluation framework	Structured EHR; CDC Diabetes, Adult, and Body Signals datasets	Synthetic health data generation and benchmarking	No	Benchmark shows that LLMs outperform GANs in modelling complex joint distributions.	High computational cost for larger LLMs limits accessibility.
Nateghi Haredasht et al. (2025) [159] *CLEAR*	Feature extractor	Structured EHR + clinical text; STARR and NeuroBlu	Six-month Opioid Use Disorder treatment retention prediction	Yes	LLM-derived features uncover psychosocial risks absent from structured codes.	Overall predictive performance remains modest.
Pamisetty et al. (2024) [173] *EHRGPT*	Synthetic data generator	Synthetic EHR (longitudinal); MIMIC-IV	Synthetic EHR generation	No	Preserves longitudinal structure while achieving a strong fidelity-privacy trade-off.	Evaluation focuses on statistical similarity rather than downstream utility.
Qiu et al. (2025) [131] *VaDeSC-EHR*	Representation learner	Structured EHR (longitudinal); UK Biobank	Survival-based patient clustering and risk prediction	Yes	Joint modelling of trajectories and survival uncovers clinically meaningful subgroups.	Limited integration of clinical text.
Shoham and Rappoport (2024) [133] *CPLLM*	Clinical predictor	Structured EHR (longitudinal); MIMIC-IV and eICU	Disease and hospital readmission prediction	Yes	LLMs effectively model long temporal sequences without medical pre-training.	High computational footprint compared with classical models.

Continued on next page

Article	Approach Role	Data Modalities and Datasets	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Wang et al. (2023) [130] *SptEHR*	Representation learner	Structured EHR; MIMIC-III	Zero-shot and few-shot disease prediction	No	Task scaling through soft prompt transfer enables efficient adaptation to rare diseases.	Performance depends on similarity between source and target tasks.
Wang et al. (2025) [162] *ColaCare*	Multi-agent DSS	Structured EHR (time-series) + medical guidelines; MIMIC-IV, CDSL, and ESRD	Mortality and readmission prediction	Yes	MDT-inspired multi-agent reasoning improves transparency and predictive performance.	System complexity and inference cost may hinder deployment.
Wu et al. (2024) [132] *Llemr*	Instruction-tuned EHR framework	Structured EHR (longitudinal); MIMIC-Instr	EHR question answering and clinical prediction	Yes	Large-scale instruction tuning bridges the schema gap and enables general-purpose EHR reasoning.	Requires substantial curated data and multi-stage training.
Xu et al. (2025) [148] *DearLLM*	Feature interaction modeller	Structured EHR; MIMIC-III and MIMIC-IV	Mortality prediction	Yes	LLM-deduced feature correlations enable explicit modelling of non-linear clinical interactions.	Correlation extraction relies on LLM inference rather than causal validation.
Yang et al. (2025) [147] *TabPF*	Prompt-based clinical predictor	Structured EHR; eICU and CECMed	Severity, mortality, and length-of-stay prediction	No	Prompt learning exploits heterogeneous tabular data without architecture redesign.	Performance depends on prompt engineering and task-specific tuning.

Continued on next page

Article	Approach Role	Data Modalities and Datasets	Clinical Tasks	Public Code	Primary Contribution	Main Limitation
Zhong et al. (2025) [174] *MedDiTPro*	Synthetic data generator	Synthetic EHR (multimodal, longitudinal); MIMIC-III, MIMIC-IV, eICU, and Breast Cancer Trial	Multimodal data synthesis and risk-prediction augmentation	Yes	Diffusion-based generation preserves temporal coherence across heterogeneous modalities.	Evaluation of downstream clinical utility remains limited.
Zhu et al. (2024) [88] *EMERGE*	Retrieval-augmented multimodal predictive framework	Structured EHR (time-series) + clinical text; MIMIC-III and MIMIC-IV	In-hospital mortality and readmission prediction	Yes	RAG-driven integration of external medical knowledge improves robustness under data sparsity.	Increased pipeline complexity and dependence on high-quality knowledge graphs.

3.4 Critical Review of the Literature

This section critically integrates the two state-of-the-art analyses reviewed in this chapter: Medical ICU data Structuring Pipelines and Generative AI Models applied to Medical Data. It examines their conceptual and methodological relationships, identifies assumptions limiting clinical applicability, and articulates the gaps that motivate this dissertation.

The scope of this review is shaped by its search strategy. Only studies retrieved through the predefined search strings were considered, meaning that relevant works using alternative terminology may not have been captured. For example, while EHR was included, related terms such as Electronic Medical Record (EMR) were not systematically incorporated. Expanding the search to encompass all adjacent terminology would have substantially increased the number of retrieved studies beyond the practical scope of this dissertation. Incomplete coverage of related literature is therefore acknowledged as a limitation.

The two bodies of literature converge on enabling reliable, generalisable, and clinically meaningful decision support from complex ICU data. Data structuring pipelines have evolved from institution-specific preprocessing routines into increasingly reusable frameworks, including CDMs, extraction layers, and benchmarking workflows. In parallel, generative AI and LLMs have demonstrated substantial capacity to model longitudinal, heterogeneous, and multimodal data while supporting abstraction, reasoning, and explanation. Nevertheless, integration between robust data pipelines and generative clinical intelligence remains fragmented, limiting the translational potential of both areas.

From the perspective of ICU data structuring, a central strength of the state of the art lies in the explicit formalisation of harmonisation processes as reusable artefacts. Frameworks such as OMOP and openEHR-based pipelines, as well as reusable extraction layers such as `ricu` and multi-database pipelines such as METRE, demonstrate that semantic alignment, temporal consistency, and unit standardisation can be systematically enforced across heterogeneous critical care databases [100], [103], [106], [107]. These approaches have substantially improved reproducibility and enabled cross-dataset benchmarking. Nevertheless, they often prioritise interoperability and structural consistency over downstream clinical expressiveness, resulting in representations that are technically aligned but not optimally suited for advanced representation learning or clinical reasoning.

Conversely, task-driven pipelines designed for phenotype definition, benchmarking, or specific prediction objectives achieve strong internal consistency and clinical validity, as illustrated by OpenSep, pyAKI, FIDDLE, and YAIB [109]–[112]. Yet, their deliberate narrowing of scope constrains reuse beyond the original task formulation. At the opposite extreme, pipelines tightly coupled to specific graph-based, mechanistic, or representation-learning architectures achieve coherent end-to-end solutions but sacrifice generalisability and modularity [114], [116], [118]. Across these categories, a persistent gap emerges: few pipelines are explicitly designed to serve as stable, clinically grounded substrates for modern generative or reasoning-based models.

The literature on generative AI applied to medical data addresses many of the representational

limitations inherent to structured EHR modelling. Transformer-based architectures and LLMs have shown clear benefits in preserving temporal dependencies, encoding hierarchical medical knowledge, and adapting across tasks under limited supervision [129], [130], [133]. Instruction-tuned and knowledge-enriched models further demonstrate that LLMs can bridge the schema gap between structured clinical data and natural language reasoning, enabling more flexible interaction with longitudinal patient records [132], [134]. However, these advances are predominantly model-centric, data representations are frequently reshaped to suit the architecture, rather than emerging from clinically principled, reusable data pipelines.

Multimodal, knowledge-guided, and retrieval-augmented frameworks further highlight this tension. Systems such as FHIR-Former, TabPF, DearLLM, and EMERGE position LLMs as semantic mediators across heterogeneous data sources and external medical knowledge [88], [145], [147], [148]. While these approaches can improve predictive performance, particularly under data sparsity, they introduce additional complexity and computational overhead, and remain highly sensitive to prompt design, knowledge source quality, and institutional context. Importantly, the underlying data pipelines are rarely decoupled from the modelling logic, complicating reproducibility and limiting transfer across settings.

A similar pattern is observed in workflow-integrated clinical DSSs. Frameworks such as CLARA, CARE-AD, ColaCare, and MEREDITH demonstrate that LLMs deliver the greatest clinical value when augmenting, contextualising, and synthesising evidence derived from structured models and curated guidelines, rather than acting as standalone predictors [158], [160], [162], [163]. These systems move decisively towards interpretability, human–AI collaboration, and alignment with real clinical workflows.

Across both reviews, a fundamental disconnect becomes apparent. Data harmonisation pipelines are largely designed independently of generative reasoning requirements, while LLM-based systems frequently assume that data representations can be flexibly transformed without compromising clinical validity. This lack of co-design results in systems that may be interoperable but limited in clinical expressiveness, or expressive but difficult to reproduce and transfer across institutions. As a consequence, few existing approaches simultaneously achieve interoperability, clinical grounding, interpretability, and adaptability in ICU decision support.

These observations point towards a clear research opportunity. Future ICU DSSs should build upon clinically principled and reusable data pipelines that explicitly support advanced predictive and generative models, rather than treating harmonisation as a peripheral preprocessing step. At the same time, LLMs should be positioned as mediators of evidence, explanation, and synthesis, grounded in structured and validated clinical signals, rather than as opaque end-to-end decision engines. Frameworks that separate quantitative evidence extraction from generative clinical reasoning, while aligning both with multidisciplinary workflows, offer a promising direction [162]. This dissertation addresses a foundational part of this broader challenge by investigating reusable ICU data harmonisation, benchmark-oriented modelling, and cross-database evaluation, while examining the feasibility of connecting an external model architecture to harmonised pipeline outputs.

Chapter 4

Methodology

This chapter describes the methodological framework followed in this dissertation, from data preparation to model evaluation. In line with the objectives defined in Chapter 1, it focuses on the practical steps required to process and harmonise heterogeneous ICU databases, generate structured representations suitable for model training, define clinically relevant prediction tasks, and evaluate predictive models under a unified validation and cross-dataset comparison framework.

The chapter begins by presenting the research approach and data sources. It then describes the adaptation of the data processing workflow, the definition of the prediction tasks, and the modelling strategies explored, including baseline models, METRE-derived architectures, and the adapted external model, EMERGE. The final sections address hyperparameter optimisation, internal and external evaluation, and the procedures used to assess model robustness across databases, providing the methodological basis for the experimental analysis presented in the following chapter.

4.1 Research Approach

Based on the objectives of this dissertation and on the findings of the literature review, the methodology was structured around two complementary stages. First, an open-source pipeline for the extraction and harmonisation of structured ICU data was selected and improved. Second, a benchmark-oriented modelling workflow was built on top of the generated representation, allowing the same processed data representation to be evaluated under the same experimental protocol and across different databases.

The selection of the base data pipeline was guided by three criteria: availability of open-source code, support for cross-dataset validation across more than one critical care database, and sufficient methodological documentation to allow reproduction and adaptation. The selected pipeline also needed to operate as a reusable harmonisation framework rather than being tightly coupled to a single model architecture, while supporting clinically relevant prediction tasks and comparative experiments for model generalisation.

Among the pipelines reviewed in Section 3.2.3, `ricu` and METRE were the most directly relevant, as both provide reusable approaches for processing heterogeneous ICU data [100], [103]. `ricu` offers a comprehensive extraction framework, supporting multiple databases and clinical concepts with extensive documentation and active community development. However, it functions primarily as a harmonisation and concept-mapping layer rather than as a task-oriented modelling pipeline. METRE was therefore selected as the primary pipeline because it combines reusable data harmonisation with explicitly defined clinical prediction tasks and a modelling framework designed for cross-database evaluation between MIMIC-IV and eICU.

After the adaptation of METRE, an external model was selected to assess whether the structured outputs generated by the pipeline could be reused beyond the models originally implemented by its authors. This provided an additional benchmarking layer, in which the pipeline was evaluated not only through its native models, but also through its ability to support an independent modelling architecture. Candidate external models were considered according to their compatibility with longitudinal structured EHR data, availability of implementation details or reusable code, and feasibility of adaptation to the METRE-generated TS representation within the scope and timeframe of this dissertation.

Within the reviewed models, EMERGE was selected because it had been explicitly evaluated in a TS-only configuration using structured EHR data from MIMIC-IV, included overlapping prediction tasks, and provided a published comparison against its own multimodal variants [88]. This allowed the model to be assessed under controlled conditions using the same structured data representation as the METRE models. Other open-source models reviewed were not prioritised because their input assumptions were not directly aligned with the outputs of the adapted METRE pipeline: FHIR-Former [145] depends on FHIR-formatted data, MedDiTPro [174] addresses synthetic data generation, and CLARA [160], VaDeSC-EHR [131], and CLEAR [159] target tasks not directly comparable with the selected METRE prediction tasks. Among the remaining candidates, ColaCare [162] required externally generated clinical reports and repeated LLM inference per patient, whereas CPLLM [133] and DearLLM [148] operate on diagnostic code sequences rather than continuous multivariate time series, making EMERGE the most compatible option for controlled reuse of the METRE representation. These models remain relevant for future extensions, but would require additional modalities, alternative task definitions, or architectural adaptations beyond the structured TS benchmark implemented in this dissertation.

4.2 Datasets

This dissertation used two publicly available critical care databases: MIMIC-IV [17] and the eICU Collaborative Research Database [13]. These datasets were selected because they are among the most widely used resources for intensive care research and because they are supported by the original METRE pipeline. Their combined use was particularly relevant for this work because they differ in institutional origin, database structure, and clinical documentation practices, while still providing overlapping types of structured ICU data.

4.2.1 MIMIC-IV

MIMIC-IV is a publicly available, de-identified clinical database developed by the MIT Laboratory for Computational Physiology and Beth Israel Deaconess Medical Center [17]. The database contains data from patients admitted to the emergency department or intensive care units at Beth Israel Deaconess Medical Center between 2008 and 2022. It includes data for over 65,000 patients admitted to an intensive care unit and over 200,000 patients admitted to the emergency department.

MIMIC-IV is organised into modules that reflect the provenance of the data. The `hosp` module contains hospital-wide information, including patient demographics, admissions, transfers, diagnoses, procedures, laboratory measurements, microbiology events, medication prescriptions, pharmacy records, and billing-related codes. The `icu` module contains data sourced from the MetaVision ICU clinical information system, including ICU stays, charted observations, intravenous and fluid inputs, outputs, procedures, date-time events, and other time-stamped bedside measurements. These modules are linked through de-identified identifiers such as `subject_id`, `hadm_id`, and `stay_id`, allowing patient trajectories to be reconstructed across hospital and ICU-level events.

Direct patient identifiers are removed, and dates are shifted into the future using a patient-specific offset while preserving temporal intervals within each patient record. This allows longitudinal analyses to be performed while preventing direct identification of admission dates. Access is provided through PhysioNet and requires completion of certified training in human subjects research and acceptance of a data use agreement. In this dissertation, MIMIC-IV version 3.1 was used.

4.2.2 eICU

The eICU Collaborative Research Database is a publicly available, multi-centre critical care database developed by the MIT Laboratory for Computational Physiology in collaboration with Philips Healthcare [13]. The database contains de-identified data associated with ICU admissions across the United States between 2014 and 2015.

In contrast to MIMIC-IV, which originates from a single academic medical centre, eICU includes data from multiple hospitals and care units. The database comprises over 200,000 patient unit encounters from over 139,000 unique patients admitted to 335 units across 208 hospitals. Patients were selected for inclusion in the public database from hospital discharges, using a stratified sampling strategy based on hospital representation in the original private repository.

The database includes a wide range of structured critical care information, including vital signs, laboratory measurements, medications, admission diagnoses, patient history, APACHE severity of illness components, care plan documentation, time-stamped diagnoses, and treatment information. Data are stored in a relational format and linked primarily through identifiers such as `patientUnitStayID`. As in MIMIC-IV, the relational structure allows multiple data streams to be combined into patient-level temporal representations, although the underlying schema and documentation conventions differ between the two databases.

All tables were de-identified in accordance with the safe harbour provisions of the Health Insurance Portability and Accountability Act, including removal of protected health information and hospital identifiers. Like MIMIC-IV, access is provided through PhysioNet and requires completion of human research training and acceptance of the corresponding data use agreement. In this dissertation, eICU version 2.0 was used.

4.3 METRE-Based Data Pipeline

As described in Section 4.1, METRE was selected as the base data pipeline because it provides an open-source workflow for multi-database extraction, harmonisation, and validation in critical care research. The original repository, proposed by Liao and Voldman, includes BigQuery-based extraction, harmonised table construction, train-development-test splitting, and downstream clinical prediction models [103].

In this dissertation, the METRE codebase was adapted to support the selected database versions, reconstruct required intermediate components, generate model-ready files, and standardise the downstream training and evaluation workflow. These adaptations were necessary because some implementation details required for full end-to-end execution were not directly available in the public repository. In particular, some database tables expected by the extraction scripts were not directly present in the source databases and therefore had to be reconstructed from available clinical tables, suggesting that they had been generated as auxiliary intermediate tables in the original workflow. In addition, the transition between extracted HDF5 files and the NumPy Array (NPY) inputs expected by the training code was not provided, and some model architectures reported in the original article were not directly implemented in the released training pipeline. Clarification was sought from the original authors, but no additional implementation artefacts were made available during the development of this dissertation. Therefore, the methodology included both the use of the original METRE workflow and the implementation of additional processing steps needed to obtain reproducible model inputs for all selected tasks and databases.

This section first summarises the original METRE pipeline and then details the adaptations required for its use in this dissertation, including database query updates, reconstruction of auxiliary tables, HDF5-to-NPY conversion, integrity validation, and preparation of the training workflow.

4.3.1 Original Pipeline

As introduced in the literature review, METRE is organised around two main stages: data extraction and model training. The extraction stage is executed after access to MIMIC-IV and eICU has been configured through PhysioNet and Google Cloud. The released repository assumes access to both databases through Google BigQuery and provides scripts in which the user specifies the database to extract, the Google Cloud billing project, cohort restrictions such as minimum age and ICU length of stay, optional patient groups, and the temporal aggregation window. The repository documentation refers to MIMIC-IV v2.2 and eICU v2.0 as the supported database sources at the time of release.

Operationally, the extraction pipeline starts from cohort selection and then queries static, time-dependent, and intervention-related records from the source database. The retrieved variables are semantically grouped, temporally aligned relative to ICU admission, and aggregated into a common hourly representation. This process produces three harmonised outputs, corresponding to the same structure illustrated in Figure 3.1.

- **Static table**: contains patient-level and admission-level variables, including demographics, admission characteristics, comorbidities, and outcome labels.
- **Vital table**: contains time-dependent physiological, laboratory, microbiology, urine output, and clinical assessment variables. These variables are aggregated into hourly bins by default. For numerical variables, METRE stores both the aggregated value and a measurement indicator, allowing missingness patterns to be represented explicitly. Categorical microbiology variables are semantically grouped and encoded to preserve compatibility across databases.
- **Intervention table**: contains binary time-dependent variables indicating whether a treatment or procedure was active at each hour of the ICU stay. These variables include mechanical ventilation, antibiotics, vasoactive agents, heparin, continuous renal replacement therapy, red blood cell transfusion, platelet transfusion, fresh frozen plasma transfusion, colloid bolus, and crystalloid bolus.

After extraction, the pipeline applies outlier removal, missing data handling, and normalisation. Numerical measurements are filtered using predefined physiologically plausible ranges, with the same outlier-removal criteria applied across MIMIC-IV and eICU after the variable mapping and harmonisation steps. This is important because equivalent clinical concepts may be recorded differently across databases, including differences in item definitions, documentation practices, or measurement units. Missing values are handled after temporal aggregation, and measurement indicators are converted into binary variables reflecting whether a value was observed. Continuous variables are standardised, with the original workflow allowing eICU variables to be normalised using statistics derived from MIMIC-IV. The processed tables are then split into train, development, and test partitions. In the released code, this corresponds to a 70/10/20 split, although the training stage subsequently combines the train and development partitions for cross-validation. The final extraction outputs are saved as HDF5 files containing the `vital`, `inv`, and `static` tables for each split.

The model input produced by the pipeline is a structured multivariate time series for each ICU stay. Each sample corresponds to one ICU stay and contains the time-dependent variables extracted during the selected observation window. In practice, the hourly vital, laboratory, microbiology, missingness, and intervention variables are arranged as a feature-by-time matrix, while the corresponding task label is obtained from the static table. This representation allows the same extracted data to be used by different temporal models without returning to the raw relational databases.

The original pipeline also defines the clinical prediction tasks used for downstream model evaluation. These tasks are part of the operational design of the pipeline, since each task determines the eligible cohort, observation window, target label, and temporal exclusion criteria applied before model training. Five prediction settings were considered: hospital mortality based on the first 48 hours of ICU data, ARF based on either the first 4 or 12 hours of ICU data, and shock based on either the first 4 or 12 hours of ICU data. In each setting, these initial hours constitute the observation window made available to the model, while the gap period separates the end of this window from the earliest eligible event onset. The original METRE study adopted a 6-hour gap for its reported experiments, and the same value was maintained in this dissertation.

Hospital Mortality

Hospital mortality was formulated as a binary classification task in which the model estimates the probability of death before hospital discharge. Following the original configuration, the input consisted of the first 48 hours of ICU data. ICU stays with insufficient length-of-stay (LOS) to provide the required observation window and gap period were excluded. The target label was obtained from the hospital mortality information available in the structured admission records. Positive cases corresponded to patients who died during the hospital admission, whereas negative cases corresponded to patients discharged alive.

Acute Respiratory Failure

ARF was defined as the subsequent need for positive-pressure respiratory support. Two prediction settings were evaluated, using either the first 4 hours or the first 12 hours of ICU data as input. ARF labels are generated from the extracted ventilation-related variables, including mechanical ventilation and positive end-expiratory pressure. Patients who already met the respiratory support criteria during the observation window or gap period were excluded from the corresponding task. Positive cases corresponded to patients whose first eligible respiratory support event occurred after the observation window and gap period, whereas negative cases corresponded to patients with no such onset during the ICU stay.

Shock

Shock was defined through the subsequent initiation of vasoactive therapy. As with ARF, two observation windows were considered: 4 hours and 12 hours after ICU admission. The label was generated from the vasoactive intervention variables extracted by METRE, including norepinephrine, epinephrine, dopamine, vasopressin, and phenylephrine. Patients already receiving vasoactive therapy during the observation window or gap period were excluded from the corresponding task. Positive cases corresponded to patients whose first vasoactive treatment occurred after the observation window and gap period, whereas negative cases corresponded to patients with no vasoactive treatment onset during the ICU stay.

After defining the extracted representation and prediction tasks, the original METRE workflow proceeds to model training and evaluation. The repository includes several modelling approaches, although they are not all integrated into a single execution workflow. LR and RF baselines are provided in a separate notebook, with hyperparameter search performed through Bayesian optimisation. The main training script implements NN models, including a TCN, RNN, and a Transformer encoder. The models evaluated in this dissertation and the adaptations made to standardise their execution are described in Section 4.4.

For model evaluation, the original training workflow combines the training and development partitions and applies a 10-fold cross-validation procedure during model development. For each fold, a model is trained on the corresponding training subset and monitored on the validation subset. The trained models are then evaluated on the held-out test set of the source database and, for cross-database validation, on the other database. The original reported workflow focuses primarily on AUROC and AUPRC, with 95% confidence intervals estimated from 1000 bootstrap iterations, each drawing a resampled subset of 1000 test instances. Based on inspection of the released training code, the final reported test results are obtained from a selected fold model rather than from all cross-validation models jointly. This evaluation strategy was reviewed and modified in this dissertation, as described in Section 4.6.

4.3.2 Reproduction and Adaptation

The implementation first followed the workflow described in the original METRE repository, from database access and extraction to the execution of the training scripts. Access to MIMIC-IV and eICU was configured through PhysioNet and Google Cloud, and the extraction pipeline was executed through BigQuery using the command-line interface provided by the authors. The extraction stage retained the original structure of METRE, from cohort filtering and temporal aggregation to normalisation and generation of split HDF5 files for the static, vital, and intervention tables. Since the released repository did not include a fully pinned requirements file, a stable execution environment was also configured.

Direct reproduction of the released workflow required several adaptations. These were mainly related to changes in MIMIC-IV since the release of the original repository, missing or differently structured derived tables, incompatibilities between the extraction outputs and the training inputs, and the need to make the final training workflow reproducible across all tasks and models. The following subsections describe these adaptations in the order in which they occur in the pipeline.

4.3.2.1 Adaptation of MIMIC-IV queries

The original METRE repository was developed for an earlier release of MIMIC-IV, v2.2. In this dissertation, MIMIC-IV v3.1 was used, as it was the most recent version selected during methodology of the work. The SQL extraction scripts were therefore updated to reference the corresponding v3.1 BigQuery schemas, including `physionet-data.mimiciv_3_1_hosp`, `physionet-data.mimiciv_3_1_icu`, and `physionet-data.mimiciv_3_1_derived`.

These updates affected cohort selection, static patient information, laboratory and charted measurements, derived clinical variables, and intervention-related queries.

Several schema-level changes were required to preserve compatibility with the original METRE extraction logic. In the MIMIC-IV static queries, demographic extraction was updated to use the `race` field, available in the v3.1 derived tables, instead of `ethnicity`. The urine output query was also updated to retrieve urine output rate information from `mimiciv_3_1_derived.urine_output_rate`. In addition, the combined chart and laboratory query was adjusted so that unit-of-measure fields such as `valueuom` were removed before downstream aggregation, since these string-valued columns were not required for the numerical feature construction and could interfere with group-level operations.

These changes were restricted to compatibility with the updated database schema. No new clinical prediction task or additional feature family was introduced, the objective was just to preserve the original METRE feature definitions and extraction structure while making the pipeline executable against MIMIC-IV v3.1.

4.3.2.2 Reconstruction of auxiliary derived tables

During the adaptation to MIMIC-IV v3.1, two auxiliary tables expected by the extraction code were not available in the required form: `culture` and `heparin`. These tables could not simply be removed without changing the feature schema expected by the rest of the pipeline. The `culture`-related variables contribute to the `vital` feature representation, while `heparin` is one of the intervention variables expected in the intervention table. Removing either component would alter the feature layout used by the models and reduce comparability with the original METRE representation.

The expected interface of the `culture` table was inferred from the METRE function that queried this table, together with the downstream processing steps that consumed its columns and the JSON files defining culture-site mappings and feature ordering. Although the original SQL used by the authors to create this auxiliary table was not available, the extraction code expected fields corresponding to patient and admission identifiers, culture timing, specimen type, screening status, culture positivity, and antimicrobial sensitivity availability. The reconstructed table was therefore derived from `microbiologyevents`. Culture positivity was defined by the presence of an organism name, while sensitivity availability was defined by the presence of an antibiotic name. Since no direct equivalent for the original `screen` field was identified in the same form, this field was retained as null to preserve the expected schema.

The expected `heparin` structure was inferred from the intervention extraction function and the intervention feature ordering used by the pipeline. The reconstructed table was derived from `inputevents`, selecting records whose `itemid` values corresponded to heparin-related administrations according to the `d_items` metadata table. The resulting table retained the patient, admission, and stay identifiers, timing information, item label, amount, unit, rate, and rate unit. In the absence of an explicit auxiliary table definition from the original implementation, all identified heparin-related `itemid` values were included to preserve the intervention definition as broadly as

possible. The auxiliary SQL queries used to reconstruct the `culture` and `heparin` tables are provided in Appendix A.

4.3.2.3 Conversion from HDF5 to NPY

After successful extraction, the METRE pipeline produces HDF5 files containing the processed `vital`, `inv`, and `static` tables for each split. However, the NN training code used in the repository expects compiled NPY files containing patient-level arrays and labels. As no standalone conversion script was available in the released repository, a dedicated HDF5-to-NPY conversion script was implemented.

For each split, the converter loads the corresponding vital, intervention, and static tables. The time-dependent tables are grouped by ICU stay identifier, ordered temporally, and converted into patient-level matrixes. The vital and intervention matrixes are then concatenated along the feature axis, producing one multivariate TS array per ICU stay. The resulting arrays follow the format expected by the training code, with entries such as `train_head`, `dev_head`, and `test_head`, together with the corresponding static labels.

The feature-by-time representation was defined to match the format consumed by the METRE training code. In the final configuration, each sample was represented as a matrix with shape $(200, T)$, where T is the number of hourly time points available for the ICU stay. The feature axis contained 200 time-dependent variables, corresponding to 184 vital, laboratory, microbiology, and missingness-related features, plus 16 intervention features. The time axis is from which task-specific observation windows were later selected during training.

4.3.2.4 Integrity validation and feature alignment

A dedicated validation procedure was developed alongside the conversion script to ensure that the generated NPY files preserved the intended structure of the extracted data. These checks verified that the static, vital, and intervention tables contained consistent patient identifiers, that the temporal indices of the vital and intervention streams matched within each ICU stay, and that the expected number of features was present before the final arrays were saved.

Feature ordering was also explicitly validated. This was particularly important for cross-database evaluation, since MIMIC-IV and eICU must share the same feature positions for a model trained on one database to be evaluated on the other. The eICU vital features were therefore aligned to the canonical MIMIC-IV feature order using explicit feature-name mappings, and intervention features were checked against the expected intervention ordering. The validation scripts also inspected schema consistency, patient-label alignment, numeric compatibility, and the position of key intervention variables used by the clinical tasks.

4.3.2.5 Training workflow preparation

The original METRE repository separated the baseline experiments from the main NN training workflow. LR and RF baselines were provided in a separate notebook, whereas the RNN, TCN,

and Transformer encoder models were implemented in the main training script. In this dissertation, this workflow was reorganised so that all evaluated models could be executed systematically across tasks, training databases, and evaluation settings.

The final METRE-based workflow included the original LR and RF baselines, two additional classical baseline models implemented in this dissertation, SVM and NB, and the three METRE-derived NN models, RNN, TCN, and Transformer encoder. This allowed the same task definitions, data splits, hyperparameter optimisation procedure, and evaluation protocol to be applied consistently across the full model set.

Further changes were made to support the evaluation design adopted in this dissertation, including stratified cross-validation, systematic saving of fold-level predictions, checkpoint retention, additional metrics, Optuna-based hyperparameter optimisation, and post-training aggregation of predictions. These changes are described in Section 4.6, since they concern the modelling and evaluation workflow rather than the extraction and conversion of the data itself.

4.4 Baseline Models

This section describes the models evaluated after the adapted METRE workflow was standardised. Since the theoretical basis of these algorithms was introduced in Chapter 2, the purpose here is to specify how each model was connected to the generated METRE representations and which configurable parameters were considered during model development. The complete hyperparameter search spaces, selected Optuna values, and default configurations are reported in Appendix C.

4.4.1 Classical Baseline Models

The classical baseline models included LR, RF, SVM, and NB. These models were used to assess how much predictive information could be captured from the structured METRE representation without relying on sequence-specific NN architectures. For these models, the task-specific longitudinal data were converted into fixed-size tabular inputs compatible with standard ML classifiers.

The LR configuration was controlled through the regularisation strength and elastic-net mixing parameter. For RF, the evaluated parameters included the number of trees, maximum tree depth, minimum number of samples required to split a node, minimum number of samples per leaf, number of features considered at each split, and bootstrap sample fraction. The SVM configuration included the regularisation parameter, kernel type, and kernel coefficient. For NB, only the smoothing parameter was considered. LR and RF were retained from the original METRE baseline setting, while SVM and NB were added in this dissertation to broaden the classical baseline comparison.

4.4.2 METRE-derived Neural-network Models

The METRE-derived NN models included the RNN, TCN, and Transformer encoder architectures. These models operated directly on the longitudinal representation generated by the adapted

pipeline, in which each ICU stay was represented as a feature-by-time matrix containing the 200 time-dependent variables produced during the HDF5-to-NPY conversion step. The temporal length of the input depended on the task-specific observation window.

The RNN model was evaluated as the recurrent sequence-modelling architecture. Its configurable parameters included the recurrent cell type, hidden dimension, number of recurrent layers, dropout rate, and input dropout. The TCN model was evaluated as the convolutional temporal architecture, with the number of convolutional channels, kernel size, and dropout rate considered during model configuration. The Transformer encoder was evaluated as the attention-based architecture, with configurable parameters including the embedding dimension, number of attention heads, feed-forward dimension multiplier, number of encoder layers, and dropout rate.

4.5 External Model: EMERGE

The second modelling component of this dissertation was based on EMERGE, an external multimodal EHR predictive modelling framework proposed by Zhu et al. [88]. EMERGE was selected as an external model because it includes a structured TS component compatible with the longitudinal outputs generated by METRE, while also representing a more recent retrieval-augmented and multimodal modelling approach. In this work, EMERGE was not used as a data extraction pipeline, but as an independent modelling architecture adapted to consume the structured outputs produced by the adapted METRE workflow. This adaptation supported the benchmarking objective of the dissertation, by testing whether the same harmonised data representation could be reused beyond the models originally associated with METRE.

This section first describes the original EMERGE framework and then details the adaptations required to apply it to the METRE-generated data, including conversion of NPY outputs into EMERGE-style HDF5 files, alignment of TS tensor dimensions, handling of unavailable textual and summary modalities, and preparation of a training workflow compatible with the evaluation protocol used in this dissertation.

4.5.1 Original Model

As introduced in the literature review, EMERGE is a multimodal EHR predictive modelling framework designed to combine structured TS data, clinical notes, and external medical knowledge through a retrieval-augmented generation strategy [88]. The overall architecture is illustrated in Figure 3.7. The framework processes TS EHR data with a temporal encoder, obtains text representations from clinical notes using language-model embeddings, retrieves relevant medical knowledge from PrimeKG, generates task-relevant patient health summaries using a LLM, and fuses the resulting representations through a multimodal fusion network. PrimeKG is a biomedical knowledge graph that integrates information about diseases, drugs, genes, biological processes, and their relationships, and is used by EMERGE as the external knowledge source for retrieval [154].

In the original formulation, each patient may be represented by multiple modalities. The structured TS component contains multivariate temporal EHR measurements, which is the component

most closely aligned with the METRE representation used in this dissertation. However, while METRE produces harmonised ICU TS tensors from structured database tables, EMERGE defines its own processed input format and combines structured measurements with additional textual and knowledge-derived representations. Clinical notes provide unstructured textual information describing the patient status. A further representation is produced through the retrieval-augmented generation process: clinically relevant entities are extracted from TS abnormalities and clinical notes, matched to PrimeKG, and used to generate a summary of the patient's health status. In the processed EMERGE data format, this generated summary is represented through the `Summary` field.

The TS branch of EMERGE uses a GRU encoder to obtain a temporal patient representation. The textual branches use language-model embeddings for clinical notes and the RAG-generated summaries, PrimeKG is used as the external knowledge source for retrieval, not as the embedding model itself. These modality-specific representations are then combined through a fusion module, including cross-attention-based fusion in the full EMERGE architecture. The model is trained for binary clinical prediction tasks, with the original paper evaluating in-hospital mortality and 30-day readmission on MIMIC-III and MIMIC-IV.

The EMERGE paper evaluates several modality combinations, including TS only, note only, summary only, TS plus notes, TS plus summaries, and multimodal combinations involving all available representations. This ablation structure is relevant for this dissertation because it separates the TS component from the full multimodal framework, allowing the structured-data-only variant to be considered independently from the note and summary components.

The original experimental protocol used fixed train, validation, and test partitions with 70%, 10%, and 20% split proportions, respectively. The reported metrics include AUROC, AUPRC, and $\min(+P, Se)$, where the latter summarises the minimum value between precision and sensitivity. The paper reports mean and standard deviation using bootstrap resampling on the test set rather than 10-fold cross-validation. Consequently, the original EMERGE protocol was preserved only as a secondary reference workflow in this dissertation, while the main comparative evaluation was adapted to match the cross-validation protocol used for the METRE models, as described in Section 4.6.

4.5.2 Reproduction and Adaptation

The public EMERGE repository provided the model implementation and a notebook-based execution workflow, but not the complete preprocessing pipeline or derived artefacts required to reproduce the original multimodal experiments. The adaptation was therefore restricted to the publicly available structured TS branch, while preserving compatibility with the data format expected by the original implementation.

The objective was not to reproduce the complete multimodal framework, but to evaluate whether an external architecture could use the structured representations produced by METRE. The selected `ehr_only` configuration corresponds to the TS-only ablation reported by the

EMERGE authors and represented the most faithful setting supported by the publicly available resources.

4.5.2.1 METRE-to-EMERGE Data Conversion

A dedicated conversion script was implemented to transform the compiled METRE NPY outputs into the HDF5 structure expected by EMERGE. The METRE training files contain split-specific lists of patient-level arrays and their corresponding static labels. In contrast, the EMERGE data loader expects three HDF5 files, named `data_train.h5`, `data_val.h5`, and `data_test.h5`, where each patient is stored as a separate group containing structured TS data, note embeddings, summary embeddings, labels, and a patient identifier.

For each split, the conversion script reads the corresponding METRE arrays and writes an EMERGE-compatible HDF5 file. The structured TS data are stored under the `X` field, the patient identifier under `PatientID`, and the binary target under the `Y` field. Placeholder fields for `Note` and `Summary` were also created to preserve the expected file structure. The interpretation of these fields was consistent with the clarification later provided by the EMERGE authors, who confirmed that `X` represents the structured TS tensor, `Note` the clinical-note embedding, `Summary` the RAG-generated summary embedding, `Y` the prediction labels, and `PatientID` the identifier used to align patient-level representations.

The converter was implemented to support the five METRE prediction settings: hospital mortality at 48 hours, ARF at 4 and 12 hours, and shock at 4 and 12 hours, all with the same 6-hour gap definition. In this dissertation, EMERGE TS-only was evaluated across all five METRE task definitions. However, only the mortality task is directly aligned with the original EMERGE benchmark, the remaining tasks were used to assess whether the external TS architecture could operate on METRE-generated representations under the same protocol as the other models.

4.5.2.2 Time-series Tensor Alignment

One important adaptation concerned the orientation and dimensionality of the TS tensors. The METRE NPY representation stores each ICU stay as a feature-by-time matrix with shape $(200, T)$, whereas EMERGE expects the structured TS input in time-by-feature format. Therefore, each METRE array was transposed before being written to the EMERGE HDF5 files.

The temporal length T depended on the prediction task, 48 hours for mortality and either 4 or 12 hours for ARF and shock. The EMERGE input dimensionality was also adjusted from the original 61 structured features to 200 features, matching the representation generated by the adapted METRE pipeline.

4.5.2.3 Handling Unavailable Note and Summary Modalities

The original EMERGE architecture is designed to operate with structured TS data, clinical-note embeddings, and embeddings of summaries generated through the retrieval-augmented pipeline.

However, the public repository did not provide the processed note embeddings, generated summaries, retrieved knowledge records, or the complete preprocessing code required to reproduce these modalities from the original databases. Since MIMIC-derived data are subject to credentialed access and data use restrictions, such artefacts cannot be freely redistributed outside the appropriate access framework, a reproducible alternative would require either the release of the preprocessing code or deposition of the derived artefacts through a controlled platform such as PhysioNet under the applicable data use agreement.[1] Moreover, generating new summaries from MIMIC data would require a separate local and privacy-compliant language-model pipeline, as patient data cannot be submitted to external hosted models.

To preserve compatibility with the EMERGE data loader while restricting the model to the structured-data modality, placeholder vectors were created for the unavailable text-based fields. Specifically, the `Note` and `Summary` fields were filled with zero vectors of dimension 768. These placeholders maintained the expected HDF5 structure but were not used by the model when the `ehr_only` modality was selected. In this configuration, the prediction is based exclusively on the structured TS encoder, corresponding to the TS-only variant reported in the EMERGE ablation experiments.

4.5.2.4 Training Workflow Adaptation

The original EMERGE implementation was mainly notebook-based and was not directly suitable for systematic execution with the outputs of the METRE pipeline. Therefore, the required data loading, model initialisation, training, checkpointing, and prediction-export steps were adapted into executable Python scripts. This allowed EMERGE to be run from the command line and on the High-Performance Computing (HPC) environment, while preserving the original TS model components used by the public implementation.

A fixed-split execution mode was first retained as a sanity-check setting, since it resembles the original EMERGE workflow more closely. However, the final results reported in this dissertation follow the common evaluation protocol described in Section 4.6, in order to make the EMERGE results comparable with the METRE models.

4.6 Performance Evaluation

The original METRE and EMERGE implementations used different evaluation strategies. METRE combined cross-validation with test-set evaluation and reported bootstrapped confidence intervals, but the final reported test result in the released workflow was associated with a selected fold model rather than with an aggregation of all cross-validation models. EMERGE, in contrast, used fixed train-validation-test partitions and reported performance over repeated bootstrap evaluations. Since this dissertation compares models derived from both frameworks, a unified

[1] `https://physionet.org/content/mimiciv/view-dua/3.1/` (accessed 7 June 2026).

evaluation protocol was implemented to ensure that all final results were obtained under the same experimental logic.

All experiments were submitted through a Slurm-managed HPC environment using GPU-enabled jobs. Each job requested one GPU, eight CPU cores, and 40 GB of memory. Separate Conda environments were used for the adapted METRE and EMERGE workflows, and the execution scripts were standardised to run each task-model configuration, save checkpoints, export predictions, and generate the corresponding tabular result files.

This section describes the experimental protocol used for hyperparameter optimisation, cross-validation, internal and external testing, bootstrap-based confidence interval estimation, metric calculation, and result artefacts.

4.6.1 Hyperparameter Optimisation

Two experimental configurations were considered whenever applicable: default configurations and optimised configurations. Default configurations corresponded to the reference hyperparameters available in the adapted workflows or to standard baseline settings when no task-specific default was provided. Optimised configurations were obtained using Optuna, an automatic hyperparameter optimisation framework [81].

Hyperparameter optimisation was performed separately for each task, training database, and model. The objective metric was validation AUROC. Each optimisation study used 50 trials. To reduce computational cost during tuning, each trial was trained on a stratified subset corresponding to 35% of the development data and evaluated on a validation fold. The selected hyperparameters were then used in the complete cross-validation and test evaluation procedure described below.

The same optimisation logic was applied to the classical baseline models, the METRE-derived temporal models, and the adapted EMERGE TS-only model. Optuna was used only for model selection during development; the final reported results were obtained from the full cross-validation, internal testing, external testing, and bootstrap aggregation procedure. The search ranges, selected Optuna values, and default values for each model are documented in Appendix C, allowing the experimental configurations to be reproduced.

4.6.2 Evaluation Procedure

After hyperparameter selection, all final models were evaluated using a unified protocol designed to support internal testing, external testing, and direct comparison between the baseline and external model workflows. The procedure combined stratified cross-validation, test-set evaluation, ensemble-based bootstrap estimation, and systematic export of predictions and result summaries.

Figure 4.1 provides a schematic representation of the data partitions and their use during model development, internal testing, and external cross-database evaluation.

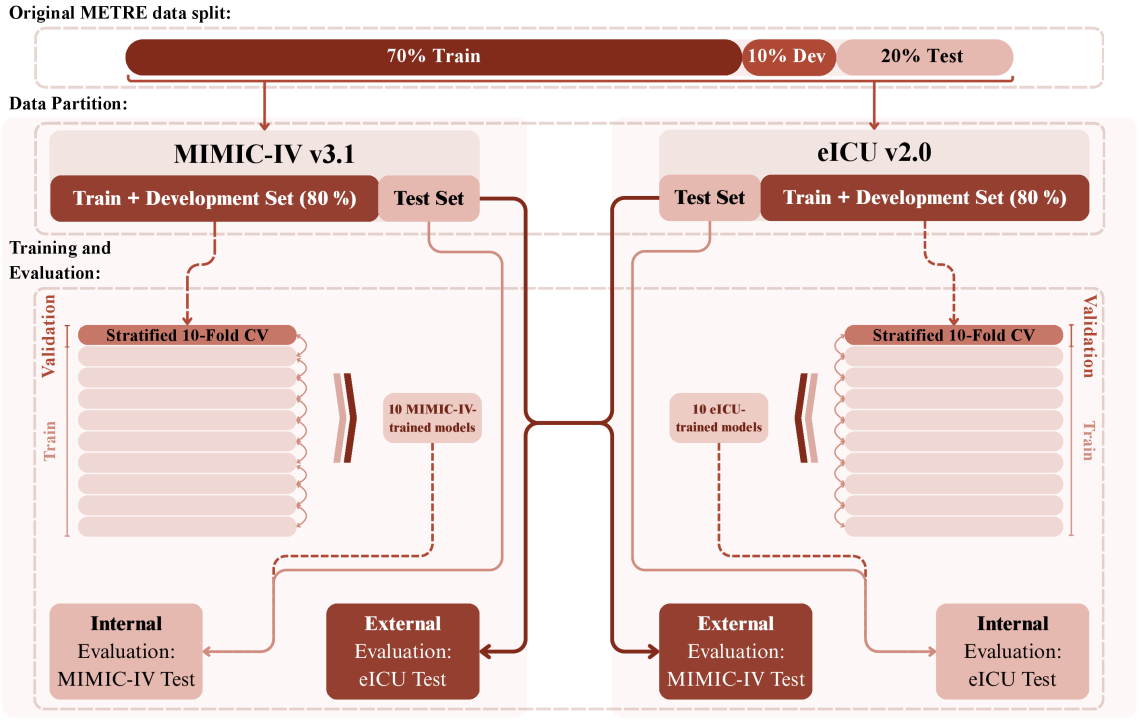


Figure 4.1: Schematic representation of the data partitioning and evaluation protocol. The original 70/10/20 train-development-test split was retained, with the train and development partitions combined for stratified 10-fold cross-validation. Models trained on each database were evaluated on the held-out internal test partition and on the held-out test partition of the other database.

4.6.2.1 Cross-validation

For each experiment, the original train and development partitions of the training database were combined into a single development cohort, defined here as the subset used for model training, validation, and hyperparameter selection before any final test-set evaluation. This cohort was then divided into ten folds using stratified cross-validation, preserving the proportion of positive and negative cases in each fold. Stratification was used because several tasks, particularly ARF and shock, are class-imbalanced and therefore sensitive to unequal fold composition. This strategy differs from the released METRE workflow, where cross-validation is performed during development but the final reported test results are obtained from a selected fold model rather than from the joint use of all trained folds.

For each fold, a new model was trained on nine folds and validated on the remaining fold. This produced ten independently trained models for each task-model configuration. The held-out validation fold was used only for model monitoring and internal validation within the development cohort. The dedicated test partition was never used during training, validation, early stopping, or hyperparameter selection.

Validation performance was summarised from the fold-level validation results as mean and standard deviation across the ten folds. This summary was used to characterise model behaviour within the development cohort, independently from the final test-set evaluation.

4.6.2.2 Internal and External Testing

After training each fold-specific model, evaluation was performed on two independent test sets. The internal test set corresponded to the held-out test partition of the same database used for training. The external test set corresponded to the held-out test partition of the other database not used in training. Thus, models trained on MIMIC-IV were tested internally on MIMIC-IV and externally on eICU, while the reverse setting was used for models trained on eICU.

Only the dedicated test partition of the external database was used for external evaluation. This differs from the original METRE cross-database evaluation, where the full external cohort was used. Restricting external evaluation to the test partition ensured that the same distinction between development and test data was maintained for both internal and external assessment.

For each fold-specific model, predicted probabilities and ground-truth labels were saved for the validation fold, the internal test set, and the external test set. These stored predictions allowed all final metrics, confidence intervals, and additional analyses to be computed without retraining the models.

4.6.2.3 Bootstrap Aggregation

The final test-set results were computed using a cross-validation ensemble bootstrap. For each test set, the ten models trained during cross-validation generated one predicted probability per patient. These ten probabilities were averaged to obtain a single ensemble probability for each patient.

Bootstrap resampling was then applied at patient level. In each bootstrap iteration, 1000 patients were sampled with replacement from the relevant test set. For every sampled patient, the previously stored ensemble probability was used, and all metrics were computed on the resulting bootstrap sample. This process was repeated 1000 times. The reported test performance corresponds to the mean and standard deviation of the bootstrap estimates, and the empirical 95% confidence interval was obtained from the 2.5th and 97.5th percentiles of the bootstrap distribution.

This approach differs from averaging metrics or confidence intervals across folds. Instead, the ten fold-specific models are treated as an ensemble, producing one final probability per patient before the metric is calculated. It also differs from the original METRE procedure, where the final bootstrapped result was associated with a selected fold model. In the present protocol, all ten cross-validation models contribute to the final test estimate.

4.6.2.4 Metrics

The evaluation protocol included AUROC, AUPRC, accuracy, precision, sensitivity, specificity, F1-score, and $\min(+P, Se)$. These metrics were introduced in Section 2.3.4.6, here, they are specified only in terms of their role in the experimental workflow. AUROC and AUPRC were computed from predicted probabilities and were therefore threshold-independent. The remaining metrics were computed after converting predicted probabilities into binary predictions using a fixed threshold of 0.5.

The $\min(+P, Se)$ metric was retained because it is reported in EMERGE and allowed the final metric set to remain consistent across the METRE-based models and the adapted EMERGE TS-only model. This metric was computed using the same thresholded predictions as precision and sensitivity.

4.6.2.5 Result Artefacts

The evaluation pipeline exported both intermediate and final result files. Fold-level Comma-Separated Values (CSV) files stored validation and test metrics for each of the ten cross-validation models. Prediction files stored patient-level ground-truth labels and predicted probabilities for validation, internal test, and external test sets. Summary CSV files stored mean and standard deviation across folds, including validation summaries. Final bootstrap CSV files stored the ensemble-based test results, including mean, standard deviation, 95% confidence intervals, number of patients, and number of folds.

Consequently, validation results reported in the dissertation were obtained from the fold-level validation summaries, whereas final internal and external test results were obtained from the ensemble bootstrap summaries.

4.7 Methodological Workflow Summary

Figure 4.2 summarises the complete methodological workflow implemented in this dissertation, from database access and METRE adaptation to model-ready data generation, model configuration, and final cross-database evaluation. Dashed borders identify externally developed pipeline or model components reproduced and adapted in this work, whereas the darker boxes highlight the main adaptations and new implementations introduced in this dissertation.

The workflow began with MIMIC-IV v3.1 and eICU v2.0, accessed through Google BigQuery. Applying METRE to MIMIC-IV v3.1 required updates to the released SQL queries and reconstruction of the auxiliary `culture` and `heparin` tables expected by the extraction functions. These database-specific inputs were then supplied to the adapted METRE pipeline, which performed cohort construction, extraction, preprocessing, temporal aggregation, harmonisation, and consistency checks across the `static`, `vital`, and `intervention` data streams.

The pipeline generated split-specific HDF5 outputs containing the `static`, `vital`, and `inv` tables for the train, development, and test partitions. Because the released training workflow expected compiled NPY inputs, a dedicated HDF5-to-NPY conversion was implemented. The resulting files were subsequently validated against their HDF5 sources to verify patient-label alignment, temporal structure, feature dimensions and ordering, and cross-database schema compatibility. This produced the common harmonised NPY representation used throughout the modelling experiments.

Two modelling branches were constructed from these validated outputs. The first used the NPY files directly to train the classical baselines, LR, RF, NB, and SVM, and the METRE-derived temporal models, RNN, TCN, and Transformer. In the second branch, the same NPY outputs

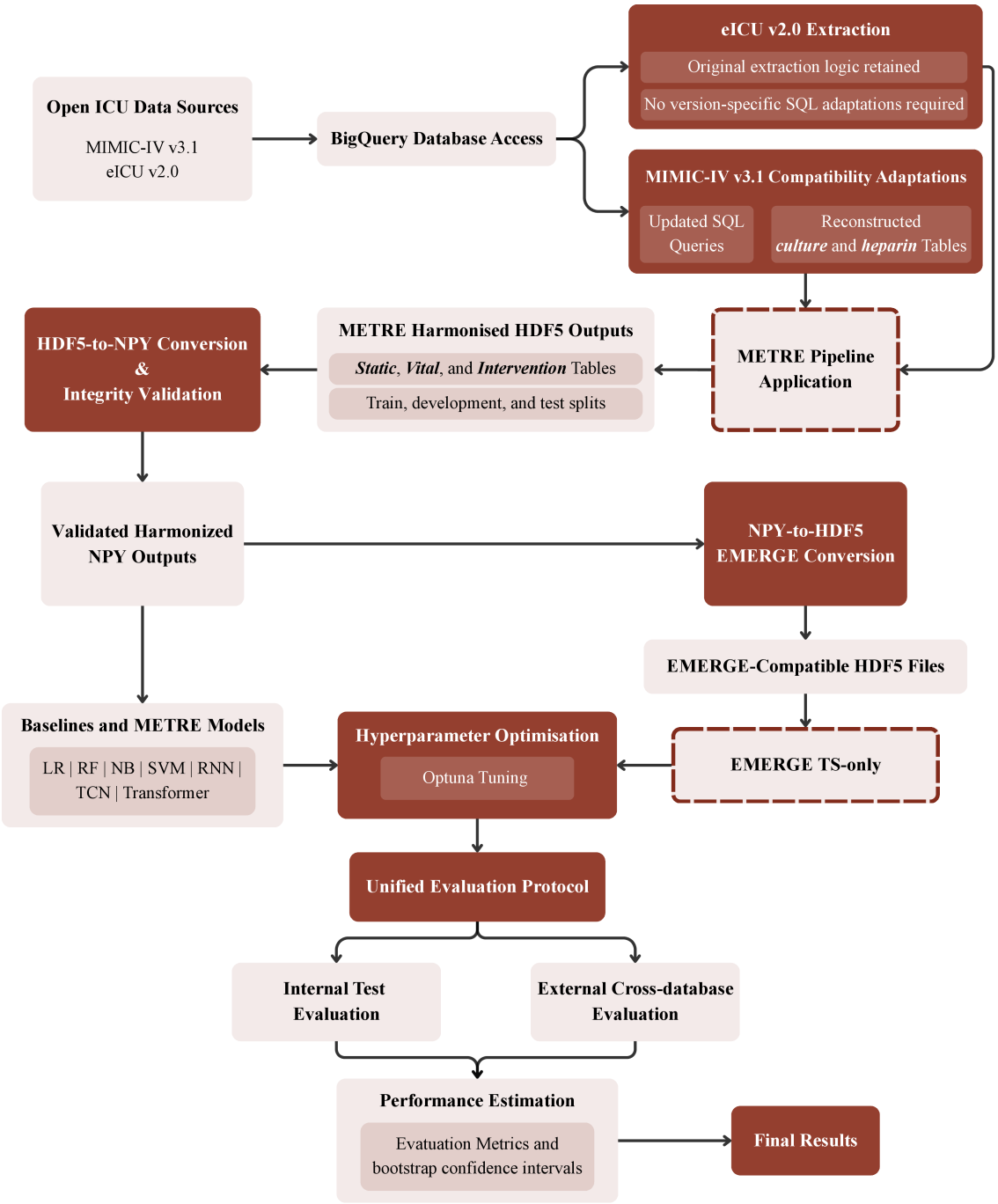


Figure 4.2: Overview of the methodological workflow implemented in this dissertation. MIMIC-IV and eICU data were processed through the adapted METRE pipeline, converted into validated model-ready representations, and used to evaluate the baseline, METRE-derived, and EMERGE TS-only models under a unified internal and external evaluation protocol. Dashed borders identify externally developed components reproduced and adapted in this work, while darker boxes highlight the main adaptations and new implementations introduced in this dissertation.

were transposed and converted into the HDF5 structure expected by EMERGE. This enabled the EMERGE TS-only architecture to be evaluated using the same five task definitions, patient partitions, and structured input variables, while excluding the unavailable note and generated-summary modalities.

Hyperparameter optimisation was performed separately for each model, task, and training database using 50 Optuna trials and validation AUROC as the objective. The selected configurations were subsequently evaluated through the unified protocol illustrated in Figure 4.1. The train and development partitions were combined and divided using stratified 10-fold cross-validation, producing ten independently trained models. Each model generated predictions for its validation fold, the held-out internal test set, and the held-out test set from the other database.

For final testing, the ten fold-specific probabilities were averaged for each patient before performance estimation. Internal evaluation measured performance within the training database, whereas external evaluation assessed transfer to the other database. Performance uncertainty was estimated using 1000 patient-level bootstrap resamples, from which the mean, standard deviation, and empirical 95% confidence intervals were obtained for all reported metrics. The resulting artefacts formed the benchmark used to compare model families, clinical tasks, and both directions of cross-database generalisation.

Chapter 5

Results and Discussion

This chapter presents and discusses the results obtained after adapting the data processing workflow, defining the task-specific cohorts, training the selected models, and evaluating them under the unified protocol described in Chapter 4. The results are organised progressively, beginning with the effect of the MIMIC-IV version update and the composition of the prediction cohorts, followed by internal validation and final test set evaluation. The evaluated models include classical baseline models, the METRE-derived NN architectures, and the external EMERGE TS-only adaptation. For the final evaluation, results are reported for both internal testing and cross-database testing, allowing model performance to be analysed within the source database and after transfer to an external database. The chapter concludes by summarising the main findings across all experiments, including the effect of hyperparameter optimisation, task-specific behaviour, model comparison, and the impact of cross-database evaluation.

5.1 Cohort and Task Composition

This section presents the cohort-level results obtained after adapting the data extraction workflow and applying the task-specific filtering criteria. These results provide the numerical context for the subsequent model evaluation, since the number of available ICU stays and the prevalence of positive cases directly affect training conditions, class imbalance, and the interpretation of performance metrics such as AUPRC.

Table 5.1 compares the number of ICU stays available in the original METRE extraction and in the extraction obtained after adapting the workflow to MIMIC-IV v3.1.

Table 5.1: Comparison between the MIMIC-IV cohort size reported in the original METRE pipeline and the cohort obtained after adapting the extraction workflow to MIMIC-IV v3.1.

Source	MIMIC-IV version	Extracted ICU stays
Original METRE pipeline	v2.2	38,766
Adapted extraction	v3.1	47,438

The adaptation to MIMIC-IV v3.1 increased the extracted MIMIC-IV cohort from 38,766 to 47,438 ICU stays, corresponding to an additional 9,672 stays. This increase was expected because newer major versions of MIMIC-IV include additional admissions and updated database content. From a modelling perspective, the larger extracted cohort is relevant because it increases the amount of data available before task-specific filtering, which may improve the stability of model training and cross-validation, particularly for tasks with fewer positive cases. Therefore, the update to MIMIC-IV v3.1 was not only a compatibility requirement, but also provided a larger data basis for the experiments conducted in this dissertation.

Table 5.2 presents the task-specific cohort composition after applying the observation windows and the 6-hour gap defined in the evaluation protocol. For each database and task, the table reports the number of negative and positive ICU stays in the training cohort and in the held-out test set. The training cohort corresponds to the original train and development partitions combined for cross-validation.

Table 5.2: Task-specific cohort composition after applying the observation windows and 6-hour gap. Train corresponds to the train and development partitions combined for cross-validation.

Database	Task	Window	Train (Negative cases)	Train (Positive cases)	Train Prevalence	Test (Negative cases)	Test (Positive cases)	Test Prevalence
MIMIC-IV	Mortality	48h	16,587	2,198	11.7%	4,214	520	11.0%
	ARF	4h	7,359	2,627	26.3%	1,839	654	26.2%
		12h	7,359	1,530	17.2%	1,839	366	16.6%
	Shock	4h	26,089	1,800	6.5%	6,573	403	5.8%
		12h	26,089	1,193	4.4%	6,573	257	3.8%
eICU	Mortality	48h	43,123	5,326	11.0%	10,701	1,318	11.0%
	ARF	4h	83,435	1,896	2.2%	21,983	363	1.6%
		12h	83,435	1,348	1.6%	21,983	251	1.1%
	Shock	4h	83,810	3,914	4.5%	21,802	880	3.9%
		12h	83,810	2,733	3.2%	21,802	608	2.7%

Within each task and database, the number of negative cases remains constant across the 4 and 12 hour settings because negative cases correspond to ICU stays with no eligible event recorded throughout the stay. Changing the observation window primarily alters which event-positive stays remain eligible after applying the 6-hour gap. The difference between the ARF and shock cohort sizes reflects their distinct task-specific filters and intervention variables. In MIMIC-IV, respiratory-support indicators exclude substantially more stays from the ARF cohort than vasoactive-treatment indicators exclude from the shock cohort. In eICU, the corresponding negative-case counts are closer, reflecting differences in the availability and recording of these intervention-derived variables.

The task-specific cohorts show substantial differences in class prevalence across tasks and databases. Mortality was the most stable outcome, with a positive prevalence of approximately 11% in both MIMIC-IV and eICU, across training and test partitions. In contrast, ARF and shock were considerably more imbalanced, particularly in eICU. Within eICU, ARF was especially sparse, positive cases represented only 2.2% of the training cohort in the 4-hour setting and 1.6% in the 12-hour setting, decreasing further to 1.6% and 1.1% in the corresponding test sets. This

low prevalence is important for later interpretation of AUPRC, since precision-recall metrics are strongly affected by the proportion of positive cases in the evaluated cohort. Therefore, lower AUPRC values in highly imbalanced tasks should be interpreted with awareness of the underlying class distribution rather than in isolation.

The comparison between MIMIC-IV and eICU also shows that the same task definition can produce substantially different positive-case distributions across databases. This is particularly relevant for cross-database evaluation, since differences in prevalence, clinical practice, intervention recording, and cohort composition may affect both discrimination metrics and threshold-dependent metrics. These cohort-level differences provide essential context for the internal validation and external evaluation results presented in the following sections.

5.2 Internal Validation

This section presents the internal validation results obtained during model development. Its purpose is to evaluate the behaviour of each model under cross-validation and compare the reference configurations reported in the corresponding studies or released implementations with the hyperparameters selected through Optuna, as detailed in Section 4.6.1, before proceeding to the final held-out test evaluation. The analysis is therefore restricted to development-stage performance and is not intended to replace the final internal and external test results presented in the following section.

Table 5.3 summarises the validation performance obtained across the 10 cross-validation folds. For each fold, the model was trained on the corresponding training set and evaluated on the held-out validation set. Results are shown for models trained on MIMIC-IV and eICU, using both default configurations and hyperparameters selected using Optuna. The main table reports AUROC, AUPRC, and $\min(+P, Se)$, while the remaining validation metrics, including accuracy, precision, sensitivity, specificity, and F1-score, are provided in Appendix B.1. The reported values correspond to the mean and standard deviation across folds.

Table 5.3: Internal validation performance for models trained on MIMIC-IV and eICU using default and selected Optuna hyperparameters. Values are reported as mean ± standard deviation. Trans denotes the Transformer model. Best results within each training database, task, and metric are shown in bold.

Train database	Config	Model	Mortality 48h			ARF 4h			ARF 12h			Shock 4h			Shock 12h		
			AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)
MIMIC-IV	Default	LR	87,02±0,85	52,68±3,63	**37,33±1,27**	73,57±1,98	53,32±2,84	44,25±1,68	73,26±1,89	42,33±3,37	32,15±1,43	76,59±2,12	21,75±3,13	14,65±0,65	78,50±1,59	19,18±3,87	12,20±0,68
		RF	86,89±1,15	49,80±3,40	32,85±3,98	**79,05±2,12**	**66,80±3,11**	44,73±4,49	**78,15±1,41**	**54,74±2,27**	27,12±2,84	**88,05±1,28**	**43,24±3,29**	36,39±3,41	**89,06±2,03**	41,06±3,34	31,60±3,99
		NB	69,32±2,51	19,46±1,31	20,78±1,43	69,71±2,25	48,88±3,56	34,14±2,99	68,48±2,01	32,48±2,07	**39,00±1,49**	69,18±1,44	10,74±0,53	9,61±0,28	50,86±1,84	4,46±0,16	4,39±0,15
		SVM	89,03±1,00	54,24±3,15	35,08±2,70	75,62±1,84	56,17±1,94	30,68±3,15	74,92±1,38	43,38±2,27	13,99±2,50	80,41±1,92	25,88±2,68	4,94±1,56	82,25±1,42	24,00±2,13	6,71±1,31
		RNN	88,02±0,94	54,22±4,28	36,07±3,44	76,40±1,96	61,62±2,74	50,69±2,79	74,80±1,33	49,49±2,68	34,73±3,53	82,06±1,93	29,81±3,90	24,93±4,74	82,18±1,51	25,15±3,30	21,23±5,96
		TCN	88,09±0,97	54,09±4,43	34,74±3,24	75,92±1,71	60,46±3,25	49,40±3,35	74,96±1,85	49,46±2,91	35,90±5,30	81,32±1,61	27,51±3,80	18,85±1,10	81,32±1,81	24,01±2,88	16,40±4,60
		Trans	87,36±0,86	50,84±3,11	28,82±4,96	76,25±2,47	60,25±3,44	47,28±4,23	75,62±1,82	49,92±3,07	36,24±5,64	81,54±1,94	27,30±4,27	14,88±2,05	83,12±1,52	21,89±3,37	11,21±1,16
	Optuna	LR	87,77±0,98	53,94±3,81	36,54±1,05	73,51±1,94	52,95±2,83	43,91±1,23	73,30±2,01	42,09±3,57	31,72±1,92	76,62±2,33	20,38±3,08	14,00±0,70	79,17±1,52	19,65±3,39	11,53±0,48
		RF	86,87±1,24	49,57±3,67	32,03±3,60	79,02±2,04	66,63±3,21	40,58±3,13	77,64±1,57	53,36±2,20	22,29±3,13	87,87±1,38	43,00±3,28	**40,56±3,74**	**89,06±2,05**	**41,24±3,41**	**32,78±3,69**
		NB	70,69±2,28	21,17±1,51	24,28±1,56	69,65±2,23	48,45±3,33	35,21±3,10	68,52±1,97	32,64±1,96	38,37±1,75	70,44±1,82	11,63±0,89	11,39±0,67	64,86±2,98	6,48±0,48	6,26±0,45
		SVM	89,08±1,05	54,04±3,21	34,35±2,82	75,45±2,04	56,47±2,38	31,44±2,65	74,83±1,86	43,66±3,09	15,62±2,93	81,54±2,26	24,86±2,86	2,72±1,15	83,42±1,54	22,23±2,97	3,10±1,68
		RNN	**89,35±0,72**	56,16±3,71	32,21±1,49	78,27±1,55	64,63±2,49	50,18±3,14	77,03±2,05	52,49±2,66	37,12±2,53	83,10±2,20	30,54±3,51	16,25±1,33	84,48±1,64	27,24±3,87	16,02±3,88
		TCN	89,00±0,93	55,67±3,44	36,67±2,18	77,24±1,85	62,76±2,69	**51,29±4,92**	76,74±1,97	52,05±2,93	38,24±4,64	82,62±1,85	29,82±3,81	20,85±4,33	82,31±1,63	23,84±1,49	14,52±2,39
		Trans	89,20±0,99	**56,83±3,85**	36,20±2,24	77,87±2,08	64,83±2,47	50,76±4,20	76,24±1,70	51,69±3,32	32,52±3,74	83,10±1,49	29,89±2,93	15,52±1,67	82,77±2,19	21,34±1,28	12,36±0,93
eICU	Default	LR	85,00±0,81	50,43±1,61	32,22±1,02	79,08±0,76	10,12±0,63	6,33±0,16	79,21±1,70	9,03±1,39	5,10±0,25	75,32±1,82	15,57±0,96	10,40±0,46	75,09±1,54	12,54±1,45	7,82±0,41
		RF	85,75±0,53	51,01±2,04	40,41±2,35	91,08±1,17	27,89±3,02	26,17±1,76	**90,60±0,88**	21,27±2,23	20,84±2,35	**82,40±0,97**	26,52±2,19	**22,92±2,04**	81,44±1,52	21,28±2,35	18,88±2,31
		NB	69,04±1,18	17,61±0,57	18,09±0,57	50,48±0,43	2,24±0,02	2,24±0,02	50,08±0,70	1,59±0,02	1,59±0,02	50,17±0,14	4,48±0,02	4,48±0,02	50,29±0,14	3,18±0,01	3,18±0,01
		SVM	86,67±0,56	51,10±1,68	**42,60±1,00**	84,71±1,11	21,05±3,18	23,22±1,85	84,27±2,12	18,12±2,26	**22,03±2,68**	77,58±1,07	17,38±1,47	17,64±0,84	77,50±1,04	14,96±1,62	**19,53±2,00**
		RNN	85,05±0,69	50,26±1,93	30,98±2,94	86,77±0,74	22,20±2,92	24,72±4,42	85,77±1,23	16,76±3,49	15,20±4,23	77,00±1,25	18,10±1,93	14,01±2,82	76,34±1,68	13,85±1,47	11,70±3,67
		TCN	85,27±0,81	50,91±2,23	29,96±3,92	86,45±0,96	19,75±2,52	14,18±2,61	84,52±1,44	16,02±1,55	15,06±2,70	76,60±1,16	17,48±1,26	12,03±1,83	75,81±1,56	13,18±1,56	12,38±3,21
		Trans	84,02±1,05	46,49±2,58	25,42±2,60	88,66±1,41	21,17±3,57	8,34±0,99	88,35±1,03	17,51±1,99	5,58±1,10	78,81±1,39	17,64±1,71	10,66±0,83	78,14±2,10	12,98±1,46	7,26±0,97
	Optuna	LR	85,61±0,83	51,41±1,55	32,21±0,97	79,54±0,73	9,95±0,67	6,23±0,14	80,07±1,57	8,40±1,10	4,94±0,25	75,34±1,83	15,32±1,01	10,32±0,39	75,53±1,75	12,24±1,46	7,72±0,39
		RF	86,00±0,71	51,83±2,04	36,73±2,13	**91,69±1,07**	**30,77±2,59**	**34,08±3,17**	89,69±0,88	18,13±1,78	6,73±0,40	82,31±0,83	**26,95±2,13**	18,32±2,06	**82,06±1,60**	**22,40±2,30**	16,83±1,69
		NB	72,23±1,49	21,09±1,12	23,72±1,15	66,90±2,50	3,45±0,24	2,60±0,08	56,90±2,19	1,86±0,09	1,75±0,08	64,80±1,71	6,44±0,31	4,90±0,22	58,84±1,04	3,84±0,10	3,47±0,10
		SVM	86,89±0,64	52,99±1,63	34,13±0,81	85,67±0,93	16,74±2,36	9,55±0,33	85,63±1,48	13,63±1,36	8,06±0,36	78,75±1,35	17,34±1,09	12,00±0,43	79,82±1,46	14,51±1,60	10,86±0,57
		RNN	86,22±0,67	52,56±2,47	27,61±1,24	90,12±0,60	27,41±3,68	13,02±2,78	87,53±1,34	20,38±3,31	10,42±3,10	79,37±0,94	19,91±1,96	9,74±0,72	77,99±1,65	15,95±1,82	9,11±3,84
		TCN	85,89±0,78	51,87±2,05	30,07±0,98	87,80±1,04	24,06±3,28	13,57±4,18	85,58±1,54	16,84±2,66	14,82±7,64	79,97±1,19	19,47±1,68	11,69±0,93	79,34±1,67	14,62±1,29	9,17±0,70
		Trans	**87,48±0,58**	**56,68±1,94**	35,14±1,59	89,64±1,04	26,52±2,15	9,82±0,97	90,08±0,98	**27,13±3,50**	14,31±2,90	80,34±1,12	20,23±1,52	11,85±0,60	80,42±1,68	16,59±1,55	10,36±1,32

Overall, the validation results show that hyperparameter optimisation had a measurable but not uniform effect across models, tasks, and databases. In several cases, particularly among the METRE-derived NN models, Optuna improved the main validation metrics. For models trained on MIMIC-IV, the Transformer improved in mortality prediction from 87.36% to 89.20% AUROC and from 50.84% to 56.83% AUPRC, while the TCN also improved in mortality across all three reported metrics. Similar gains were observed for ARF 4h, where the Transformer increased from 76.25% to 77.87% AUROC and from 60.25% to 64.83% AUPRC. For models trained on eICU, the effect of optimisation was also visible in several settings, such as the Transformer in ARF 12h, where AUROC increased from 88.35% to 90.08% and AUPRC from 17.51% to 27.13%. These examples indicate that the optimisation procedure was able to identify more suitable configurations than the reference settings reported in the corresponding studies.

However, the effect of Optuna was not uniformly positive. This may be explained by the fact that Optuna maximised validation AUROC over a finite search space using a stratified subset of the development data. Consequently, the selected configuration may reflect task-specific sampling variability, provide only marginal gains over an already suitable reference configuration, or favour AUROC at the expense of other metrics. Its performance may also vary slightly when subsequently evaluated through complete 10-fold cross-validation. Some classical baseline models showed only marginal changes or even small decreases after optimisation, particularly when their default configurations already produced stable validation performance. For example, RF models trained on MIMIC-IV remained very similar before and after optimisation in mortality and shock prediction, with only small variations in AUROC and AUPRC. In some tasks, improvements in one metric were accompanied by decreases in another, especially for $\min(+P, Se)$, which is sensitive to the balance between precision and sensitivity. This behaviour is expected in imbalanced clinical prediction tasks, where optimising for AUROC does not necessarily maximise threshold-dependent or precision-recall-based metrics. Therefore, the validation results should be interpreted as evidence of task- and model-specific tuning effects, rather than as a universal improvement across all metrics. Comparisons with the baseline performance reported in the original METRE study should nevertheless be interpreted cautiously. In particular, this dissertation used MIMIC-IV v3.1 rather than the originally supported v2.2 release, requiring adaptations to the extraction workflow. Differences in database versions, extracted cohorts, and available variables may therefore also contribute to deviations from the published baseline results.

The comparison between model families also provides useful information before the final test evaluation. The classical baselines were not merely trivial references, RF and SVM achieved competitive validation performance in several tasks, particularly in mortality and shock prediction. In MIMIC-IV validation, SVM obtained the highest default AUROC for mortality, while RF achieved strong validation results in both shock tasks. This suggests that, for some structured tabular representations, classical models can capture substantial predictive signal without requiring more complex temporal architectures. Conversely, NB showed weaker and less stable behaviour, especially in tasks derived from the Intervention table, where its AUPRC and $\min(+P, Se)$ values were frequently low. LR behaved as a more stable but generally less expressive baseline, providing

moderate performance across most tasks.

Across tasks, mortality prediction appeared to be the most stable validation setting, with consistently high AUROC values across most models and both training databases. This is consistent with the cohort composition reported in Table 5.2, where mortality showed similar positive prevalence in MIMIC-IV and eICU and was less sparse than ARF and shock. In contrast, ARF and shock showed more heterogeneous behaviour across databases and metrics. In MIMIC-IV, ARF produced relatively high AUPRC values for several models, particularly in the 4-hour setting, reflecting the higher positive-case proportion in MIMIC-IV compared with eICU.

In eICU, where ARF positive cases were much rarer, AUPRC and $\min(+P, Se)$ remained low despite reasonable AUROC values in some models. This pattern anticipates the need to interpret the final test results relative to task prevalence, rather than from AUROC alone.

The additional validation metrics reported in Appendix B.1 support this interpretation. In several sparse tasks, particularly ARF and shock in eICU, models could achieve high accuracy or specificity while maintaining low precision, sensitivity, or F1-score. This indicates that good overall classification performance was often influenced by the correct identification of negative cases, rather than by reliable detection of positive events. These results do not replace the held-out test evaluation, but they already show why positive-class metrics are essential for interpreting imbalanced ICU prediction tasks.

Taken together, the validation stage provided the basis for selecting the final configurations used in the held-out evaluation. Although Optuna did not improve every metric in every setting, it offered a systematic and reproducible tuning procedure and produced relevant improvements across several model and task combinations. For this reason, the final test set analyses focus on the selected Optuna configurations for all models and tasks, while retaining the default results as development-stage evidence of the effect of hyperparameter optimisation.

5.3 External Evaluation

This section presents the final held-out test evaluation of the optimised models. Unlike the internal validation results, which were used to characterise model behaviour during development, the results reported here correspond to the final evaluation stage and were obtained after hyperparameter selection.

For each training database, results are presented in a separate subsection. Models trained on MIMIC-IV are evaluated on the internal MIMIC-IV test set and externally on eICU, whereas models trained on eICU are evaluated on the internal eICU test set and externally on MIMIC-IV. This structure allows the results to be interpreted in two complementary ways, comparing model behaviour within the same data source, and by examining the extent to which performance is preserved under cross-database transfer.

Only the optimised Optuna configurations are reported in the main external evaluation tables. The main tables report AUROC, AUPRC, and $\min(+P, Se)$, while the remaining metrics, including accuracy, precision, sensitivity, specificity, and F1-score, are provided in Appendix B.2 and

Appendix B.3.

5.3.1 MIMIC-IV Training

Table 5.4 presents the final test performance of the models trained on MIMIC-IV. Results are reported for the internal MIMIC-IV test set and for the external eICU test set. Each cell contains the bootstrap mean and standard deviation, with the corresponding 95% confidence interval shown in parentheses below.

Table 5.4: Test performance for models trained on MIMIC-IV and evaluated on the internal MIMIC-IV test set and the external eICU test set. Values are reported as mean ± standard deviation, with the 95% confidence interval shown below. Trans denotes the Transformer model. Best results within each test database, task, and metric are shown in bold.

Test database	Model	Mortality 48h			ARF 4h			ARF 12h			Shock 4h			Shock 12h		
		AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)
MIMIC-IV	LR	87,41±1,72 (83,97, 90,50)	51,62±5,05 (42,05, 61,47)	35,85±3,16 (29,83, 42,26)	73,79±1,78 (70,31, 77,16)	53,41±3,24 (46,89, 59,74)	43,44±2,50 (38,54, 48,16)	72,52±2,16 (68,74, 76,64)	40,92±3,79 (33,25, 48,38)	29,91±2,41 (25,29, 34,63)	75,99±3,07 (68,91, 81,72)	20,28±4,36 (12,87, 29,12)	12,25±1,85 (8,78, 16,04)	77,88±4,07 (69,31, 85,18)	18,05±5,44 (9,06, 29,62)	9,79±1,93 (6,28, 13,78)
	RF	87,40±1,53 (84,16, 90,29)	48,24±5,04 (38,62, 58,13)	31,12±4,51 (22,33, 40,18)	**79,03±1,65** (75,77, 82,19)	**66,95±2,56** (61,71, 71,86)	41,99±3,05 (35,82, 47,92)	**76,98±2,09** (73,12, 81,19)	54,03±3,60 (46,97, 60,97)	23,78±3,24 (17,65, 30,07)	**87,04±2,48** (81,60, 91,57)	**42,10±6,59** (28,99, 54,60)	**39,81±5,96** (27,45, 51,35)	**88,24±3,10** (81,75, 93,69)	**42,88±8,33** (25,35, 57,82)	**37,53±7,68** (22,58, 52,28)
	NB	71,24±2,52 (66,23, 75,90)	20,05±2,32 (15,77, 25,09)	21,62±2,31 (17,10, 26,35)	70,09±1,94 (66,18, 74,06)	52,54±3,07 (46,39, 58,63)	36,53±2,98 (30,90, 42,38)	68,44±2,46 (63,50, 73,27)	34,46±3,30 (28,12, 41,25)	**38,80±3,49** (31,61, 45,24)	70,79±2,95 (64,98, 76,39)	10,74±1,78 (7,38, 14,46)	10,34±1,56 (7,20, 13,35)	66,37±3,58 (58,83, 72,71)	5,84±1,10 (3,87, 8,07)	5,48±0,94 (3,72, 7,41)
	SVM	89,02±1,41 (86,12, 91,76)	52,03±5,04 (42,14, 62,22)	30,84±4,44 (22,14, 39,25)	75,49±1,72 (71,92, 78,77)	56,06±3,20 (49,56, 62,11)	30,82±2,80 (25,11, 36,37)	73,13±2,16 (69,28, 77,31)	39,26±3,80 (31,68, 46,62)	13,08±2,65 (7,78, 18,50)	81,02±2,73 (75,15, 86,13)	26,28±5,30 (16,93, 37,00)	4,44±2,70 (0,00, 10,34)	81,92±3,58 (74,83, 88,50)	22,04±6,16 (11,28, 35,43)	4,08±3,40 (0,00, 11,63)
	RNN	89,37±1,39 (86,55, 92,13)	56,01±4,80 (46,72, 65,31)	30,91±2,69 (25,52, 36,05)	78,06±1,69 (74,63, 81,33)	66,33±2,60 (60,91, 71,30)	48,44±2,70 (42,99, 53,62)	76,16±2,20 (71,55, 80,31)	**54,52±3,55** (47,20, 61,34)	35,53±2,87 (29,73, 41,01)	82,75±2,72 (76,99, 87,63)	33,91±6,26 (21,78, 46,10)	14,90±2,04 (10,89, 18,75)	84,80±3,40 (77,70, 91,13)	32,14±7,56 (18,56, 47,85)	16,66±3,21 (10,61, 23,36)
	TCN	89,31±1,42 (86,58, 92,10)	55,34±4,73 (46,36, 64,22)	36,33±3,13 (30,30, 43,10)	78,34±1,65 (75,07, 81,51)	64,91±2,69 (59,29, 70,04)	**52,62±2,85** (46,69, 57,98)	76,50±2,13 (72,40, 80,73)	52,52±3,66 (44,83, 59,59)	35,67±2,95 (29,63, 41,50)	83,57±2,77 (77,51, 88,71)	34,58±6,14 (22,56, 46,45)	21,89±3,06 (16,06, 28,05)	84,39±3,49 (77,07, 90,89)	32,05±7,51 (18,83, 47,72)	15,73±2,91 (10,26, 21,92)
	Trans	89,74±1,38 (86,90, 92,48)	56,92±4,74 (47,41, 65,92)	36,33±3,08 (30,69, 42,68)	77,82±1,68 (74,56, 80,98)	65,86±2,61 (60,65, 70,87)	52,35±2,82 (46,57, 58,08)	75,85±2,20 (71,59, 80,10)	52,98±3,58 (45,47, 59,57)	30,99±2,46 (26,19, 35,82)	82,93±2,70 (77,26, 87,81)	32,69±6,02 (20,93, 44,15)	13,86±1,92 (10,24, 17,61)	82,88±3,42 (75,57, 89,10)	26,80±6,85 (14,81, 41,72)	10,88±1,95 (7,00, 14,90)
	EMERGE	**90,05±1,34** (87,44, 92,69)	**58,95±4,54** (50,20, 67,62)	**40,22±4,68** (31,35, 49,52)	77,69±1,66 (74,45, 80,66)	63,91±2,75 (58,29, 69,38)	40,18±3,00 (34,37, 45,83)	74,88±2,23 (70,45, 79,15)	50,68±3,69 (43,14, 57,66)	31,52±3,53 (24,65, 38,61)	82,99±2,76 (77,17, 88,03)	34,95±6,19 (22,97, 47,08)	15,39±4,67 (6,90, 24,59)	84,32±3,49 (76,94, 90,66)	33,39±7,79 (18,66, 49,02)	16,70±6,27 (5,56, 30,31)
	TS-only															
eICU	LR	75,35±2,60 (70,41, 80,27)	36,29±4,82 (27,02, 46,09)	24,66±2,74 (19,48, 29,97)	63,12±7,45 (47,94, 76,05)	3,98±2,20 (1,41, 10,13)	2,04±0,59 (0,95, 3,19)	65,08±8,25 (47,34, 80,62)	3,28±2,40 (0,94, 9,91)	1,60±0,56 (0,62, 2,71)	64,86±4,65 (55,66, 73,29)	8,89±2,63 (4,76, 14,83)	6,71±1,37 (3,96, 9,21)	64,90±5,57 (53,39, 75,70)	7,72±3,23 (3,17, 15,66)	5,53±1,69 (2,62, 8,95)
	RF	78,11±2,20 (73,49, 82,17)	31,44±4,35 (23,22, 40,42)	9,30±2,78 (4,30, 15,18)	**80,94±3,94** (72,63, 87,98)	6,66±3,24 (2,66, 15,21)	**4,16±1,14** (2,14, 6,43)	**73,64±6,39** (60,08, 86,11)	3,98±2,58 (1,22, 12,34)	**3,19±2,87** (0,00, 9,76)	**73,47±4,08** (65,05, 81,34)	**11,70±3,44** (6,23, 19,64)	1,11±1,72 (0,00, 5,56)	**72,75±5,02** (62,26, 82,32)	**9,70±3,74** (4,24, 18,51)	0,11±0,68 (0,00, 3,03)
	NB	49,02±2,15 (45,04, 53,36)	11,47±1,25 (9,24, 14,06)	11,67±3,07 (5,82, 17,82)	75,39±5,40 (64,00, 84,73)	3,89±1,14 (1,89, 6,19)	3,93±1,11 (1,90, 6,23)	48,74±10,38 (29,53, 68,80)	1,86±0,87 (0,70, 4,01)	2,21±1,22 (0,00, 4,80)	63,25±4,92 (53,33, 72,55)	6,49±1,39 (3,90, 9,29)	6,19±1,19 (3,84, 8,51)	54,08±5,30 (43,80, 64,44)	3,30±0,72 (2,05, 4,89)	3,40±0,75 (2,06, 5,02)
	SVM	**78,98±2,22** (74,56, 83,28)	34,22±4,56 (25,84, 43,69)	11,23±3,03 (5,98, 17,70)	66,64±6,66 (52,81, 78,83)	4,75±3,12 (1,56, 13,63)	2,94±1,08 (1,12, 5,20)	66,73±7,78 (51,33, 81,30)	3,66±2,82 (0,99, 11,61)	2,37±1,69 (0,00, 6,16)	68,36±4,44 (59,41, 76,86)	9,23±2,57 (4,83, 14,76)	0,35±0,97 (0,00, 3,03)	68,39±5,05 (57,61, 77,43)	7,07±2,58 (3,35, 12,89)	0,00±0,00 (0,00, 0,00)
	RNN	74,58±2,49 (69,83, 79,38)	30,34±4,20 (22,35, 38,45)	20,80±2,21 (16,62, 25,24)	71,86±6,21 (58,40, 83,32)	7,43±4,53 (1,86, 18,94)	2,39±0,64 (1,20, 3,67)	68,31±7,69 (52,29, 82,94)	4,96±4,35 (1,05, 16,67)	1,78±0,61 (0,68, 3,06)	66,61±4,66 (57,47, 75,27)	8,85±2,73 (4,60, 15,69)	6,88±1,44 (4,07, 9,63)	67,06±4,95 (57,75, 76,50)	6,03±2,11 (3,00, 11,13)	5,65±1,97 (2,19, 9,74)
	TCN	77,18±2,41 (72,55, 81,82)	36,52±4,69 (27,55, 45,43)	22,82±2,42 (18,10, 27,49)	68,43±6,39 (55,40, 80,80)	4,94±3,13 (1,58, 13,71)	2,42±0,67 (1,16, 3,77)	67,39±8,11 (49,99, 82,30)	4,17±3,65 (0,98, 14,70)	1,90±0,68 (0,74, 3,38)	67,36±4,47 (58,13, 75,65)	9,04±2,70 (4,73, 15,48)	**8,15±1,94** (4,24, 12,18)	66,68±5,03 (56,48, 75,81)	6,47±2,48 (2,96, 12,24)	5,69±1,79 (2,40, 9,39)
	Trans	78,08±2,35 (73,70, 82,65)	**37,68±4,72** (29,04, 47,15)	**28,27±3,15** (22,32, 34,54)	69,71±6,69 (56,13, 82,87)	**9,81±6,63** (1,83, 24,77)	2,34±0,64 (1,19, 3,62)	69,67±7,78 (54,14, 84,97)	5,19±4,42 (1,04, 16,98)	1,68±0,54 (0,73, 2,78)	69,28±4,39 (60,52, 77,38)	10,31±3,14 (5,44, 17,55)	7,44±1,53 (4,32, 10,37)	67,97±4,88 (58,58, 77,19)	7,58±3,00 (3,41, 15,02)	5,47±1,57 (2,54, 8,78)
	EMERGE	75,90±2,35 (71,42, 80,38)	29,16±4,06 (22,06, 37,43)	26,82±4,12 (19,15, 34,87)	68,78±6,83 (55,42, 81,61)	5,81±3,84 (1,76, 15,99)	3,07±1,00 (1,27, 5,15)	65,78±9,02 (45,52, 82,54)	**5,41±4,69** (0,96, 17,68)	2,66±1,31 (0,63, 5,70)	67,37±4,68 (57,57, 76,08)	9,30±2,75 (4,67, 15,65)	4,11±3,17 (0,00, 11,43)	66,15±4,89 (56,39, 75,78)	6,17±2,20 (3,09, 11,40)	**5,97±4,26** (0,00, 14,81)
	TS-only															

The internal MIMIC-IV test results show that the harmonised representation extracted from MIMIC-IV supported clinically meaningful prediction across the five tasks, but also that model performance depended strongly on the outcome being predicted. Mortality was the most stable task, with the METRE-derived NN models reaching AUROC values close to 90% and AUPRC values between 55.34% and 56.92%. The EMERGE TS-only adaptation obtained the highest mortality AUROC and AUPRC in the internal test set, at 90.05% and 58.95%, respectively. SVM also performed strongly in mortality, reaching 89.02% AUROC, but this did not translate into the best AUPRC or $\min(+P, Se)$, with values of 52.03% and 30.84%, respectively. These results indicate that early ICU TS data contain a relatively consistent signal for hospital mortality prediction and that this signal can be captured by different model families. The internal test results were also broadly consistent with the preceding cross-validation results. For example, the optimised RNN, TCN, and Transformer achieved validation AUROC values of 89.35%, 89.00%, and 89.20%, respectively, compared with 89.37%, 89.31%, and 89.74% on the held-out test set. Their AUPRC values were similarly close across both stages. This agreement suggests that the selected configurations generalised consistently from the development cohort to previously unseen MIMIC-IV cases.

The comparison between classical baselines and more complex models is particularly informative. RF was not only a reference baseline but one of the strongest models in several tasks, especially ARF and shock. In ARF 4h, RF obtained the highest AUPRC, 66.95%, while the RNN, TCN, and Transformer obtained similar values between 64.91% and 66.33%. In shock prediction, RF achieved the best AUROC and AUPRC in both settings, reaching 87.04% AUROC and 42.10% AUPRC for shock 4h, and 88.24% AUROC and 42.88% AUPRC for shock 12h. This suggests that, after METRE preprocessing, a substantial part of the predictive signal is already available in a structured representation that can be used by classical models. In other words, the added complexity of NN architectures did not systematically translate into better internal performance for all tasks. This does not make the temporal models unnecessary, but it shows that the strength of the data representation and task definition can be as important as the model architecture itself.

SVM added a different pattern among the baselines. It was highly competitive for mortality, particularly in AUROC, and later achieved the best external mortality AUROC on eICU. However, it was less consistent in ARF and shock, especially when considering positive-class metrics. For example, internally in shock 12h, SVM reached 81.92% AUROC but only 4.08% $\min(+P, Se)$. This indicates that SVM could rank patients reasonably well in some settings, but often struggled to provide a useful balance between precision and sensitivity for rarer positive events. The remaining baselines behaved as expected. LR provided stable but generally less competitive results, performing reasonably in mortality and ARF but showing limited positive-class performance in shock. NB was consistently weaker, particularly in the shock tasks, where both AUPRC and $\min(+P, Se)$ were low. In shock 4h, NB obtained 10.74% AUPRC and 10.34% $\min(+P, Se)$, while in shock 12h these values decreased to 5.84% and 5.48%, respectively. This behaviour reinforces its role as a low-complexity reference rather than a competitive final model.

Across tasks, ARF and shock showed different types of difficulty. ARF 4h presented relatively strong AUPRC values for several models, suggesting that respiratory support onset remained detectable from the MIMIC-IV representation. Although the ARF 12h task provided more observations, the 6-hour gap excluded earlier events, reducing the number and prevalence of positive cases. Moreover, relevant signs of deterioration may be concentrated in the latest hours, while earlier measurements may be redundant or noisy. Therefore, a longer observation window did not necessarily improve performance. Shock prediction presented a different pattern, RF achieved high AUROC and AUPRC, whereas several METRE-derived temporal models had more modest $\min(+P, Se)$. For shock 12h, RNN, TCN, and Transformer obtained values of 16.66%, 15.73%, and 10.88%, respectively, despite AUROC values above 82%. This indicates that these models preserved discriminative ranking without achieving the same balance between precision and sensitivity.

A direct comparison with the original METRE results is not methodologically fair, since the original study used a different evaluation protocol, different database versions, and a different model selection procedure. Nevertheless, the comparison is still informative as a contextual reference. Under the protocol used in this dissertation, mortality prediction reached AUROC values comparable to those reported for METRE-MIMIC-IV. In the original article, the METRE-MIMIC-IV mortality AUROC values ranged from 86.7% to 87.6% across LR, RF, CNN, and LSTM, while in this dissertation the METRE-derived temporal models and EMERGE TS-only were all close to 90% in the internal MIMIC-IV test set [103]. Shock prediction also remained competitive, for example, the original METRE-MIMIC-IV shock 4h AUROC values ranged from 76.3% to 88.0%, while the adapted workflow reached 87.04% with RF and 83.57% with TCN. These results suggest that the adapted extraction, conversion, and training workflow preserved a substantial part of the predictive information present in the original METRE representation.

The main exception was ARF, not because performance was uniformly poor, but because this task can be particularly sensitive to how respiratory support variables are extracted, temporally aligned, and encoded. Internally, ARF 4h obtained strong AUPRC values, such as 66.95% for RF and 66.33% for RNN, which are close to or above the corresponding METRE-MIMIC-IV values reported in the original article. However, the interpretation of ARF remains less straightforward than mortality or shock because its labels depend directly on respiratory support variables extracted into the Intervention table. Unlike the published article, the public METRE repository did not provide the complete HDF5-to-NPY conversion workflow required to reproduce the exact model-ready tensors used by the authors. Consequently, the conversion file had to be reconstructed in this dissertation, including assumptions about feature ordering, tensor organisation, and compatibility with the downstream training code. Any residual discrepancy in ARF performance may therefore reflect task-specific sensitivity to the extraction, timing, or encoding of respiratory support variables, rather than a global failure of the adapted pipeline.

EMERGE TS-only

The EMERGE TS-only adaptation provides a separate perspective, since it was not part of the original METRE modelling framework. Internally, it achieved the strongest mortality performance among the tested models, with 90.05% AUROC and 58.95% AUPRC. However, this result corresponds only to the structured TS branch that could be adapted using the available data and released implementation. The full EMERGE framework combines TS data with clinical notes, generated summaries, retrieval-augmented knowledge, and multimodal fusion. The artefacts and preprocessing components required to include these additional modalities were not available for the implementation developed in this dissertation.

The original EMERGE article reported 89.50% AUROC, 63.11% AUPRC, and 59.95% $\min(+P, Se)$ for the complete multimodal framework on MIMIC-IV mortality [88]. In the corresponding ablation study, the original TS-only configuration achieved 87.96%, 55.62%, and 55.02%, respectively. These values cannot be directly compared with those obtained in this dissertation because the evaluation protocol and data processing workflow differed. Nevertheless, they provide useful context. Compared with the original TS-only configuration, the adapted model showed higher AUROC and AUPRC, but a substantially lower $\min(+P, Se)$ of 40.22%. This indicates that the model discriminated mortality risk well, but did not reproduce the same precision-sensitivity balance reported for the full multimodal EMERGE framework. The difference is consistent with the absence of notes, RAG summaries, and multimodal fusion.

For ARF and shock, EMERGE TS-only should be interpreted only as an external architecture evaluated under the METRE task definitions, since these tasks were not part of the same EMERGE benchmark. In these settings, EMERGE TS-only was competitive but not consistently superior to RF or to the METRE-derived temporal models. Its main contribution in this dissertation was therefore methodological, it showed that a model originally developed outside METRE could be connected to METRE outputs and evaluated within the same protocol.

Cross-database Evaluation on eICU

The external eICU test set provides a stricter assessment of model generalisation, since all models were trained on MIMIC-IV and then evaluated on a different critical care database. As expected, performance decreased relative to the internal MIMIC-IV test set, but the degradation was task-dependent rather than uniform. Mortality remained the most transferable outcome, whereas ARF and shock showed clearer evidence of sensitivity to database shift. This pattern was observed not only for the METRE-derived NN models, but also for the adapted EMERGE TS-only model, indicating that the difficulty was not specific to a single architecture.

For mortality, several models retained moderate external performance. SVM obtained the highest external AUROC, 78.98%, followed closely by RF and the Transformer, with 78.11% and 78.08%, respectively. However, the Transformer achieved the strongest external AUPRC and $\min(+P, Se)$, with 37.68% and 28.27%, indicating a better balance between ranking performance and positive-class detection. Although AUPRC decreased relative to the internal test set, mortality

performance remained more stable than the intervention tasks, particularly when considering the joint behaviour of AUROC, AUPRC, and $\min(+P, Se)$.

ARF showed the strongest external degradation. In both the 4-hour and 12-hour settings, AUPRC and $\min(+P, Se)$ were low across almost all models, including EMERGE TS-only. Some models retained moderate AUROC values, but this did not translate into reliable positive-event detection. For example, the Transformer obtained the highest ARF 4h external AUPRC, 9.81%, while its $\min(+P, Se)$ remained only 2.34%. Given the much lower ARF prevalence in eICU, these results suggest that respiratory support prediction was particularly affected by differences in event frequency, recording practices, and the operational definition of mechanical ventilation across databases.

Shock prediction showed an intermediate behaviour under external evaluation. External AUROC values remained moderate for several models, especially RF and the METRE-derived temporal models, but AUPRC and $\min(+P, Se)$ decreased markedly. EMERGE TS-only followed the same general pattern, with external shock performance below its internal MIMIC-IV results. RF illustrates the main issue clearly, despite achieving the best external AUROC in both shock tasks, with 73.47% for shock 4h and 72.75% for shock 12h, its $\min(+P, Se)$ was only 1.11% and 0.11%, respectively. This indicates that the model retained some ability to rank patients externally, but failed to provide a useful balance between precision and sensitivity for positive shock cases. The additional test metrics in Appendix B.2 support this interpretation, showing that several external models preserved high specificity while sensitivity, precision, or F1-score remained limited.

Overall, the eICU external evaluation shows that the models preserved some transferable signal, especially for mortality, but struggled with tasks whose labels depend on recorded treatment or support initiation. The loss of performance was not restricted to a single architecture, since it affected both METRE-derived models and the adapted EMERGE TS-only model. This points to database-specific outcome prevalence, recording practices, and label construction as central factors in cross-database ICU prediction.

5.3.2 eICU Training

The eICU training results provide the complementary direction of the external evaluation, showing how models trained on eICU behave when evaluated internally on eICU and externally on MIMIC-IV. Since the main interpretation axes were already introduced for the MIMIC-IV-trained models, this subsection focuses on whether the same patterns remain visible when the training database is reversed. Table 5.5 reports the main test metrics, while the remaining metrics, including accuracy, precision, sensitivity, specificity, and F1-score, are provided in Appendix B.3.

Table 5.5: Test performance for models trained on eICU and evaluated on the internal eICU test set and the external MIMIC-IV test set. Values are reported as mean ± standard deviation, with the 95% confidence interval shown below. Trans denotes the Transformer model. Best results within each test database, task, and metric are shown in bold.

Test database	Model	Mortality 48h			ARF 4h			ARF 12h			Shock 4h			Shock 12h		
		AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)	AUROC (%)	AUPRC (%)	min(+P,Se) (%)
eICU	LR	85,36±1,82	50,30±5,06	31,79±2,96	80,81±5,61	9,40±4,15	5,11±1,53	80,47±7,34	7,37±4,12	3,76±1,33	74,22±4,22	15,12±4,72	8,52±1,63	74,53±5,02	12,66±4,78	6,37±1,44
		(81,77, 88,84)	(40,71, 60,14)	(25,62, 37,45)	(69,27, 90,53)	(3,39, 19,66)	(2,34, 8,38)	(64,44, 92,84)	(2,04, 17,68)	(1,49, 6,70)	(65,98, 81,95)	(7,59, 25,56)	(5,35, 11,71)	(64,17, 83,93)	(5,20, 23,24)	(3,66, 9,20)
	RF	86,00±1,84	51,70±5,07	**36,24±4,83**	**93,59±2,79**	34,79±11,10	**30,07±9,25**	**91,37±3,95**	25,54±12,05	4,99±1,54	**81,73±3,72**	**26,52±7,05**	**19,95±6,63**	**82,44±4,36**	22,88±7,53	**16,74±7,24**
		(82,00, 89,41)	(41,16, 61,65)	(26,92, 46,50)	(86,88, 97,61)	(14,34, 57,39)	(11,76, 47,62)	(82,54, 97,55)	(5,64, 50,51)	(2,27, 8,08)	(74,23, 88,66)	(13,19, 40,73)	(7,14, 33,33)	(73,08, 90,41)	(10,35, 39,56)	(3,70, 33,33)
	NB	71,49±2,43	20,87±2,49	22,87±2,65	70,24±4,60	2,87±0,76	1,94±0,49	57,63±6,15	1,38±0,43	1,21±0,36	65,28±3,46	5,87±1,01	4,32±0,69	60,51±3,42	3,50±0,70	2,94±0,57
		(66,86, 76,24)	(15,98, 25,61)	(17,74, 28,03)	(60,24, 77,47)	(1,47, 4,44)	(1,00, 2,93)	(43,95, 66,43)	(0,60, 2,27)	(0,58, 1,96)	(58,25, 71,68)	(3,86, 7,84)	(2,92, 5,61)	(53,00, 66,12)	(2,24, 5,01)	(1,92, 4,11)
	SVM	86,49±1,75	51,18±5,21	33,16±3,11	86,78±4,12	15,72±7,17	7,26±2,17	85,70±5,29	14,72±8,43	6,33±2,34	77,82±3,71	17,54±5,13	10,39±1,91	78,65±4,66	15,23±5,53	9,24±2,22
		(83,06, 89,79)	(40,69, 61,97)	(26,86, 39,01)	(77,96, 93,79)	(5,06, 32,87)	(3,31, 11,88)	(74,63, 95,05)	(3,39, 35,50)	(2,25, 11,32)	(70,09, 84,81)	(9,07, 28,90)	(6,61, 14,04)	(69,51, 86,91)	(6,32, 27,61)	(5,20, 13,58)
	RNN	86,14±1,83	52,00±5,20	27,44±2,60	92,54±3,28	34,25±11,42	13,03±3,54	89,52±5,50	25,19±11,13	12,15±4,33	79,20±3,87	21,58±6,09	8,83±1,51	79,89±4,66	19,47±6,94	9,17±2,16
		(82,50, 89,73)	(41,26, 62,09)	(22,15, 32,35)	(84,68, 97,65)	(13,40, 57,77)	(6,32, 20,00)	(76,16, 97,70)	(6,73, 49,37)	(4,44, 21,43)	(70,91, 86,39)	(10,53, 34,40)	(5,92, 11,97)	(70,20, 87,94)	(8,16, 34,63)	(5,08, 13,54)
	TCN	85,60±1,84	50,50±5,04	29,49±2,80	90,33±3,82	28,11±10,44	13,44±3,99	86,73±6,01	25,64±12,08	**17,83±7,68**	79,92±3,82	20,44±5,67	11,10±1,99	79,29±4,62	15,11±5,42	8,36±1,96
		(82,16, 89,25)	(40,44, 60,60)	(24,03, 35,10)	(81,32, 96,49)	(9,22, 48,45)	(5,97, 21,05)	(72,70, 96,42)	(5,55, 51,64)	(4,35, 34,78)	(72,14, 86,88)	(10,58, 32,63)	(7,11, 15,04)	(69,86, 87,19)	(6,50, 27,96)	(4,66, 12,17)
	Trans	**87,38±1,79**	**55,65±5,10**	34,59±3,28	92,10±3,20	**35,49±11,58**	8,49±2,29	91,19±4,49	**32,75±13,25**	13,53±4,92	80,12±3,68	20,77±5,77	10,54±1,88	80,15±4,49	18,21±6,47	9,05±2,15
		(83,93, 90,76)	(45,76, 65,52)	(27,98, 40,83)	(84,98, 97,23)	(13,50, 57,36)	(4,12, 13,05)	(80,82, 98,07)	(7,83, 59,85)	(5,00, 23,91)	(72,24, 86,57)	(10,47, 32,66)	(6,93, 14,03)	(70,81, 88,47)	(7,14, 33,83)	(5,29, 13,51)
	EMERGE TS-only	86,90±1,79	54,95±5,12	32,46±4,55	89,08±4,34	28,66±10,74	14,17±9,19	88,24±5,76	25,07±11,48	14,33±10,88	79,78±3,87	22,95±6,35	6,06±3,88	79,78±3,87	**22,95±6,35**	6,06±3,88
		(83,49, 90,37)	(44,77, 65,01)	(23,71, 41,77)	(79,46, 96,59)	(10,46, 51,12)	(0,00, 33,33)	(74,49, 97,40)	(5,43, 50,74)	(0,00, 37,50)	(71,39, 86,67)	(11,91, 36,84)	(0,00, 15,00)	(71,39, 86,67)	(11,91, 36,84)	(0,00, 15,00)
MIMIC-IV	LR	77,88±2,57	38,77±4,82	36,13±4,06	58,45±2,14	35,18±2,84	32,85±2,47	58,35±2,51	24,52±2,80	21,99±2,12	62,63±4,02	10,99±2,47	9,77±1,75	65,39±4,83	7,98±2,29	6,26±1,31
		(72,53, 82,75)	(29,12, 47,91)	(27,92, 43,98)	(53,81, 62,66)	(29,93, 40,67)	(28,15, 37,68)	(53,40, 63,35)	(19,32, 30,80)	(18,15, 26,08)	(55,04, 70,57)	(6,94, 16,74)	(6,29, 13,41)	(56,18, 74,19)	(4,20, 13,26)	(3,79, 8,92)
	RF	81,19±2,12	38,82±4,63	9,04±2,77	**67,30±2,06**	48,65±3,21	13,28±2,08	**67,15±2,34**	36,69±3,67	30,68±2,77	**77,67±2,93**	**20,42±4,52**	3,59±2,43	**78,97±3,45**	15,39±4,41	1,27±1,93
		(76,88, 84,85)	(29,36, 47,41)	(3,92, 14,55)	(63,13, 71,43)	(42,39, 54,80)	(9,56, 17,60)	(62,70, 71,57)	(29,67, 43,51)	(25,00, 36,19)	(71,68, 83,24)	(12,22, 30,53)	(0,00, 9,09)	(72,33, 85,25)	(8,32, 25,23)	(0,00, 6,46)
	NB	39,44±2,36	9,56±0,92	6,37±1,19	46,08±2,10	25,70±1,62	22,83±1,61	45,37±2,45	16,37±1,44	13,99±1,35	36,96±3,52	5,07±0,66	3,47±0,71	35,19±3,92	3,24±0,53	1,98±0,57
		(34,92, 44,23)	(7,80, 11,41)	(4,02, 8,76)	(42,16, 50,03)	(22,49, 28,87)	(19,74, 25,92)	(40,58, 50,01)	(13,50, 19,14)	(11,40, 16,62)	(29,97, 43,98)	(3,83, 6,36)	(2,15, 4,96)	(27,72, 43,17)	(2,30, 4,33)	(0,99, 3,21)
	SVM	81,26±2,24	41,85±4,96	35,36±3,75	59,43±2,08	36,43±2,80	32,17±2,87	60,68±2,49	25,54±2,84	21,00±2,99	67,50±3,69	12,40±2,71	**12,33±2,24**	68,96±4,55	9,83±3,05	**8,20±2,10**
		(76,63, 85,29)	(31,87, 51,42)	(28,16, 42,65)	(55,30, 63,44)	(30,65, 41,72)	(26,61, 37,83)	(55,77, 65,40)	(20,15, 31,37)	(14,96, 26,47)	(60,49, 74,25)	(7,92, 18,78)	(8,20, 16,82)	(59,01, 77,66)	(4,92, 16,89)	(4,23, 12,43)
	RNN	83,60±1,90	46,22±4,74	25,68±2,49	62,51±2,14	42,60±3,18	33,48±2,95	63,27±2,49	32,56±3,46	30,38±3,28	69,91±3,66	15,80±3,78	10,31±1,64	70,82±4,48	13,89±4,55	7,75±1,62
		(79,65, 87,18)	(37,15, 55,29)	(20,75, 30,70)	(58,33, 66,66)	(36,67, 48,89)	(27,71, 39,48)	(58,19, 68,10)	(25,77, 39,00)	(23,98, 36,61)	(62,41, 76,96)	(9,29, 24,20)	(7,24, 13,55)	(62,06, 79,01)	(6,30, 24,37)	(4,79, 10,93)
	TCN	82,37±2,08	46,32±4,82	28,01±2,74	62,71±2,09	41,02±3,00	23,50±2,62	61,04±2,51	32,43±3,45	16,68±2,92	65,16±3,89	13,32±3,32	11,66±2,38	65,77±4,74	9,36±3,11	7,48±2,10
		(78,02, 86,25)	(37,00, 55,37)	(22,78, 33,33)	(58,49, 66,67)	(35,22, 47,12)	(18,58, 28,68)	(55,82, 65,88)	(25,69, 38,93)	(11,19, 22,52)	(57,60, 72,70)	(8,01, 21,27)	(7,29, 16,41)	(56,44, 74,73)	(4,59, 16,30)	(3,68, 11,85)
	Trans	**83,74±2,01**	**48,20±4,94**	**37,35±3,74**	64,88±2,17	**51,61±2,93**	**42,82±2,82**	64,65±2,63	**41,84±3,57**	**32,01±3,50**	69,53±3,45	12,93±2,90	11,94±2,21	69,28±4,49	10,28±3,21	7,77±1,68
		(79,47, 87,39)	(38,40, 57,72)	(29,87, 44,85)	(60,68, 69,19)	(45,97, 57,15)	(37,30, 48,39)	(59,40, 69,83)	(34,77, 48,50)	(25,29, 39,01)	(62,45, 76,14)	(8,19, 19,38)	(7,66, 16,53)	(60,56, 77,92)	(5,39, 18,04)	(4,93, 11,25)
	EMERGE TS-only	80,31±2,12	41,33±4,80	28,23±4,38	59,28±2,16	37,26±2,95	4,27±1,24	60,31±2,49	24,88±2,67	2,68±1,21	68,40±3,81	15,50±3,86	1,28±1,51	68,40±3,81	**15,50±3,86**	1,28±1,51
		(76,08, 84,30)	(32,08, 51,27)	(19,49, 37,26)	(55,15, 63,73)	(31,47, 43,24)	(2,07, 6,83)	(55,36, 65,02)	(19,85, 30,44)	(0,59, 5,26)	(60,90, 75,77)	(8,98, 23,88)	(0,00, 5,17)	(60,90, 75,77)	(8,98, 23,88)	(0,00, 5,17)

The internal eICU results confirm that mortality was the most stable prediction task. Most models achieved AUROC values between 85.36% and 87.38%, with the Transformer, EMERGE TS-only, SVM, RNN, TCN, RF, and LR showing broadly comparable discrimination. These values were slightly lower than those obtained internally on MIMIC-IV, where the strongest models approached or exceeded 89% AUROC and EMERGE TS-only reached 90.05%. Nevertheless, mortality performance remained comparatively stable across both training databases, consistent with their similar outcome prevalence. RF remained one of the strongest eICU baselines, particularly for $\min(+P, Se)$, where it reached 36.24%, compared with 31.12% in the internal MIMIC-IV test. This further shows that simpler models can remain competitive when the structured feature representation is informative.

For ARF and shock, the internal eICU results were more mixed and differed substantially from those obtained after MIMIC-IV training. eICU produced higher maximum AUROC values for ARF, with RF reaching 93.59% for ARF 4h and 91.37% for ARF 12h. However, positive-class performance remained weaker. For example, in ARF 12h, RF achieved 25.54% AUPRC and only 4.99% $\min(+P, Se)$. Conversely, internal shock prediction was stronger after MIMIC-IV training, with best AUPRC values of 42.10% and 42.88%, compared with 26.52% and 22.95% after eICU training.

This pattern is consistent with the cohort composition reported earlier, since ARF and shock had substantially lower positive prevalence in eICU, particularly in the 12-hour tasks. Therefore, the high AUROC values indicate an ability to rank patients but not necessarily balanced positive-event detection. The supplementary metrics in Appendix B.3 reinforce this interpretation: several models achieved high accuracy and specificity while maintaining limited sensitivity or F1-score, indicating that performance was partly driven by correct classification of the majority negative class.

EMERGE TS-only

The EMERGE TS-only adaptation was again evaluated as an external architecture applied to the METRE-derived structured representation. Internally, it achieved competitive results for mortality and ARF, but it did not consistently outperform RF or the METRE-derived NN models. In mortality, its AUROC was close to the best-performing models, reaching 86.90%, but its $\min(+P, Se)$ remained below RF and the Transformer, with values of 32.46%, 36.24%, and 34.59%, respectively. This suggests that, although the architecture captured useful risk information, it did not provide a clear advantage in balancing precision and sensitivity.

For ARF and shock, EMERGE TS-only followed the same general limitations observed in the other models. Its internal AUROC values were reasonable, particularly for ARF 4h and ARF 12h, where it reached 89.08% and 88.24%, respectively. However, $\min(+P, Se)$ remained low in the more imbalanced tasks, especially in shock, with values of 6.06% for both shock prediction windows. The Appendix B.3 metrics further clarify this behaviour, EMERGE TS-only often achieved very high specificity but relatively low sensitivity, especially in shock. This means that the model was conservative in identifying positive cases, which is problematic for tasks where the

clinically relevant objective is the detection of future deterioration or treatment initiation. These results support the same conclusion drawn from the MIMIC-IV training setting: in this dissertation, EMERGE TS-only demonstrates the feasibility of connecting an external TS model to the METRE data representation, but it should not be interpreted as reproducing the full multimodal EMERGE framework.

Cross-database Evaluation on MIMIC-IV

When models trained on eICU were evaluated on the held-out MIMIC-IV test set, performance decreased in most tasks relative to their internal eICU evaluation, confirming that cross-database testing was more demanding than internal testing. However, the degradation was again not uniform across outcomes. Mortality remained the most transferable task, with several models preserving substantial AUROC values, ranging from 77.88% for LR to 83.74% for the Transformer, and meaningful AUPRC values, ranging from 38.77% for LR to 48.20% for the Transformer, on the external MIMIC-IV test set. This transfer direction also preserved more performance than training on MIMIC-IV and testing on eICU, where the best mortality AUROC and AUPRC were 78.98% and 37.68%, respectively. This is consistent with the previous direction of transfer and suggests that mortality prediction relies on broader physiological patterns that are more stable across databases than tasks derived from the Intervention table.

The external ARF results were more nuanced. Compared with the MIMIC-IV-to-eICU transfer direction, ARF performance was less severely degraded when models were trained on eICU and tested on MIMIC-IV, particularly in AUPRC for ARF 4h, where RF and the Transformer reached 48.65% and 51.61%, respectively. However, this should not be interpreted as evidence of intrinsically better transfer. The MIMIC-IV ARF test cohorts had substantially higher positive prevalence than the corresponding eICU cohorts, which naturally affects AUPRC and makes positive-event detection less sparse. The low $\min(+P, Se)$ values for several models still show that balanced precision and sensitivity remained difficult to achieve, especially in the 12-hour setting.

Shock prediction again showed limited transferability. Nevertheless, transfer from eICU to MIMIC-IV was less severe than in the opposite direction. The best external AUROC values reached 77.67% and 78.97% for shock 4h and 12h, compared with 73.47% and 72.75% when models were trained on MIMIC-IV and tested on eICU. Some models retained moderate AUROC values, such as RF with 77.67% for shock 4h and 78.97% for shock 12h, but AUPRC and $\min(+P, Se)$ were generally low. For example, in shock 12h, the best AUPRC was 15.50% for EMERGE TS-only, while SVM achieved the highest $\min(+P, Se)$, at 8.20%. SVM also obtained the highest $\min(+P, Se)$ for shock 4h, at 12.33%. Nevertheless, these values remained low in absolute terms, indicating that external ranking ability did not translate into reliable positive-case detection. The supplementary metrics in Appendix B.3 support this interpretation, several models showed high specificity but weak sensitivity or F1-score, suggesting that they were better at identifying non-events than detecting future shock onset. This pattern mirrors the previous training direction and reinforces that shock prediction is highly sensitive to database-specific differences in treatment recording, cohort composition, and temporal event representation.

Overall, the eICU-trained experiments corroborate the main conclusions from the MIMIC-IV-trained evaluation while adding an important nuance. Mortality was consistently the most robust task, RF remained a strong baseline, and the METRE-derived NN models were competitive but not uniformly superior. In contrast, ARF and shock remained the most fragile tasks under external evaluation, particularly when judged by positive-class metrics rather than AUROC alone. This confirms that cross-database validation is essential not only for estimating performance loss, but also for revealing whether apparently strong models are genuinely detecting positive clinical events or mainly learning database-specific patterns.

5.4 Overall Experimental Findings

This section summarises the main experimental patterns observed across cohort composition, internal validation, and external evaluation. Rather than repeating each individual table, it provides an integrated view of the results, focusing on the effect of hyperparameter optimisation, the relative behaviour of the evaluated models, and the impact of cross-database evaluation.

The internal validation stage was essential for understanding model behaviour before the final test evaluation. In particular, it showed that the selected Optuna hyperparameters generally improved performance over the default configurations, as reported in Table 5.3. These improvements were not equally large across all models and tasks, and some metrics remained sensitive to the selected operating point. Nevertheless, the validation results supported the use of the optimised configurations in the final internal and external test evaluation.

Table 5.6 summarises the best-performing model for each task, training database, test database, and metric. The best model was selected independently for AUROC, AUPRC, and $\min(+P, Se)$, since these metrics capture different aspects of performance. Therefore, the table should not be interpreted as identifying a single best model overall, but rather as summarising which models performed best under each metric and evaluation setting.

Across the full set of experiments, mortality was the most stable prediction task. Its best AUROC values remained high in internal testing and decreased less sharply under external evaluation than those observed for ARF and shock. By contrast, ARF and shock were more fragile, particularly when evaluated through AUPRC and $\min(+P, Se)$. This pattern is consistent with the cohort composition and with the dependence of these tasks on support and treatment variables derived from the Intervention table.

The model comparison showed that higher architectural complexity did not guarantee better performance. RF was consistently one of the strongest models, particularly for ARF and shock, and often matched or exceeded the METRE-derived NN models. SVM also became relevant after optimisation, it achieved the best external AUROC for mortality when trained on MIMIC-IV and the highest external $\min(+P, Se)$ for both shock windows when trained on eICU. However, these advantages were metric-specific, and the shock results remained low in absolute terms. These findings reinforce the importance of analysing AUROC together with AUPRC and $\min(+P, Se)$, rather than selecting models based on discrimination alone.

Table 5.6: Best-performing model for each task and evaluation setting. Values are reported in percentage. The best model was selected independently for each metric.

Train database	Test database	Task	Best AUROC (%)	Best AUPRC (%)	Best $\min(+P, Se)$ (%)
MIMIC-IV	MIMIC-IV	Mortality 48h	EMERGE TS-only (90.05)	EMERGE TS-only (58.95)	EMERGE TS-only (40.22)
		ARF 4h	RF (79.03)	RF (66.95)	TCN (52.62)
		ARF 12h	RF (76.98)	RNN (54.52)	NB (38.80)
		Shock 4h	RF (87.04)	RF (42.10)	RF (39.81)
		Shock 12h	RF (88.24)	RF (42.88)	RF (37.53)
	eICU	Mortality 48h	SVM (78.98)	Trans (37.68)	EMERGE TS-only (32.89)
		ARF 4h	RF (80.94)	Trans (9.81)	RF (4.16)
		ARF 12h	RF (73.64)	EMERGE TS-only (5.41)	RF (3.19)
		Shock 4h	RF (73.47)	RF (11.70)	TCN (8.15)
		Shock 12h	RF (72.75)	RF (9.70)	EMERGE TS-only (5.97)
eICU	eICU	Mortality 48h	Trans (87.38)	Trans (55.65)	RF (36.24)
		ARF 4h	RF (93.59)	Trans (35.49)	RF (30.07)
		ARF 12h	RF (91.37)	Trans (32.75)	TCN (17.83)
		Shock 4h	RF (81.73)	RF (26.52)	RF (19.95)
		Shock 12h	RF (82.44)	EMERGE TS-only (22.95)	RF (16.74)
	MIMIC-IV	Mortality 48h	Trans (83.74)	Trans (48.20)	Trans (37.35)
		ARF 4h	RF (67.30)	Trans (51.61)	Trans (42.82)
		ARF 12h	RF (67.15)	Trans (41.84)	Trans (32.01)
		Shock 4h	RF (77.67)	RF (20.42)	SVM (12.33)
		Shock 12h	RF (78.97)	EMERGE TS-only (15.50)	SVM (8.20)

The METRE-derived NN models remained competitive, especially in mortality and in some external evaluation settings, but their advantage was not systematic. EMERGE TS-only provided an additional perspective by showing that an external TS architecture could be connected to the METRE-derived representation and evaluated under the same protocol. However, its results should be interpreted as the performance of the structured TS component only, rather than as a reproduction of the full multimodal EMERGE framework.

The strongest overall pattern was the difference between internal and external evaluation. Several models achieved strong internal AUROC values, but external performance decreased substantially, particularly for positive-class metrics in ARF and shock. In several cases, models preserved moderate ranking ability while showing low AUPRC or $\min(+P, Se)$, indicating that discrimination did not necessarily translate into balanced positive-event detection. Taken together, these results show that cross-database evaluation substantially changes the interpretation of model performance and is essential for assessing whether ICU prediction models are likely to generalise beyond their source database.

Chapter 6

Conclusions

This dissertation addressed the challenge of developing and evaluating clinical predictive models for ICU data under conditions of data heterogeneity, incomplete reproducibility, and cross-database variability. The work was motivated by the observation that robust AI-based clinical DSSs depend not only on model architecture, but also on reusable data pipelines, clinically meaningful task definitions, and evaluation protocols capable of assessing generalisation beyond a single database.

One of the major contributions of this dissertation was the structured review of the literature supporting the methodological choices of the work. Following a transparent search and screening strategy inspired by PRISMA, two complementary research directions were analysed: medical ICU data structuring pipelines and Generative AI models applied to medical data. The first review showed that harmonisation pipelines have substantially improved reproducibility, interoperability, and cross-dataset processing, but often remain focused on structural consistency, specific clinical tasks, or model-dependent workflows. The second review showed that Transformer- and LLM-based approaches have expanded the capacity to model longitudinal, heterogeneous, and multimodal clinical data, but are frequently model-centric and weakly connected to reusable, clinically grounded data pipelines.

The critical synthesis of these two reviews revealed a central methodological gap. Current ICU data pipelines and modern clinical AI models are often developed in parallel rather than as part of an integrated ecosystem. Pipelines may provide structured and interoperable data representations without being designed for flexible downstream modelling, while recent generative or reasoning-based models often assume customised representations that limit reproducibility and transferability. This observation motivated the experimental direction of the dissertation: to investigate whether a reusable ICU harmonisation pipeline could support benchmark-oriented predictive modelling across multiple model families and databases.

To address this gap, the open-source METRE pipeline was reproduced and improved to current MIMIC-IV and eICU data, and a unified modelling and evaluation workflow was built around its outputs. This required updating database extraction procedures, reconstructing auxiliary tables, implementing the missing HDF5-to-NPY conversion step, validating the generated model-ready

tensors, and standardising the training workflow across all selected tasks and models. The resulting benchmark included classical baseline models, namely LR, RF, SVM, and NB, the METRE-derived NN models, RNN, TCN, and Transformer encoder, and an adapted external EMERGE TS-only model. This design allowed the same harmonised representation to be evaluated across different modelling families, supporting the objective of testing whether the adapted pipeline could serve as a reusable substrate for predictive modelling.

The experimental results demonstrated that the main objectives of the dissertation were achieved. The adapted workflow supported training, validation, and testing across the selected prediction tasks, including hospital mortality, ARF, and shock. Hyperparameter optimisation through Optuna improved performance in several model-task configurations, although the effect was not uniform across all metrics or models. Mortality was consistently the most robust and transferable task, with stronger internal performance and less severe degradation under external evaluation. In contrast, ARF and shock were more fragile, especially when assessed through AUPRC and $\min(+P, Se)$, reflecting their stronger dependence on support and treatment variables, fewer positive cases in some settings, and sensitivity to database-specific documentation practices.

The comparison between model families also produced important findings. RF was consistently one of the strongest models, particularly for ARF and shock, showing that higher architectural complexity did not necessarily translate into superior performance when the structured representation already retained relevant predictive signal. The METRE-derived NN models remained competitive, particularly in mortality and some external settings, but their advantage was not systematic. The EMERGE TS-only adaptation demonstrated that an external TS architecture could be connected to the METRE-derived representation and evaluated under the same protocol. However, it did not reproduce the full multimodal EMERGE framework and did not consistently outperform the other models. Together, these results reinforce that model performance in clinical prediction depends on the interaction between data representation, task definition, model architecture, and evaluation setting.

Another central contribution of this dissertation was the emphasis on external evaluation. Several models achieved strong internal results, but performance decreased substantially under cross-database testing, especially for positive-class metrics in ARF and shock. This confirms that internal validation alone can overestimate the robustness of clinical predictive models. Cross-database evaluation was therefore essential for revealing whether models were learning transferable clinical patterns or database-specific regularities. In this sense, the work contributes not only a structured synthesis of the literature and a set of experimental results, but also a reproducible benchmark-oriented evaluation framework for studying generalisation in ICU prediction.

6.1 Limitations

Several limitations should be acknowledged. First, the adapted METRE workflow depended on reconstructing components that were not fully available in the public repository. Some auxiliary tables expected by the extraction scripts had to be recreated from available database tables, and

the conversion from extracted HDF5 files to the NPY inputs required by the training code had to be implemented in this dissertation. Although validation checks were performed, residual differences in feature ordering, temporal alignment, or intervention encoding may have influenced task-specific results, particularly for ARF, whose labels depend directly on respiratory support variables.

Second, the EMERGE adaptation was restricted to the TS-only configuration. The public EMERGE repository provided the model implementation, but not the complete preprocessing artefacts required to reproduce the full multimodal framework, including processed note embeddings, generated summaries, and retrieved knowledge records. As a result, this dissertation evaluated whether the structured TS component of EMERGE could operate on METRE-derived data, but did not reproduce the complete multimodal and retrieval-augmented EMERGE architecture.

Third, although the evaluation protocol was unified across models, the experiments remained limited to two public critical care databases and five prediction settings. The results therefore provide evidence of cross-database generalisation between MIMIC-IV and eICU, but not across all possible ICU populations, healthcare systems, or clinical tasks. In addition, the selected 6-hour gap and predefined observation windows were retained to remain aligned with the original METRE configuration, rather than exhaustively exploring alternative temporal definitions.

The literature review also has limitations. Although the search strategy was structured and designed to provide broad coverage of the two research directions, the search strings had to remain sufficiently focused to keep the review feasible within the scope of the dissertation. As a result, relevant works using alternative terminology, addressing adjacent clinical tasks, or focusing on related modelling settings may not have been captured. Nevertheless, the review provided a broad and coherent basis for identifying the methodological gap addressed in this dissertation.

6.2 Future Directions

One immediate direction is the further validation of the reconstructed METRE conversion workflow, particularly for tasks derived from the Intervention table. A detailed audit of respiratory support variables, vasoactive treatment variables, temporal alignment rules, and task-label generation would help determine whether the weaker or less stable ARF results reflect model limitations or remaining differences in the adapted preprocessing workflow. Additional experiments could also assess alternative observation windows and gap periods, since the temporal distance between observed data and event onset may substantially affect ARF and shock prediction.

A second direction concerns the extension of the EMERGE adaptation beyond the TS-only setting. Incorporating MIMIC-IV-Note [183], local clinical text embeddings, or privacy-compliant summary generation would allow the multimodal and retrieval-augmented components of EMERGE to be evaluated more faithfully. This would also make it possible to investigate whether clinical notes, generated summaries, and external medical knowledge improve cross-database robustness beyond structured TS data alone.

The benchmark could also be expanded to additional models reviewed in the literature, including architectures based on FHIR, diagnostic code sequences, clinical notes, generated reports, or multimodal fusion. Incorporating models such as FHIR-Former [145], ColaCare [162], CPLLM [133], DearLLM [148], CLARA [160], VaDeSC-EHR [131], or CLEAR [159] would require extending the current METRE-based workflow beyond structured TS tensors. Depending on the model, this could involve mapping data into FHIR-compatible resources, adding diagnostic code sequences, defining new prediction tasks, integrating clinical notes, or adapting model input layers to accept the METRE representation. Future work could also compare the METRE-based representation with alternative ICU harmonisation frameworks to assess how different data representations affect model generalisation.

More broadly, future research on AI-based ICU DSSs should continue to prioritise external validation, uncertainty estimation, interpretability, and clinical usability. The results of this dissertation show that strong internal performance is not sufficient evidence of clinical robustness. Models intended for critical care must be evaluated across heterogeneous databases, aligned with clinical workflows, and designed around transparent data processing assumptions. By combining structured literature analysis, reusable ICU data harmonisation, benchmark-oriented modelling, and systematic cross-database evaluation, this dissertation provides a foundation for developing more reliable and generalisable clinical AI systems.

References

[1] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2017.2767063.

[2] G. S. Birkhead, M. Klompas, and N. R. Shah, "Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health," en, *Annual Review of Public Health*, vol. 36, no. 1, pp. 345–359, Mar. 2015, ISSN: 0163-7525, 1545-2093. DOI: 10.1146/annurev-publhealth-031914-122747.

[3] M. Kreuzthaler, S. Schulz, and A. Berghold, "Secondary use of electronic health records for building cohort studies through top-down information extraction," en, *Journal of Biomedical Informatics*, vol. 53, pp. 188–195, Feb. 2015, ISSN: 15320464. DOI: 10.1016/j.jbi.2014.10.010.

[4] R. S. Ledley and L. B. Lusted, "Reasoning Foundations of Medical Diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason," en, *Science*, vol. 130, no. 3366, pp. 9–21, Jul. 1959, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.130.3366.9.

[5] E. S. Berner and T. J. La Lande, "Overview of Clinical Decision Support Systems," in *Clinical Decision Support Systems*, K. J. Hannah, M. J. Ball, and E. S. Berner, Eds., Series Title: Health Informatics, New York, NY: Springer New York, 2007, pp. 3–22, ISBN: 978-0-387-38319-4. DOI: 10.1007/978-0-387-38319-4_1.

[6] J. Timiliotis, B. Blümke, P. D. Serfözö, *et al.*, "A Novel Diagnostic Decision Support System for Medical Professionals: Prospective Feasibility Study," en, *JMIR Formative Research*, vol. 6, no. 3, e29943, Mar. 2022, ISSN: 2561-326X. DOI: 10.2196/29943.

[7] B. Middleton, D. F. Sittig, and A. Wright, "Clinical Decision Support: A 25 Year Retrospective and a 25 Year Vision," eng, *Yearbook of Medical Informatics*, vol. Suppl 1, no. Suppl 1, S103–116, Aug. 2016, ISSN: 2364-0502. DOI: 10.15265/IYS-2016-s034.

[8] C. Ohmann and W. Kuchinke, "Future Developments of Medical Informatics from the Viewpoint of Networked Clinical Research: Interoperability and Integration," en, *Methods of Information in Medicine*, vol. 48, no. 01, pp. 45–54, 2009, ISSN: 0026-1270, 2511-705X. DOI: 10.3414/ME9137.

[9] J. Walker, E. Pan, D. Johnston, J. Adler-Milstein, D. W. Bates, and B. Middleton, "The Value Of Health Care Information Exchange And Interoperability: There is a business case to be made for spending money on a fully standardized nationwide system.," en, *Health Affairs*, vol. 24, no. Suppl1, W5–10–W5–18, Jan. 2005, ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.W5.10.

[10] L. M. Etheredge, "A Rapid-Learning Health System: What would a rapid-learning health system look like, and how might we get there?" en, *Health Affairs*, vol. 26, no. Suppl1, w107–w118, Jan. 2007, ISSN: 0278-2715, 1544-5208. DOI: 10.1377/hlthaff.26.2.w107.

[11] C. P. Friedman, A. K. Wong, and D. Blumenthal, "Achieving a Nationwide Learning Health System," en, *Science Translational Medicine*, vol. 2, no. 57, Nov. 2010, ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.3001456.

[12] M.-H. Luo, D.-L. Huang, J.-C. Luo, *et al.*, "Data science in the intensive care unit," eng, *World Journal of Critical Care Medicine*, vol. 11, no. 5, pp. 311–316, Sep. 2022, ISSN: 2220-3141. DOI: 10.5492/wjccm.v11.i5.311.

[13] T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi, "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," eng, *Scientific Data*, vol. 5, p. 180 178, Sep. 2018, ISSN: 2052-4463. DOI: 10.1038/sdata.2018.178.

[14] J. M. Boss, G. Narula, C. Straessle, *et al.*, "ICU Cockpit: A platform for collecting multimodal waveform data, AI-based computational disease modeling and real-time decision support in the intensive care unit," eng, *Journal of the American Medical Informatics Association: JAMIA*, vol. 29, no. 7, pp. 1286–1291, Jun. 2022, ISSN: 1527-974X. DOI: 10.1093/jamia/ocac064.

[15] J. Bajwa, U. Munir, A. Nori, and B. Williams, "Artificial intelligence in healthcare: Transforming the practice of medicine," eng, *Future Healthcare Journal*, vol. 8, no. 2, e188–e194, Jul. 2021, ISSN: 2514-6645. DOI: 10.7861/fhj.2021-0095.

[16] S. Barbieri, J. Kemp, O. Perez-Concha, *et al.*, *Benchmarking Deep Learning Architectures for Predicting Readmission to the ICU and Describing Patients-at-Risk*, Version Number: 3, 2019. DOI: 10.48550/ARXIV.1905.08547.

[17] A. E. W. Johnson, L. Bulgarelli, L. Shen, *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.

[18] L. Lijović and P. Elbers, "Leveraging the power of routinely collected ICU data," en, *Intensive Care Medicine*, vol. 51, no. 1, pp. 163–166, Jan. 2025, ISSN: 0342-4642, 1432-1238. DOI: 10.1007/s00134-024-07745-5.

[19] A. Jagesar, T. Dam, T. Struja, *et al.*, "Sharing is caring: A systematic review of publicly available intensive care data sets," en, *Journal of Critical Care*, vol. 90, p. 155 205, Dec. 2025, ISSN: 08839441. DOI: 10.1016/j.jcrc.2025.155205.

[20] J. H. Yoon, M. R. Pinsky, and G. Clermont, "Artificial Intelligence in Critical Care Medicine," en, *Critical Care*, vol. 26, no. 1, p. 75, Mar. 2022, ISSN: 1364-8535. DOI: 10.1186/s13054-022-03915-3.

[21] M. Imhoff, "Acquisition of ICU data: Concepts and demands," en, *International Journal of Clinical Monitoring and Computing*, vol. 9, no. 4, pp. 229–237, Dec. 1992, ISSN: 0167-9945. DOI: 10.1007/BF01133618.

[22] W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "APACHE II: A severity of disease classification system," eng, *Critical Care Medicine*, vol. 13, no. 10, pp. 818–829, Oct. 1985, ISSN: 0090-3493.

[23] R. Moreno, A. Rhodes, L. Piquilloud, *et al.*, "The Sequential Organ Failure Assessment (SOFA) Score: Has the time come for an update?" en, *Critical Care*, vol. 27, no. 1, p. 15, Jan. 2023, ISSN: 1364-8535. DOI: 10.1186/s13054-022-04290-9.

[24] M. Greco, P. F. Caruso, and M. Cecconi, "Artificial Intelligence in the Intensive Care Unit," en, *Seminars in Respiratory and Critical Care Medicine*, vol. 42, no. 01, pp. 002–009, Feb. 2021, ISSN: 1069-3424, 1098-9048. DOI: 10.1055/s-0040-1719037.

[25] T. Sinuff, D. J. Cook, and M. Giacomini, "How qualitative research can contribute to research in the intensive care unit," en, *Journal of Critical Care*, vol. 22, no. 2, pp. 104–111, Jun. 2007, ISSN: 08839441. DOI: 10.1016/j.jcrc.2007.03.001.

[26] B. Foreman, I. A. Lissak, N. Kamireddi, D. Moberg, and E. S. Rosenthal, "Challenges and Opportunities in Multimodal Monitoring and Data Analytics in Traumatic Brain Injury," en, *Current Neurology and Neuroscience Reports*, vol. 21, no. 3, p. 6, Mar. 2021, ISSN: 1528-4042, 1534-6293. DOI: 10.1007/s11910-021-01098-y.

[27] G. J. Kuperman, R. M. Gardner, and T. A. Pryor, "Medical Information Bus," en, in *HELP: A Dynamic Hospital Information System*, H. F. Orthner, Ed., Series Title: Computers and Medicine, New York, NY: Springer New York, 1991, pp. 206–211, ISBN: 978-1-4612-3070-0. DOI: 10.1007/978-1-4612-3070-0_20.

[28] M. Cabeleira, A. Ercole, and P. Smielewski, "HDF5-Based Data Format for Archiving Complex Neuro-monitoring Data in Traumatic Brain Injury Patients," in *Intracranial Pressure & Neuromonitoring XVI*, T. Heldt, Ed., vol. 126, Series Title: Acta Neurochirurgica Supplement, Cham: Springer International Publishing, 2018, pp. 121–125, ISBN: 978-3-319-65797-4. DOI: 10.1007/978-3-319-65798-1_26.

[29] J. G. Klann, M. A. H. Joss, K. Embree, and S. N. Murphy, "Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model," en, *PLOS ONE*, vol. 14, no. 2, C. Lovis, Ed., e0212463, Feb. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0212463.

[30] S. Kamm, S. S. Veekati, T. Müller, N. Jazdi, and M. Weyrich, "A survey on machine learning based analysis of heterogeneous data in industrial automation," en, *Computers in Industry*, vol. 149, p. 103 930, Aug. 2023, ISSN: 01663615. DOI: 10.1016/j.compind.2023.103930.

[31] L. Zhang, X. Chen, T. Chen, Z. Wang, and B. J. Mortazavi, "DynEHR: Dynamic adaptation of models with data heterogeneity in electronic health records," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Athens, Greece: IEEE, Jul. 2021, pp. 1–4, ISBN: 978-1-6654-0358-0. DOI: 10.1109/BHI50953.2021.9508558.

[32] N. Rodemund, B. Wernly, C. Jung, C. Cozowicz, and A. Koköfer, "Harnessing Big Data in Critical Care: Exploring a new European Dataset," en, *Scientific Data*, vol. 11, no. 1, p. 320, Mar. 2024, ISSN: 2052-4463. DOI: 10.1038/s41597-024-03164-9.

[33] M. Gezimati and G. Singh, "Deep learning for multisource medical information processing," en, in *Data Fusion Techniques and Applications for Smart Healthcare*, Elsevier, 2024, pp. 45–76, ISBN: 978-0-443-13233-9. DOI: 10.1016/B978-0-44-313233-9.00009-6.

[34] M. Lorenzi, M. Deprez, I. Balelli, A. L. Aguila, and A. Altmann, "Integration of Multimodal Data," en, in *Machine Learning for Brain Disorders*, O. Colliot, Ed., vol. 197, Series Title: Neuromethods, New York, NY: Springer US, 2023, pp. 573–597, ISBN: 978-1-0716-3194-2. DOI: 10.1007/978-1-0716-3195-9_19.

[35] J. Liu, X. Cen, C. Yi, *et al.*, "Challenges in AI-driven Biomedical Multimodal Data Fusion and Analysis," en, *Genomics, Proteomics & Bioinformatics*, vol. 23, no. 1, X. Zhang, Ed., qzaf011, May 2025, ISSN: 1672-0229, 2210-3244. DOI: 10.1093/gpbjnl/qzaf011.

[36] F. Cremonesi, V. Planat, V. Kalokyri, *et al.*, "The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform," en, *Journal of Biomedical Informatics*, vol. 141, p. 104 338, May 2023, ISSN: 15320464. DOI: 10.1016/j.jbi.2023.104338.

[37] J. M. L. Alcaraz, X. L. Moran, E. D. Zaragoza, *et al.*, *A Multimodal Deep Learning Framework for Predicting ICU Deterioration: Integrating ECG Waveforms with Clinical Data and Clinician Benchmarking*, Version Number: 1, 2026. DOI: 10.48550/ARXIV.2601.06645.

[38] N. Ardic and R. Dinc, "Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review," en, *Health Informatics Journal*, vol. 31, no. 3, p. 14 604 582 251 366 141, Jul. 2025, ISSN: 1460-4582, 1741-2811. DOI: 10.1177/14604582251366141.

[39] D. J. Vreeman, C. J. McDonald, and S. M. Huff, "LOINC®: A universal catalogue of individual clinical observations and uniform representation of enumerated collections," en, *International Journal of Functional Informatics and Personalised Medicine*, vol. 3, no. 4, p. 273, 2010, ISSN: 1756-2104, 1756-2112. DOI: 10.1504/IJFIPM.2010.040211.

[40] P. Raghavan, J. L. Chen, E. Fosler-Lussier, and A. M. Lai, *How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?* Version Number: 1, 2015. DOI: 10.48550/ARXIV.1502.04049.

[41] F. Mohsen, H. Ali, N. El Hajj, and Z. Shah, "Artificial intelligence-based methods for fusion of electronic health records and imaging data," en, *Scientific Reports*, vol. 12, no. 1, p. 17 981, Oct. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-22514-4.

[42] J. Ye, J. Hai, J. Song, and Z. Wang, "Multimodal Data Hybrid Fusion and Natural Language Processing for Clinical Prediction Models," eng, *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2024, pp. 191–200, 2024, ISSN: 2153-4063.

[43] C.-C. Chiu, C.-M. Wu, T.-N. Chien, L.-J. Kao, C. Li, and C.-M. Chu, "Integrating Structured and Unstructured EHR Data for Predicting Mortality by Machine Learning and Latent Dirichlet Allocation Method," en, *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 4340, Feb. 2023, ISSN: 1660-4601. DOI: 10.3390/ijerph20054340.

[44] J. C. Smith, S. E. Davis, R. M. Reeves, *et al.*, "Characterization and comparison of structured and unstructured electronic health record data mapped to MedDRA for post-marketing surveillance," en, *JAMIA Open*, vol. 9, no. 2, ooag025, Mar. 2026, ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooag025.

[45] R. Zhang, Y. Chen, W. Yue, *et al.*, "Multimodal artificial intelligence in medicine: A task-oriented framework for clinical translation," *Frontiers in Medicine*, vol. 12, p. 1 736 272, Jan. 2026, ISSN: 2296-858X. DOI: 10.3389/fmed.2025.1736272.

[46] C. N. Vorisek, M. Lehne, S. A. I. Klopfenstein, *et al.*, "Fast Healthcare Interoperability Resources (FHIR) for Interoperability in Health Research: Systematic Review," en, *JMIR Medical Informatics*, vol. 10, no. 7, e35724, Jul. 2022, ISSN: 2291-9694. DOI: 10.2196/35724.

[47] F. FitzHenry, F. Resnic, S. Robbins, *et al.*, "Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership," en, *Applied Clinical Informatics*, vol. 06, no. 03, pp. 536–547, 2015, ISSN: 1869-0327. DOI: 10.4338/ACI-2014-12-CR-0121.

[48] World Health Organization, *International Statistical Classification of Diseases and Related Health Problems (ICD)*, 2024. [Online]. Available: https://www.who.int/standards/classifications/classification-of-diseases.

[49] M. S. C. Almeida, L. F. D. Sousa Filho, P. M. Rabello, and B. M. Santiago, "Classificação Internacional das Doenças - 11ª revisão: Da concepção à implementação," *Revista de Saúde Pública*, vol. 54, p. 104, Dec. 2020, ISSN: 1518-8787, 0034-8910. DOI: 10.11606/s1518-8787.2020054002120.

[50] J. E. Harrison, S. Weber, R. Jakob, and C. G. Chute, "ICD-11: An international classification of diseases for the twenty-first century," en, *BMC Medical Informatics and Decision Making*, vol. 21, no. S6, p. 206, Nov. 2021, ISSN: 1472-6947. DOI: 10.1186/s12911-021-01534-6.

[51] National Library of Medicine, *SNOMED CT: Overview*, 2016. [Online]. Available: https://www.nlm.nih.gov/healthit/snomedct/snomed_overview.html.

[52] SNOMED International, *Benefits and Challenges (SNOMED CT Postcoordination Guide)*, 2026. [Online]. Available: https://docs.snomed.org/snomed-ct-practical-guides/snomed-ct-postcoordination-guide/snomed-ct-expressions/benefits-and-challenges.

[53] Regenstrief Institute, Inc., *LOINC Database*, 2026. [Online]. Available: https://loinc.org/about/.

[54] A. W. Forrey, C. J. McDonald, G. DeMoor, *et al.*, "Logical observation identifier names and codes (LOINC) database: A public use set of codes and names for electronic reporting of clinical laboratory test results," en, *Clinical Chemistry*, vol. 42, no. 1, pp. 81–90, Jan. 1996, ISSN: 0009-9147, 1530-8561. DOI: 10.1093/clinchem/42.1.81.

[55] Health Level Seven International (HL7), *FHIR Specification*, 2023. [Online]. Available: https://hl7.org/fhir/.

[56] L. Wang, A. Wen, S. Fu, *et al.*, "A scoping review of OMOP CDM adoption for cancer research using real world data," en, *npj Digital Medicine*, vol. 8, no. 1, p. 189, Apr. 2025, ISSN: 2398-6352. DOI: 10.1038/s41746-025-01581-7.

[57] K. Kersting, "Machine Learning and Artificial Intelligence: Two Fellow Travelers on the Quest for Intelligent Behavior in Machines," *Frontiers in Big Data*, vol. 1, p. 6, Nov. 2018, ISSN: 2624-909X. DOI: 10.3389/fdata.2018.00006.

[58] J. McCarthy, *WHAT IS ARTIFICIAL INTELLIGENCE?* Nov. 2007.

[59] S. Brown, *Machine learning, explained*, Apr. 2021. [Online]. Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

[60] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, "Introduction to Machine Learning, Neural Networks, and Deep Learning," eng, *Translational Vision Science & Technology*, vol. 9, no. 2, p. 14, Feb. 2020, ISSN: 2164-2591. DOI: 10.1167/tvst.9.2.14.

[61] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of Artificial Neural Networks and Deep Learning," en, in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Cham: Springer International Publishing, 2022, pp. 379–425, ISBN: 978-3-030-89009-4. DOI: 10.1007/978-3-030-89010-0_10.

[62] Y. Si, J. Du, Z. Li, *et al.*, "Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review," en, *Journal of Biomedical Informatics*, vol. 115, p. 103 671, Mar. 2021, ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103671.

[63] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," en, *Cognitive Robotics*, vol. 3, pp. 54–70, 2023, ISSN: 26672413. DOI: 10.1016/j.cogr.2023.04.001.

[64] D. Bergmann, *What is fine-tuning?* Mar. 2024. [Online]. Available: https://www.ibm.com/think/topics/fine-tuning.

[65] Y. Liu, *The Backstory of Backpropagation*, Dec. 2023. [Online]. Available: https://yuxi-liu-wired.github.io/essays/posts/backstory-of-backpropagation/#abstract.

[66] W.-K. Hong, "Factors influencing network trainings," en, in *Artificial Intelligence-Based Design of Reinforced Concrete Structures*, Elsevier, 2023, pp. 15–66, ISBN: 978-0-443-15252-8. DOI: 10.1016/B978-0-443-15252-8.00001-7.

[67] R. He, J. Cao, and T. Tan, "Generative artificial intelligence: A historical perspective," en, *National Science Review*, vol. 12, no. 5, nwaf050, Apr. 2025, ISSN: 2095-5138, 2053-714X. DOI: 10.1093/nsr/nwaf050.

[68] R. S. Segall, "Generative AI (Artificial Intelligence): What Is It? & What Are Its Inter- And Transdisciplinary Applications?" en, *Journal of Systemics, Cybernetics and Informatics*, vol. 23, no. 7, pp. 116–125, Dec. 2025, ISSN: 16904524. DOI: 10.54808/JSCI.23.07.116.

[69] M. Raza, Z. Jahangir, M. B. Riaz, M. J. Saeed, and M. A. Sattar, "Industrial applications of large language models," en, *Scientific Reports*, vol. 15, no. 1, p. 13 755, Apr. 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-98483-1.

[70] H. Naveed, A. U. Khan, S. Qiu, *et al.*, *A Comprehensive Overview of Large Language Models*, Version Number: 10, 2023. DOI: 10.48550/ARXIV.2307.06435.

[71] X. Meng, X. Yan, K. Zhang, *et al.*, "The application of large language models in medicine: A scoping review," en, *iScience*, vol. 27, no. 5, p. 109 713, May 2024, ISSN: 25890042. DOI: 10.1016/j.isci.2024.109713.

[72] L. A. Biesheuvel, J. D. Workum, M. Reuland, *et al.*, "Large language models in critical care," en, *Journal of Intensive Medicine*, S2667100X24001257, Dec. 2024, ISSN: 2667100X. DOI: 10.1016/j.jointm.2024.12.001.

[73] S. Ramlakhan, R. Saatchi, L. Sabir, *et al.*, "Understanding and interpreting artificial intelligence, machine learning and deep learning in Emergency Medicine," en, *Emergency Medicine Journal*, vol. 39, no. 5, pp. 380–385, May 2022, ISSN: 1472-0205, 1472-0213. DOI: 10.1136/emermed-2021-212068.

[74] D. Bergmann, *What is self-supervised learning?* Dec. 2023. [Online]. Available: https://www.ibm.com/think/topics/self-supervised-learning.

[75] B. L. Ortiz, V. Gupta, R. Kumar, *et al.*, "Data Preprocessing Techniques for AI and Machine Learning Readiness: Scoping Review of Wearable Sensor Data in Cancer Care," en, *JMIR mHealth and uHealth*, vol. 12, e59587, Sep. 2024, ISSN: 2291-5222. DOI: 10.2196/59587.

[76] I. Landi, B. S. Glicksberg, H.-C. Lee, *et al.*, "Deep representation learning of electronic health records to unlock patient stratification at scale," en, *npj Digital Medicine*, vol. 3, no. 1, p. 96, Jul. 2020, ISSN: 2398-6352. DOI: 10.1038/s41746-020-0301-z.

[77] A. Mumuni and F. Mumuni, "Automated data processing and feature engineering for deep learning and big data applications: A survey," en, *Journal of Information and Intelligence*, vol. 3, no. 2, pp. 113–153, Mar. 2025, ISSN: 29497159. DOI: 10.1016/j.jiixd.2024.01.002.

[78] O. A. Montesinos López, A. Montesinos López, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, en. Cham: Springer International Publishing, 2022, ISBN: 978-3-030-89009-4. DOI: 10.1007/978-3-030-89010-0.

[79] K. Kunanbayev, I. Temirbek, and A. Zollanvari, "Complex Encoding," in *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China: IEEE, Jul. 2021, pp. 1–6, ISBN: 978-1-6654-3900-8. DOI: 10.1109/IJCNN52387.2021.9534094.

[80] Lucas Marnoto Figueiredo, "An Overview of the main Machine Learning Models: From Theory to Algorithms," Master's dissertation, NOVA Information Management School, Universidade Nova de Lisboa, Lisboa, Portugal, Jun. 2020.

[81] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," en, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, Jul. 2019, pp. 2623–2631, ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330701.

[82] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, Version Number: 7, 2017. DOI: 10.48550/ARXIV.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762 (visited on 03/31/2025).

[83] J. Ferrer, *How Transformers Work: A Detailed Exploration of Transformer Architecture*, Jan. 2024. [Online]. Available: https://www.datacamp.com/tutorial/how-transformers-work?.

[84] D. Soydaner, "Attention Mechanism in Neural Networks: Where it Comes and Where it Goes," 2022, Version Number: 1. DOI: 10.48550/ARXIV.2204.13154. [Online]. Available: https://arxiv.org/abs/2204.13154 (visited on 03/31/2025).

[85] R. B. B. D. SANTOS, "IMPROVING THE ROBUSTNESS OF MULTIMODAL AI WITH ASYNCHRONOUS AND MISSING INPUTS," DOCTORATE IN BIOMEDICAL ENGINEERING, NOVA University Lisbon, Sep. 2024.

[86] J. Egbert and L. Plonsky, "Bootstrapping Techniques," en, in *A Practical Handbook of Corpus Linguistics*, M. Paquot and S. T. Gries, Eds., Cham: Springer International Publishing, 2020, pp. 593–610, ISBN: 978-3-030-46215-4 978-3-030-46216-1. DOI: 10.1007/978-3-030-46216-1_24.

[87] A. E. Ayintareba, D. R. Korda, and O. Owusu-Boateng, "Performance Metrics in Machine Learning: A Comprehensive Narrative Review," *Journal of Advances in Mathematics and Computer Science*, vol. 41, no. 5, pp. 183–199, May 2026, ISSN: 2456-9968. DOI: 10.9734/jamcs/2026/v41i52145.

[88] Y. Zhu, C. Ren, Z. Wang, *et al.*, "EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation," *International Conference on Information and Knowledge Management, Proceedings*, pp. 3549–3559, 2024. DOI: 10.1145/3627673.3679582.

[89] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," en, *BMJ*, n71, Mar. 2021, ISSN: 1756-1833. DOI: 10.1136/bmj.n71.

[90] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews," en, *Systematic Reviews*, vol. 5, no. 1, p. 210, Dec. 2016, ISSN: 2046-4053. DOI: 10.1186/s13643-016-0384-4.

[91] L. Guo, S. Pfohl, J. Fries, *et al.*, "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine," *Scientific Reports*, vol. 12, no. 1, 2022. DOI: 10.1038/s41598-022-06484-1.

[92] L.-H. Fu, C. Knaplund, K. Dwain Cato, *et al.*, "Utilizing timestamps of longitudinal electronic health record data to classify clinical deterioration events," *Journal of the American Medical Informatics Association*, vol. 28, no. 9, pp. 1955–1963, 2021. DOI: 10.1093/jamia/ocab111.

[93] R. M. Thomas, I. Koh, K. Wilkinson, *et al.*, "Development of a Computable Phenotype for Hospital-Acquired Venous Thrombosis: The Medical Inpatient Thrombosis and Hemostasis (MITH) Study," *Blood*, vol. 136, pp. 40–41, Nov. 2020, ISSN: 0006-4971. DOI: https://doi.org/10.1182/blood-2020-141047.

[94] IMPACT consortium, Y. Kim, M. Park, *et al.*, "Optimizing Nursing Data for Al and CDSS: SNOMED CT Mapping Using K-MIMIC," in *Studies in Health Technology and Informatics*, M. S. Househ, Z. U. A. Tariq, M. Al-Zubaidi, U. Shah, and E. Huesing, Eds., IOS Press, Aug. 2025, ISBN: 978-1-64368-608-0. DOI: 10.3233/SHTI250799.

[95] D. Püttmann, R. de Groot, N. de Keizer, *et al.*, "Assessing the FAIRness of databases on the EHDEN portal: A case study on two Dutch ICU databases," *International Journal of Medical Informatics*, vol. 176, 2023. DOI: 10.1016/j.ijmedinf.2023.105104.

[96] D. F. Stein, M. J. Carter, J. Booth, *et al.*, "Predicting Cardiovascular deterioration in a paediatric intensive care unit (PicEWS): A machine learning modelling study of routinely collected health-care data," *eClinicalMedicine*, vol. 85, p. 103 255, Jul. 2025, ISSN: 2589-5370. DOI: https://doi.org/10.1016/j.eclinm.2025.103255.

[97] A. Jagesar, M. Otten, T. Dam, *et al.*, "Comparative performance of intensive care mortality prediction models based on manually curated versus automatically extracted electronic health record data," *International Journal of Medical Informatics*, vol. 188, p. 105 477, Aug. 2024, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2024.105477.

[98] P. J. Thoral, J. M. Peppink, R. H. Driessen, *et al.*, "Sharing ICU Patient Data Responsibly Under the Society of Critical Care Medicine/European Society of Intensive Care Medicine Joint Data Science Collaboration: The Amsterdam University Medical Centers Database (AmsterdamUMCdb) Example*," en, *Critical Care Medicine*, vol. 49, no. 6, e563–e577, Jun. 2021, ISSN: 0090-3493. DOI: 10.1097/CCM.0000000000004916.

[99] X. Wang, T. Iwashyna, H. Prescott, V. Valbuena, and S. Seelye, "Pulse oximetry and supplemental oxygen use in nationwide Veterans Health Administration hospitals, 2013-2017: A Veterans Affairs Patient Database validation study," *BMJ Open*, vol. 11, no. 10, 2021. DOI: 10.1136/bmjopen-2021-051978.

[100] N. Bennett, D. Plečko, I.-F. Ukor, N. Meinshausen, and P. Bûhlmann, "Ricu: R's interface to intensive care data," *GigaScience*, vol. 12, 2023. DOI: 10.1093/gigascience/giad041.

[101] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "MIMIC-III, a freely accessible critical care database," en, *Scientific Data*, vol. 3, no. 1, p. 160 035, May 2016, ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35.

[102] S. L. Hyland, M. Faltys, M. Hüser, *et al.*, "Early prediction of circulatory failure in the intensive care unit using machine learning," en, *Nature Medicine*, vol. 26, no. 3, pp. 364–373, Mar. 2020, ISSN: 1078-8956, 1546-170X. DOI: 10.1038/s41591-020-0789-4.

[103] W. Liao and J. Voldman, "A Multidatabase ExTRaction PipEline (METRE) for facile cross validation in critical care research," *Journal of Biomedical Informatics*, vol. 141, 2023. DOI: 10.1016/j.jbi.2023.104356.

[104] M. Oliver, J. Allyn, R. Carencotte, N. Allou, and C. Ferdynus, "Introducing the BlendedICU dataset, the first harmonized, international intensive care dataset," *Journal of Biomedical Informatics*, vol. 146, 2023. DOI: 10.1016/j.jbi.2023.104502.

[105] R. JC, L. PG, C. K, *et al.*, "A common longitudinal intensive care unit data format (CLIF) for critical illness research.," eng, *Intensive care medicine*, vol. 51, no. 3, pp. 556–569, Mar. 2025, ISSN: 1432-1238 (Electronic). DOI: 10.1007/s00134-025-07848-7.

[106] N. Paris, A. Lamer, and A. Parrot, "Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study," *JMIR Medical Informatics*, vol. 9, no. 12, Dec. 2021, ISSN: 2291-9694. DOI: https://doi.org/10.2196/30970.

[107] L. Bode, M. Mast, H. Rathert, T. Jack, and A. Wulff, "Achieving Interoperable Datasets in Pediatrics: A Data Integration Approach," *Studies in Health Technology and Informatics*, vol. 305, pp. 327–330, 2023. DOI: 10.3233/SHTI230496.

[108] L. Fleuren, T. Dam, M. Tonutti, *et al.*, "The Dutch Data Warehouse, a multicenter and full-admission electronic health records database for critically ill COVID-19 patients," *Critical Care*, vol. 25, no. 1, 2021. DOI: 10.1186/s13054-021-03733-z.

[109] M. Hofford, S. Yu, A. Johnson, A. Lai, P. Payne, and A. Michelson, "OpenSep: A generalizable open source pipeline for SOFA score calculation and Sepsis-3 classification," *JAMIA Open*, vol. 5, no. 4, 2022. DOI: 10.1093/jamiaopen/ooac105.

[110] C. Porschen, J. Ernsting, P. Brauckmann, *et al.*, "pyAKI—An open source solution to automated acute kidney injury classification," en, *PLOS ONE*, vol. 20, no. 1, M.-Y. Chothia, Ed., e0315325, Jan. 2025, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0315325.

[111] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. Sjoding, and J. Wiens, "Democratizing EHR analyses with FIDDLE: A flexible data-driven preprocessing pipeline for structured clinical data," *Journal of the American Medical Informatics Association*, vol. 27, no. 12, pp. 1921–1934, 2020. DOI: 10.1093/jamia/ocaa139.

[112] R. van de Water, H. Schmidt, P. Elbers, P. Thoral, B. Arnrich, and P. Rockenschaub, "YET ANOTHER ICU BENCHMARK: A FLEXIBLE MULTI-CENTER FRAMEWORK FOR CLINICAL ML," 2024.

[113] J. Gao, Y. Zhu, W. Wang, *et al.*, "A comprehensive benchmark for COVID-19 predictive modeling using electronic health records in intensive care," *Patterns*, vol. 5, no. 4, 2024. DOI: 10.1016/j.patter.2024.100951.

[114] Ó. Escudero-Arnanz, C. Soguero-Ruiz, and A. G. Marques, "Explainable Spatio-Temporal GCNNs for Irregular Multivariate Time Series: Architecture and Application to ICU Patient Data," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 11, pp. 1286–1301, 2025, ISSN: 2373-776X, 2373-7778. DOI: 10.1109/TSIPN.2025.3613951.

[115] F. Xu, F. Morales, and L. Nunes Amaral, "Robust extraction of pneumonia-associated clinical states from electronic health records," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, no. 45, 2024. DOI: 10.1073/pnas.2417688121.

[116] Y. Wang, J. Stroh, G. Hripcsak, *et al.*, "A methodology of phenotyping ICU patients from EHR data: High-fidelity, personalized, and interpretable phenotypes estimation," *Journal of Biomedical Informatics*, vol. 148, 2023. DOI: 10.1016/j.jbi.2023.104547.

[117] N. Jajcay, B. Bezák, A. Segev, *et al.*, "Data processing pipeline for cardiogenic shock prediction using machine learning," *Frontiers in Cardiovascular Medicine*, vol. 10, 2023. DOI: 10.3389/fcvm.2023.1132680.

[118] Z. Liu, Y. Hu, X. Wu, G. Mertes, Y. Yang, and D. Clifton, "Patient Clustering for Vital Organ Failure Using ICD Code With Graph Attention," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 8, pp. 2329–2337, 2023. DOI: 10.1109/TBME.2023.3243311.

[119] S. Gupta, S. Sharma, R. Sharma, and J. Chandra, "Healing with hierarchy: Hierarchical attention empowered graph neural networks for predictive analysis in medical data," *Artificial Intelligence in Medicine*, vol. 165, p. 103 134, Jul. 2025, ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2025.103134.

[120] N. Kotonya and F. Toni, *Explainable Automated Fact-Checking for Public Health Claims*, Version Number: 1, 2020. DOI: 10.48550/ARXIV.2010.09926.

[121] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, Version Number: 2, 2018. DOI: 10.48550/ARXIV.1810.04805.

[122] S. Nowak, H. Schneider, Y. Layer, *et al.*, "Development of image-based decision support systems utilizing information extracted from radiological free-text report databases with text-based transformers," *European Radiology*, vol. 34, no. 5, pp. 2895–2904, 2024. DOI: 10.1007/s00330-023-10373-0.

[123] T. Dam, L. Fleuren, L. Roggeveen, *et al.*, "Augmented intelligence facilitates concept mapping across different electronic health records," *International Journal of Medical Informatics*, vol. 179, 2023. DOI: 10.1016/j.ijmedinf.2023.105233.

[124] H. Fan, S. C. Rossetti, J. Thate, R. Mugoya, A. M. Lai, and P.-Y. Yen, "Semi-automated pipeline to accelerate multi-site flowsheet alignment and concept mapping in electronic health records," en, *Journal of the American Medical Informatics Association*, vol. 32, no. 7, pp. 1140–1148, Jul. 2025, ISSN: 1067-5027, 1527-974X. DOI: 10.1093/jamia/ocaf076.

[125] M. Yang, G. Kwak, T. Pollard, L. Celi, and M. Ghassemi, "Evaluating the Impact of Social Determinants on Health Prediction in the Intensive Care Unit," pp. 333–350, 2023. DOI: 10.1145/3600211.3604719.

[126] K. Mani, T. Scharfenberger, S. N. Goldman, *et al.*, "Multimodal machine learning for predicting perioperative safety indicators in spinal surgery," *The Spine Journal*, Mar. 2025, ISSN: 1529-9430. DOI: https://doi.org/10.1016/j.spinee.2025.03.021.

[127] S. H. Noteboom, E. Kho, M. Galanty, *et al.*, "From intensive care monitors to cloud environments: A structured data pipeline for advanced clinical decision support," *eBioMedicine*, vol. 111, p. 105 529, Jan. 2025, ISSN: 2352-3964. DOI: https://doi.org/10.1016/j.ebiom.2024.105529.

[128] Á. Ritoré, A. M. Oprescu, A. Estirado Bronchalo, and M. Á. Armengol de la Hoz, *COVID Data for Shared Learning (CDSL): A comprehensive, multimodal COVID-19 dataset from HM Hospitales*. DOI: 10.13026/1176-6C44.

[129] V. Dwivedi, V. Schlegel, A. Liu, *et al.*, "Representation Learning of Structured Data for Medical Foundation Models," *Proceedings of Machine Learning Research*, vol. 285, 2024.

[130] Y. Wang, X. Peng, T. Shen, *et al.*, "Soft Prompt Transfer for Zero-Shot and Few-Shot Learning in EHR Understanding," *Lecture Notes in Computer Science*, vol. 14178, pp. 18–32, 2023. DOI: 10.1007/978-3-031-46671-7_2.

[131] J. Qiu, Y. Hu, L. Li, *et al.*, "Deep representation learning for clustering longitudinal survival data from electronic health records," *Nature Communications*, vol. 16, no. 1, 2025. DOI: 10.1038/s41467-025-56625-z.

[132] Z. Wu, A. Dadu, M. Nalls, F. Faghri, and J. Sun, "Instruction Tuning Large Language Models to Understand Electronic Health Records," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[133] O. Ben Shoham and N. Rappoport, "CPLLM: Clinical prediction with large language models," *PLOS Digital Health*, vol. 3, no. 12, 2024. DOI: 10.1371/journal.pdig.0000680.

[134] Y. Kang, M. Yang, Y. Peng, *et al.*, "LLM-DG: Leveraging large language model for enhanced disease prediction via inter-patient and intra-patient modeling," *Information Fusion*, vol. 121, 2025. DOI: 10.1016/j.inffus.2025.103145.

[135] M. Wornow, R. Thapa, E. Steinberg, J. A. Fries, and N. H. Shah, *EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models*, Version Number: 3, 2023. DOI: 10.48550/ARXIV.2307.02028.

[136] K. Huang, J. Altosaar, and R. Ranganath, *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*, Version Number: 3, 2019. DOI: 10.48550/ARXIV.1904.05342.

[137] L. Zheng, W.-L. Chiang, Y. Sheng, *et al.*, *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*, Version Number: 4, 2023. DOI: 10.48550/ARXIV.2306.05685.

[138] H. Touvron, T. Lavril, G. Izacard, *et al.*, *LLaMA: Open and Efficient Foundation Language Models*, Version Number: 1, 2023. DOI: 10.48550/ARXIV.2302.13971.

[139] E. Bolton, A. Venigalla, M. Yasunaga, *et al.*, *BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text*, Version Number: 1, 2024. DOI: 10.48550/ARXIV.2403.18421.

[140] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, *QLoRA: Efficient Finetuning of Quantized LLMs*, Version Number: 1, 2023. DOI: 10.48550/ARXIV.2305.14314.

[141] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang, *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*, Version Number: 1, 2023. DOI: 10.48550/ARXIV.2312.12148.

[142] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, *Med-BERT: Pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction*, Version Number: 1, 2020.

[143] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, *RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism*, Version Number: 4, 2016. DOI: 10.48550/ARXIV.1608.05745.

[144] C. Sudlow, J. Gallacher, N. Allen, *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.

[145] M. Engelke, G. Baldini, J. Kleesiek, F. Nensa, and A. Dada, "FHIR-Former: Enhancing clinical predictions through Fast Healthcare Interoperability Resources and large language models," *Journal of the American Medical Informatics Association*, vol. 32, no. 12, pp. 1793–1801, 2025. DOI: 10.1093/jamia/ocaf165.

[146] U. Hemamalini, M. Dhamayanthi, G. Santhosh, C. Rajeshkumar, K. Rajakumari, and M. Sakthivanitha, "Multimodal Embedding Framework with Few-Shot Learning for Disease Prediction From Multimodal EHRs," pp. 1413–1418, 2025. DOI: 10.1109/ICIRCA65293.2025.11089852.

[147] X. Yang, L. Li, C. Wang, W. Garralda-Guillem, H. Liu, and W. Tang, "Leveraging heterogeneous tabular of EHRs with prompt learning for clinical prediction," *Journal of Biomedical Informatics*, vol. 168, 2025. DOI: 10.1016/j.jbi.2025.104868.

[148] Y. Xu, X. Jiang, X. Chu, *et al.*, "DearLLM: Enhancing Personalized Healthcare via Large Language Models-Deduced Feature Correlations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, pp. 941–949, 2025. DOI: 10.1609/aaai.v39i1.32079.

[149] A. Dada, A. Chen, C. Peng, *et al.*, *On the Impact of Cross-Domain Data on German Language Models*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2310.07321.

[150] T. GLM, : A. Zeng, *et al.*, *ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools*, Version Number: 2, 2024. DOI: 10.48550/ARXIV.2406.12793.

[151] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, *M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation*, Version Number: 5, 2024. DOI: 10.48550/ARXIV.2402.03216.

[152] Y. Zhu, W. Tang, H. Yang, *et al.*, *The Potential of LLMs in Medical Education: Generating Questions and Answers for Qualification Exams*, Version Number: 2, 2024. DOI: 10.48550/ARXIV.2410.23769.

[153] J. Chen, X. Wang, K. Ji, *et al.*, *HuatuoGPT-II, One-stage Training for Medical Adaption of LLMs*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2311.09774.

[154] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," en, *Scientific Data*, vol. 10, no. 1, p. 67, Feb. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-01960-3.

[155] DeepSeek-AI, A. Liu, B. Feng, *et al.*, *DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model*, Version Number: 5, 2024. DOI: 10.48550/ARXIV.2405.04434.

[156] B. Glicksberg, P. Timsina, D. Patel, *et al.*, "Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1921–1928, 2024. DOI: 10.1093/jamia/ocae103.

[157] S. Cohen, Z. Chen, J. Bian, C. Boucher, Y. Wu, and M. Prosperi, "Comparative Evaluation of Clinical Large Language Models and Machine Learning to Predict Antimicrobial Resistance in Hospital-Onset Sepsis," *Lecture Notes in Computer Science*, vol. 15734, pp. 65–76, 2025. DOI: 10.1007/978-3-031-95838-0_7.

[158] R. Li, X. Wang, D. Berlowitz, J. Mez, H. Lin, and H. Yu, "CARE-AD: A multi-agent large language model framework for Alzheimer's disease prediction using longitudinal clinical notes," *npj Digital Medicine*, vol. 8, no. 1, 2025. DOI: 10.1038/s41746-025-01940-4.

[159] F. Nateghi Haredasht, I. López, S. Tate, *et al.*, "Predicting treatment retention in medication for opioid use disorder: A machine learning approach using NLP and LLM-derived clinical features," *Journal of the American Medical Informatics Association*, vol. 32, no. 12, pp. 1865–1876, 2025. DOI: 10.1093/jamia/ocaf157.

[160] S. Laaouar and N. Yanes, "CLARA: A Context-Aware Drug Recommendation Framework Using Multi-Level Integration of Large Language Models and Graph Neural Networks," *29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025)*, vol. 270, pp. 4085–4094, Jan. 2025, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2025.09.533.

[161] S. Laaouar and N. Yanes, "Integrating Patient Context and Large Language Models for Graph-Based Drug Recommendation," *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, pp. 324–329, 2025. DOI: 10.18293/SEKE2025-111.

[162] Z. Wang, Y. Zhu, H. Zhao, *et al.*, "ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration," pp. 2255–2261, 2025. DOI: 10.1145/3696410.3714877.

[163] J. Lammert, T. Dreyer, S. Mathes, *et al.*, "Expert-Guided Large Language Models for Clinical Decision Support in Precision Oncology," *JCO Precision Oncology*, vol. 8, 2024. DOI: 10.1200/PO-24-00478.

[164] OpenAI, J. Achiam, S. Adler, *et al.*, *GPT-4 Technical Report*, Version Number: 6, 2023. DOI: 10.48550/ARXIV.2303.08774.

[165] T. Sounack, J. Davis, B. Durieux, *et al.*, *BioClinical ModernBERT: A State-of-the-Art Long-Context Encoder for Biomedical and Clinical NLP*, Version Number: 1, 2025. DOI: 10.48550/ARXIV.2506.10896.

[166] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," en, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

[167] X. Yang, A. Chen, N. PourNejatian, *et al.*, *GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records*, Version Number: 3, 2022. DOI: 10.48550/ARXIV.2203.03540.

[168] H. W. Chung, L. Hou, S. Longpre, *et al.*, *Scaling Instruction-Finetuned Language Models*, Version Number: 5, 2022. DOI: 10.48550/ARXIV.2210.11416.

[169] B. Baranker, A. Kapoor, and S. G. Sampath, *The STARR Protocol: An Automated LLM Methodology for Enhanced Systematic Literature Review Screening*, en, Jun. 2025. DOI: 10.22541/au.174897564.48312932/v1.

[170] R. Patel, S. N. Wee, R. Ramaswamy, *et al.*, "NeuroBlu, an electronic health record (EHR) trusted research environment (TRE) to support mental healthcare analytics with real-world data," en, *BMJ Open*, vol. 12, no. 4, e057227, Apr. 2022, ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2021-057227.

[171] S. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, Version Number: 2, 2017. DOI: 10.48550/ARXIV.1705.07874.

[172] M. Miletic and M. Sariyar, "Large Language Models for Synthetic Tabular Health Data: A Benchmark Study," *Studies in Health Technology and Informatics*, vol. 316, pp. 963–967, 2024. DOI: 10.3233/SHTI240571.

[173] G. Pamisetty, S. Batreddy, P. Verma, and C. Sobhan Babu, "EHRGPT: Leveraging language models for generating synthetic health records," pp. 4579–4587, 2024. DOI: 10.1109/BigData62323.2024.10825057.

[174] Y. Zhong, X. Wang, J. Wang, X. Zhang, and F. Ma, "MedDiTPro: A Prompt-Guided Diffusion Transformer for Multimodal Longitudinal Medical Data Synthesis," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2, pp. 4086–4097, 2025. DOI: 10.1145/3711896.3737045.

[175] R. Huang, H. Wu, Y. Yuan, *et al.*, "Evaluation and Bias Analysis of Large Language Models in Generating Synthetic Electronic Health Records: Comparative Study," *Journal of Medical Internet Research*, vol. 27, no. 1, 2025. DOI: 10.2196/65317.

[176] S. Biderman, H. Schoelkopf, Q. Anthony, *et al.*, *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*, Version Number: 2, 2023. DOI: 10.48550/ARXIV.2304.01373.

[177] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language Models are Few-Shot Learners*, Version Number: 4, 2020. DOI: 10.48550/ARXIV.2005.14165.

[178] A. Torfi and E. A. Fox, *CorGAN: Correlation-Capturing Convolutional Generative Adversarial Networks for Generating Synthetic Healthcare Records*, Version Number: 2, 2020. DOI: 10.48550/ARXIV.2001.09346.

[179] D. K. Park, S. Yoo, H. Bahng, J. Choo, and N. Park, *MEGAN: Mixture of Experts of Generative Adversarial Networks for Multimodal Image Generation*, Version Number: 2, 2018. DOI: 10.48550/ARXIV.1805.02481.

[180] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," en, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., vol. 9351, Series Title: Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 234–241, ISBN: 978-3-319-24573-7 978-3-319-24574-4. DOI: 10.1007/978-3-319-24574-4_28.

[181] 0. AI, : A. Young, *et al.*, *Yi: Open Foundation Models by 01.AI*, Version Number: 3, 2024. DOI: 10.48550/ARXIV.2403.04652.

[182] A. Yang, A. Li, B. Yang, *et al.*, *Qwen3 Technical Report*, Version Number: 1, 2025. DOI: 10.48550/ARXIV.2505.09388.

[183] A. Johnson, T. Pollard, S. Horng, L. A. Celi, and R. Mark, *MIMIC-IV-Note: Deidentified free-text clinical notes*. DOI: 10.13026/1N74-NE17. [Online]. Available: https://physionet.org/content/mimic-iv-note/2.2/ (visited on 04/01/2025).

Appendix A

Auxiliary SQL Queries for the METRE Adaptation

This appendix presents the auxiliary SQL queries used to reconstruct intermediate tables required by the adapted METRE extraction pipeline for MIMIC-IV v3.1.

A.1 Culture Table

```sql
CREATE OR REPLACE TABLE `project.dataset.culture` AS
SELECT
    m.subject_id,
    m.hadm_id,
    m.charttime,
    m.chartdate,
    m.spec_type_desc AS specimen,
    CAST(NULL AS STRING) AS screen,
    CASE
        WHEN m.org_name IS NOT NULL THEN 1
        ELSE 0
    END AS positive_culture,
    CASE
        WHEN m.ab_name IS NOT NULL THEN 1
        ELSE 0
    END AS has_sensitivity
FROM `physionet-data.mimiciv_3_1_hosp.microbiologyevents` m;
```

Listing A.1: SQL query used to reconstruct the culture table.

A.2 Heparin Table

```sql
CREATE OR REPLACE TABLE `project.dataset.heparin` AS
SELECT
    ie.subject_id,
    ie.hadm_id,
    ie.stay_id,
    ie.starttime,
    ie.endtime,
    ie.itemid,
    di.label,
    ie.amount,
    ie.amountuom,
    ie.rate,
    ie.rateuom
FROM `physionet-data.mimiciv_3_1_icu.inputevents` ie
INNER JOIN `physionet-data.mimiciv_3_1_icu.d_items` di
    ON ie.itemid = di.itemid
WHERE ie.itemid IN (
    225152,
    225975,
    229597,
    230044
);
```

Listing A.2: SQL query used to reconstruct the heparin table.

Appendix B

Additional Results

This appendix presents the supplementary internal validation and test metrics supporting the results discussed in Chapter 5.

B.1 Additional Internal Validation Metrics

Table B.1 reports the remaining internal validation metrics for the same model configurations. The table includes accuracy, precision, sensitivity, specificity, and F1-score for models trained on MIMIC-IV and eICU using default and Optuna-selected hyperparameters, , complementing the main results presented in Section 5.2. Values are reported as mean ± standard deviation, in percentage points.

Table B.1: Supplementary internal validation metrics for models trained on MIMIC-IV and eICU. Values are reported as mean ± standard deviation.

Train database	Config	Model	Task	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
MIMIC-IV	Default	LR	Mortality 48h	82,52 ± 0,83	37,33 ± 1,27	72,43 ± 1,59	83,85 ± 1,01	49,25 ± 1,14
			ARF 4h	69,33 ± 1,26	44,25 ± 1,68	64,03 ± 3,60	71,22 ± 1,21	52,32 ± 2,27
			ARF 12h	70,71 ± 1,16	32,15 ± 1,43	63,14 ± 3,12	72,28 ± 1,28	42,59 ± 1,82
			Shock 4h	73,06 ± 0,85	14,65 ± 0,65	65,72 ± 2,90	73,57 ± 0,92	23,96 ± 1,03
			Shock 12h	79,13 ± 0,65	12,20 ± 0,68	60,86 ± 3,72	79,97 ± 0,69	20,32 ± 1,12
		RF	Mortality 48h	89,31 ± 0,59	57,54 ± 4,22	32,85 ± 3,98	96,79 ± 0,41	41,75 ± 4,04
			ARF 4h	80,41 ± 1,43	69,88 ± 3,55	44,73 ± 4,49	93,15 ± 0,75	54,48 ± 4,28
			ARF 12h	85,71 ± 0,57	72,90 ± 4,63	27,12 ± 2,84	97,89 ± 0,43	39,46 ± 3,31
			Shock 4h	93,64 ± 0,37	51,05 ± 3,93	36,39 ± 3,41	97,59 ± 0,28	42,45 ± 3,49
			Shock 12h	95,98 ± 0,21	57,57 ± 4,72	31,60 ± 3,99	98,92 ± 0,24	40,60 ± 3,66
		NB	Mortality 48h	64,59 ± 2,19	20,78 ± 1,43	71,70 ± 3,09	63,65 ± 2,30	32,21 ± 1,96
			ARF 4h	77,66 ± 1,25	64,28 ± 4,79	34,14 ± 2,99	93,19 ± 1,19	44,54 ± 3,33
			ARF 12h	79,87 ± 0,99	41,38 ± 2,59	39,74 ± 2,18	88,22 ± 1,38	40,47 ± 1,61
			Shock 4h	46,47 ± 2,86	9,61 ± 0,28	86,67 ± 3,58	43,70 ± 3,26	17,30 ± 0,45
			Shock 12h	10,54 ± 0,90	4,39 ± 0,15	93,71 ± 2,67	6,73 ± 0,88	8,39 ± 0,28

Continued on next page

Train database	Config	Model	Task	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
		SVM	Mortality 48h	90,07 ± 0,53	63,78 ± 4,17	35,08 ± 2,70	97,35 ± 0,38	45,23 ± 3,08
			ARF 4h	77,57 ± 1,09	65,79 ± 4,04	30,68 ± 3,15	94,31 ± 0,80	41,79 ± 3,51
			ARF 12h	83,74 ± 0,50	62,20 ± 6,20	13,99 ± 2,50	98,25 ± 0,31	22,78 ± 3,62
			Shock 4h	93,52 ± 0,13	47,14 ± 10,89	4,94 ± 1,56	99,63 ± 0,07	8,94 ± 2,75
			Shock 12h	95,71 ± 0,13	60,52 ± 14,26	6,71 ± 1,31	99,78 ± 0,11	12,02 ± 2,26
		RNN	Mortality 48h	80,93 ± 2,71	36,07 ± 3,44	77,84 ± 3,91	81,34 ± 3,52	49,08 ± 2,58
			ARF 4h	74,22 ± 1,82	51,13 ± 3,19	58,17 ± 4,37	79,96 ± 3,15	54,26 ± 2,26
			ARF 12h	73,09 ± 4,02	34,73 ± 3,53	60,20 ± 6,06	75,77 ± 6,01	43,65 ± 1,75
			Shock 4h	85,89 ± 5,30	25,79 ± 5,92	50,11 ± 16,15	88,35 ± 6,75	31,82 ± 2,06
			Shock 12h	88,36 ± 5,01	22,07 ± 6,96	48,11 ± 16,68	90,20 ± 5,96	27,39 ± 2,79
		TCN	Mortality 48h	79,74 ± 2,54	34,74 ± 3,24	80,25 ± 4,37	79,68 ± 3,34	48,29 ± 2,43
			ARF 4h	73,45 ± 2,51	50,14 ± 4,19	59,23 ± 5,05	78,53 ± 4,60	54,01 ± 2,24
			ARF 12h	73,81 ± 5,49	35,90 ± 5,30	59,08 ± 6,45	76,87 ± 7,81	44,09 ± 3,14
			Shock 4h	79,80 ± 1,92	18,85 ± 1,10	63,94 ± 5,79	80,90 ± 2,39	29,04 ± 1,22
			Shock 12h	84,05 ± 4,75	16,55 ± 5,03	56,78 ± 11,85	85,30 ± 5,46	24,39 ± 2,76
		Trans	Mortality 48h	72,16 ± 7,13	28,82 ± 4,96	84,90 ± 8,26	70,47 ± 9,14	42,44 ± 4,27
			ARF 4h	72,37 ± 3,85	49,11 ± 6,02	60,00 ± 8,34	76,79 ± 7,25	53,30 ± 3,40
			ARF 12h	74,26 ± 5,87	36,78 ± 6,32	57,84 ± 10,76	77,67 ± 9,15	43,85 ± 2,91
			Shock 4h	68,96 ± 6,86	14,88 ± 2,05	77,33 ± 8,52	68,38 ± 7,81	24,76 ± 2,55
			Shock 12h	71,03 ± 3,98	11,21 ± 1,16	79,89 ± 3,78	70,63 ± 4,28	19,62 ± 1,75
	Optuna	LR	Mortality 48h	81,80 ± 0,79	36,54 ± 1,05	75,07 ± 1,69	82,69 ± 1,02	49,13 ± 0,90
			ARF 4h	69,03 ± 0,97	43,91 ± 1,23	64,03 ± 2,80	70,81 ± 1,15	52,08 ± 1,66
			ARF 12h	70,28 ± 1,57	31,72 ± 1,92	62,88 ± 3,60	71,82 ± 1,62	42,15 ± 2,38
			Shock 4h	70,57 ± 0,60	14,00 ± 0,70	69,28 ± 4,48	70,66 ± 0,72	23,29 ± 1,21
			Shock 12h	76,44 ± 0,73	11,53 ± 0,48	65,72 ± 3,01	76,93 ± 0,80	19,61 ± 0,80
		RF	Mortality 48h	89,32 ± 0,64	57,85 ± 5,12	32,03 ± 3,60	96,91 ± 0,44	41,18 ± 4,05
			ARF 4h	80,73 ± 1,17	74,56 ± 3,67	40,58 ± 3,13	95,07 ± 0,72	52,52 ± 3,41
			ARF 12h	85,74 ± 0,46	81,59 ± 4,34	22,29 ± 3,13	98,93 ± 0,38	34,85 ± 3,90
			Shock 4h	93,14 ± 0,35	46,41 ± 3,17	40,56 ± 3,74	96,77 ± 0,24	43,26 ± 3,37
			Shock 12h	95,88 ± 0,18	55,03 ± 3,48	32,78 ± 3,69	98,77 ± 0,20	40,96 ± 3,34
		NB	Mortality 48h	73,05 ± 1,15	24,28 ± 1,56	61,51 ± 4,53	74,58 ± 1,17	34,81 ± 2,25
			ARF 4h	77,19 ± 1,30	61,75 ± 4,67	35,21 ± 3,10	92,17 ± 1,29	44,78 ± 3,34
			ARF 12h	80,27 ± 0,94	42,18 ± 2,56	38,63 ± 2,03	88,93 ± 1,27	40,26 ± 1,61
			Shock 4h	64,41 ± 1,34	11,39 ± 0,67	66,56 ± 4,34	64,26 ± 1,48	19,45 ± 1,14
			Shock 12h	41,86 ± 4,52	6,26 ± 0,45	87,43 ± 2,30	39,78 ± 4,72	11,67 ± 0,79
		SVM	Mortality 48h	90,11 ± 0,54	64,59 ± 4,44	34,35 ± 2,82	97,50 ± 0,39	44,81 ± 3,22
			ARF 4h	77,83 ± 0,89	66,59 ± 3,17	31,44 ± 2,65	94,39 ± 0,47	42,69 ± 3,02
			ARF 12h	83,97 ± 0,61	63,59 ± 5,70	15,62 ± 2,93	98,18 ± 0,19	25,04 ± 4,20
			Shock 4h	93,51 ± 0,14	47,73 ± 19,52	2,72 ± 1,15	99,78 ± 0,10	5,13 ± 2,14
			Shock 12h	95,65 ± 0,09	48,92 ± 21,93	3,10 ± 1,68	99,88 ± 0,05	5,82 ± 3,11
		RNN	Mortality 48h	76,97 ± 1,59	32,21 ± 1,49	87,04 ± 1,73	75,64 ± 1,94	46,99 ± 1,51
			ARF 4h	73,65 ± 2,11	50,18 ± 3,14	62,80 ± 4,15	77,52 ± 3,49	55,65 ± 2,34
			ARF 12h	75,25 ± 2,64	37,12 ± 2,53	60,59 ± 6,34	78,30 ± 4,45	45,74 ± 0,95
			Shock 4h	72,32 ± 2,94	16,25 ± 1,33	78,33 ± 4,63	71,91 ± 3,29	26,87 ± 1,85
			Shock 12h	81,68 ± 7,22	16,02 ± 3,88	65,89 ± 10,26	82,40 ± 7,98	25,21 ± 4,29

Continued on next page

Train database	Config	Model	Task	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
		TCN	Mortality 48h	81,35 ± 1,96	36,67 ± 2,18	79,75 ± 4,49	81,56 ± 2,71	50,11 ± 1,60
			ARF 4h	75,32 ± 3,38	53,88 ± 5,68	57,14 ± 6,15	81,80 ± 5,97	54,95 ± 2,99
			ARF 12h	75,93 ± 3,76	38,24 ± 4,64	58,63 ± 7,11	79,52 ± 5,75	45,72 ± 2,58
			Shock 4h	80,17 ± 5,88	20,85 ± 4,33	66,89 ± 10,49	81,08 ± 6,92	31,15 ± 4,13
			Shock 12h	80,82 ± 5,26	14,52 ± 2,39	64,88 ± 9,41	81,55 ± 5,91	23,40 ± 2,50
		Trans	Mortality 48h	81,02 ± 1,84	36,20 ± 2,24	79,57 ± 5,11	81,21 ± 2,70	49,60 ± 1,25
			ARF 4h	75,58 ± 3,26	54,46 ± 6,15	58,12 ± 6,68	81,80 ± 6,09	55,62 ± 2,84
			ARF 12h	69,29 ± 5,54	32,52 ± 3,74	68,04 ± 8,70	69,55 ± 8,33	43,47 ± 2,05
			Shock 4h	69,78 ± 4,50	15,52 ± 1,67	81,22 ± 5,28	68,99 ± 5,09	25,99 ± 2,21
			Shock 12h	76,06 ± 2,27	12,36 ± 0,93	73,17 ± 7,34	76,20 ± 2,59	21,11 ± 1,51
eICU	Default	LR	Mortality 48h	80,34 ± 0,70	32,22 ± 1,02	71,31 ± 1,54	81,45 ± 0,75	44,37 ± 1,14
			ARF 4h	77,44 ± 0,52	6,33 ± 0,16	66,30 ± 2,08	77,70 ± 0,56	11,55 ± 0,29
			ARF 12h	80,78 ± 0,53	5,10 ± 0,25	62,99 ± 3,01	81,07 ± 0,54	9,44 ± 0,47
			Shock 4h	73,68 ± 0,37	10,40 ± 0,46	64,31 ± 2,83	74,11 ± 0,33	17,90 ± 0,78
			Shock 12h	76,23 ± 0,43	7,82 ± 0,41	60,48 ± 3,60	76,75 ± 0,46	13,84 ± 0,73
		RF	Mortality 48h	89,92 ± 0,31	55,72 ± 2,03	40,41 ± 2,35	96,03 ± 0,28	46,81 ± 2,03
			ARF 4h	95,91 ± 0,19	26,17 ± 1,76	46,00 ± 3,51	97,05 ± 0,20	33,34 ± 2,15
			ARF 12h	96,65 ± 0,26	20,84 ± 2,35	39,24 ± 4,24	97,58 ± 0,26	27,17 ± 2,77
			Shock 4h	94,85 ± 0,23	37,46 ± 3,69	22,92 ± 2,04	98,21 ± 0,18	28,41 ± 2,52
			Shock 12h	96,30 ± 0,08	34,41 ± 2,07	18,88 ± 2,31	98,83 ± 0,12	24,33 ± 2,31
		NB	Mortality 48h	59,10 ± 1,62	18,09 ± 0,57	77,02 ± 3,00	56,89 ± 2,04	29,29 ± 0,83
			ARF 4h	4,21 ± 0,44	2,24 ± 0,02	98,52 ± 0,97	2,07 ± 0,46	4,37 ± 0,03
			ARF 12h	3,98 ± 0,38	1,59 ± 0,02	97,63 ± 1,32	2,47 ± 0,39	3,13 ± 0,04
			Shock 4h	4,96 ± 0,10	4,48 ± 0,02	99,80 ± 0,25	0,53 ± 0,10	8,57 ± 0,03
			Shock 12h	3,95 ± 0,08	3,18 ± 0,01	99,74 ± 0,29	0,83 ± 0,09	6,16 ± 0,02
		SVM	Mortality 48h	86,64 ± 0,34	42,60 ± 1,00	61,88 ± 1,77	89,70 ± 0,38	50,45 ± 1,12
			ARF 4h	95,93 ± 0,12	23,22 ± 1,85	36,24 ± 4,09	97,28 ± 0,13	28,29 ± 2,56
			ARF 12h	97,88 ± 0,10	28,50 ± 3,03	22,03 ± 2,68	99,10 ± 0,10	24,80 ± 2,65
			Shock 4h	89,45 ± 0,24	17,64 ± 0,84	37,17 ± 1,87	91,89 ± 0,23	23,92 ± 1,12
			Shock 12h	94,59 ± 0,21	19,53 ± 2,00	22,87 ± 2,69	96,93 ± 0,21	21,04 ± 2,22
		RNN	Mortality 48h	78,62 ± 3,08	30,98 ± 2,94	74,22 ± 4,45	79,16 ± 3,96	43,51 ± 2,33
			ARF 4h	96,00 ± 1,56	25,80 ± 5,12	34,75 ± 8,95	97,39 ± 1,78	28,43 ± 2,99
			ARF 12h	94,54 ± 3,31	15,20 ± 4,23	41,02 ± 12,79	95,41 ± 3,56	21,06 ± 4,19
			Shock 4h	82,46 ± 5,96	14,49 ± 3,71	51,42 ± 16,95	83,90 ± 7,01	20,76 ± 1,01
			Shock 12h	84,93 ± 6,45	11,70 ± 3,67	47,89 ± 12,46	86,14 ± 7,05	17,77 ± 3,00
		TCN	Mortality 48h	77,09 ± 4,28	29,96 ± 3,92	76,23 ± 5,63	77,20 ± 5,47	42,66 ± 2,99
			ARF 4h	90,96 ± 2,06	14,18 ± 2,61	57,21 ± 6,75	91,73 ± 2,24	22,42 ± 2,58
			ARF 12h	95,27 ± 1,02	15,06 ± 2,70	40,27 ± 7,26	96,16 ± 1,12	21,55 ± 2,45
			Shock 4h	77,96 ± 4,91	12,03 ± 1,83	59,58 ± 6,74	78,81 ± 5,43	19,83 ± 2,12
			Shock 12h	86,87 ± 5,47	12,38 ± 3,21	43,98 ± 12,01	88,27 ± 6,03	18,36 ± 2,50
		Trans	Mortality 48h	71,63 ± 4,37	25,42 ± 2,60	79,20 ± 5,14	70,70 ± 5,49	38,30 ± 2,40
			ARF 4h	79,09 ± 3,55	8,34 ± 0,99	82,28 ± 5,11	79,01 ± 3,72	15,10 ± 1,58
			ARF 12h	76,09 ± 5,97	5,58 ± 1,10	83,61 ± 5,05	75,97 ± 6,14	10,43 ± 1,90
			Shock 4h	71,90 ± 3,14	10,66 ± 0,83	71,08 ± 4,24	71,94 ± 3,45	18,51 ± 1,17
			Shock 12h	68,76 ± 5,74	7,26 ± 0,97	73,52 ± 5,12	68,60 ± 6,05	13,18 ± 1,58

Continued on next page

Train database	Config	Model	Task	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
	Optuna	LR	Mortality 48h	80,08 ± 0,75	32,21 ± 0,97	73,38 ± 1,70	80,90 ± 0,87	44,76 ± 1,06
			ARF 4h	76,68 ± 0,49	6,23 ± 0,14	67,51 ± 1,38	76,89 ± 0,52	11,40 ± 0,24
			ARF 12h	78,82 ± 0,61	4,94 ± 0,25	67,44 ± 3,29	79,01 ± 0,62	9,20 ± 0,45
			Shock 4h	73,40 ± 0,36	10,32 ± 0,39	64,51 ± 2,70	73,81 ± 0,37	17,79 ± 0,68
			Shock 12h	74,98 ± 0,37	7,72 ± 0,39	63,19 ± 3,48	75,37 ± 0,39	13,76 ± 0,70
		RF	Mortality 48h	90,46 ± 0,34	61,08 ± 2,88	36,73 ± 2,13	97,10 ± 0,32	45,83 ± 2,08
			ARF 4h	97,16 ± 0,12	35,83 ± 2,31	35,08 ± 3,76	98,57 ± 0,15	35,37 ± 2,68
			ARF 12h	81,97 ± 0,67	6,73 ± 0,40	80,27 ± 3,04	82,00 ± 0,66	12,42 ± 0,71
			Shock 4h	95,31 ± 0,10	43,78 ± 2,96	18,32 ± 2,06	98,90 ± 0,08	25,80 ± 2,48
			Shock 12h	96,59 ± 0,15	40,91 ± 4,77	16,83 ± 1,69	99,19 ± 0,16	23,76 ± 2,04
		NB	Mortality 48h	74,67 ± 0,60	23,72 ± 1,15	58,94 ± 3,64	76,61 ± 0,56	33,83 ± 1,73
			ARF 4h	19,54 ± 2,41	2,60 ± 0,08	96,36 ± 1,09	17,80 ± 2,46	5,06 ± 0,16
			ARF 12h	14,26 ± 2,67	1,75 ± 0,08	96,14 ± 1,96	12,93 ± 2,70	3,45 ± 0,15
			Shock 4h	15,69 ± 4,94	4,90 ± 0,22	97,04 ± 1,72	11,89 ± 5,23	9,33 ± 0,39
			Shock 12h	14,70 ± 3,29	3,47 ± 0,10	96,89 ± 1,55	12,02 ± 3,43	6,70 ± 0,18
		SVM	Mortality 48h	81,41 ± 0,58	34,13 ± 0,81	74,18 ± 1,47	82,31 ± 0,69	46,75 ± 0,85
			ARF 4h	84,91 ± 0,38	9,55 ± 0,33	68,36 ± 2,45	85,29 ± 0,41	16,76 ± 0,56
			ARF 12h	88,72 ± 0,45	8,06 ± 0,36	58,46 ± 1,74	89,21 ± 0,45	14,16 ± 0,59
			Shock 4h	77,27 ± 0,38	12,00 ± 0,43	64,64 ± 2,56	77,86 ± 0,39	20,24 ± 0,74
			Shock 12h	83,60 ± 0,40	10,86 ± 0,57	58,10 ± 2,91	84,43 ± 0,39	18,29 ± 0,94
		RNN	Mortality 48h	74,49 ± 1,58	27,61 ± 1,24	80,94 ± 1,76	73,70 ± 1,89	41,15 ± 1,32
			ARF 4h	87,56 ± 4,29	13,02 ± 2,78	73,27 ± 9,33	87,88 ± 4,59	21,79 ± 3,63
			ARF 12h	89,99 ± 4,07	10,42 ± 3,10	59,44 ± 10,96	90,48 ± 4,31	17,22 ± 3,80
			Shock 4h	66,74 ± 3,75	9,74 ± 0,72	77,31 ± 3,42	66,24 ± 4,05	17,28 ± 1,10
			Shock 12h	74,82 ± 10,04	9,11 ± 3,84	61,87 ± 16,43	75,24 ± 10,87	14,69 ± 3,47
		TCN	Mortality 48h	77,76 ± 1,08	30,07 ± 0,98	76,87 ± 2,50	77,87 ± 1,43	43,20 ± 0,96
			ARF 4h	88,62 ± 5,00	13,57 ± 4,18	63,54 ± 11,10	89,19 ± 5,35	21,64 ± 4,83
			ARF 12h	91,94 ± 5,77	15,50 ± 8,47	45,52 ± 18,71	92,69 ± 6,15	19,16 ± 6,20
			Shock 4h	74,83 ± 3,84	11,69 ± 0,93	69,70 ± 7,64	75,07 ± 4,36	19,94 ± 1,02
			Shock 12h	78,42 ± 2,67	9,17 ± 0,70	64,80 ± 5,41	78,86 ± 2,91	16,03 ± 0,96
		Trans	Mortality 48h	81,91 ± 1,24	35,14 ± 1,59	75,42 ± 3,13	82,72 ± 1,72	47,87 ± 1,07
			ARF 4h	83,06 ± 2,52	9,82 ± 0,97	79,64 ± 4,38	83,13 ± 2,65	17,45 ± 1,50
			ARF 12h	93,08 ± 1,95	14,31 ± 2,90	61,95 ± 7,10	93,58 ± 2,09	22,91 ± 3,30
			Shock 4h	75,15 ± 1,69	11,85 ± 0,60	70,69 ± 2,83	75,36 ± 1,85	20,28 ± 0,87
			Shock 12h	81,49 ± 3,38	10,36 ± 1,32	61,73 ± 4,10	82,14 ± 3,60	17,67 ± 1,86

B.2 Additional Test Metrics for Models Trained on MIMIC

Table B.2 reports the supplementary test metrics for the models trained on MIMIC-IV. The table includes accuracy, precision, sensitivity, specificity, and F1-score for the internal MIMIC-IV test set and the external eICU test set, complementing the main results presented in Section 5.3. Values are reported as mean ± standard deviation, with the 95% confidence interval shown in parentheses below.

Table B.2: Supplementary test metrics for models trained on MIMIC-IV and evaluated on the internal MIMIC-IV test set and the external eICU test set. Values are reported as mean ± standard deviation, with the 95% confidence interval shown below.

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
MIMIC-IV	Mortality 48h	LR	82,61 ± 1,21 (80,30, 85,00)	35,85 ± 3,16 (29,83, 42,26)	75,31 ± 4,11 (67,24, 83,04)	83,51 ± 1,25 (81,17, 86,06)	48,50 ± 3,32 (41,80, 55,10)
		RF	90,09 ± 0,96 (88,20, 91,90)	58,55 ± 6,53 (45,45, 70,70)	31,12 ± 4,51 (22,33, 40,18)	97,31 ± 0,53 (96,20, 98,22)	40,50 ± 4,90 (29,94, 49,69)
		NB	71,58 ± 1,48 (68,50, 74,40)	21,62 ± 2,31 (17,10, 26,35)	61,17 ± 4,65 (51,85, 69,70)	72,86 ± 1,53 (69,89, 75,70)	31,89 ± 2,90 (26,21, 37,95)
		SVM	90,31 ± 0,95 (88,40, 92,10)	60,98 ± 6,83 (47,54, 74,00)	30,84 ± 4,44 (22,14, 39,25)	97,58 ± 0,52 (96,55, 98,54)	40,82 ± 4,90 (31,08, 49,69)
		RNN	77,64 ± 1,33 (74,80, 80,10)	30,91 ± 2,69 (25,52, 36,05)	85,02 ± 3,38 (78,12, 92,06)	76,74 ± 1,44 (73,71, 79,50)	45,27 ± 3,06 (38,96, 51,00)
		TCN	82,69 ± 1,20 (80,30, 85,10)	36,33 ± 3,13 (30,30, 43,10)	77,99 ± 3,98 (70,19, 85,59)	83,27 ± 1,25 (80,66, 85,65)	49,49 ± 3,27 (42,69, 56,21)
		Trans	82,53 ± 1,20 (80,00, 84,90)	36,33 ± 3,08 (30,69, 42,68)	80,04 ± 3,69 (72,83, 87,23)	82,83 ± 1,28 (80,24, 85,31)	49,91 ± 3,20 (43,80, 56,43)
		EMERGE TS-only	91,24 ± 0,90 (89,50, 92,90)	66,20 ± 5,85 (54,83, 77,27)	40,22 ± 4,68 (31,35, 49,52)	97,49 ± 0,52 (96,36, 98,43)	49,89 ± 4,59 (40,70, 58,65)
	ARF 4h	LR	68,60 ± 1,45 (65,80, 71,40)	43,44 ± 2,50 (38,54, 48,16)	64,91 ± 3,00 (58,97, 70,29)	69,92 ± 1,69 (66,80, 73,20)	52,01 ± 2,37 (47,26, 56,45)
		RF	80,93 ± 1,25 (78,40, 83,30)	74,15 ± 3,49 (67,10, 80,43)	41,99 ± 3,05 (35,82, 47,92)	94,79 ± 0,81 (93,10, 96,27)	53,56 ± 2,97 (47,57, 59,09)
		NB	78,59 ± 1,29 (76,00, 81,10)	66,84 ± 3,81 (59,15, 74,39)	36,53 ± 2,98 (30,90, 42,38)	93,56 ± 0,86 (91,72, 95,19)	47,18 ± 3,06 (41,01, 53,21)
		SVM	77,50 ± 1,32 (74,90, 79,90)	65,10 ± 4,34 (56,30, 73,21)	30,82 ± 2,80 (25,11, 36,37)	94,11 ± 0,90 (92,28, 95,76)	41,77 ± 3,10 (35,48, 47,94)
		RNN	72,69 ± 1,40 (70,10, 75,50)	48,44 ± 2,70 (42,99, 53,62)	62,84 ± 3,10 (56,77, 69,08)	76,20 ± 1,57 (73,24, 79,06)	54,66 ± 2,47 (49,73, 59,43)
		TCN	75,34 ± 1,36 (72,60, 78,00)	52,80 ± 2,98 (46,69, 58,93)	56,88 ± 3,17 (50,73, 62,78)	81,90 ± 1,42 (79,11, 84,65)	54,72 ± 2,61 (49,30, 59,65)
		Trans	75,14 ± 1,33 (72,60, 77,90)	52,38 ± 2,85 (46,57, 58,18)	58,41 ± 3,02 (52,65, 64,38)	81,10 ± 1,46 (78,20, 83,95)	55,19 ± 2,48 (50,09, 60,31)
		EMERGE TS-only	79,64 ± 1,28 (77,10, 82,20)	69,36 ± 3,78 (61,68, 76,64)	40,18 ± 3,00 (34,37, 45,83)	93,68 ± 0,92 (91,80, 95,38)	50,83 ± 2,98 (44,99, 56,72)
	ARF 12h	LR	69,64 ± 1,46 (66,80, 72,40)	29,91 ± 2,41 (25,29, 34,63)	61,84 ± 3,74 (55,03, 69,43)	71,19 ± 1,55 (68,15, 74,03)	40,28 ± 2,68 (35,01, 45,50)
		RF	86,77 ± 1,03 (84,80, 88,80)	86,95 ± 5,16 (76,59, 96,15)	23,78 ± 3,24 (17,65, 30,07)	99,29 ± 0,29 (98,69, 99,76)	37,25 ± 4,18 (29,19, 45,23)
		NB	80,68 ± 1,24 (78,30, 83,10)	41,36 ± 3,85 (33,75, 49,14)	39,41 ± 3,78 (31,79, 47,06)	88,89 ± 1,07 (86,69, 90,92)	40,29 ± 3,41 (33,56, 46,65)
		SVM	84,16 ± 1,12 (82,00, 86,30)	60,34 ± 8,36 (42,86, 75,76)	13,08 ± 2,65 (7,78, 18,50)	98,29 ± 0,46 (97,35, 99,15)	21,41 ± 3,90 (13,46, 28,96)
		RNN	75,58 ± 1,38 (73,10, 78,40)	35,53 ± 2,87 (29,73, 41,01)	57,97 ± 3,76 (50,83, 65,42)	79,09 ± 1,41 (76,50, 81,96)	44,00 ± 2,92 (38,18, 49,66)
		TCN	76,20 ± 1,32 (73,60, 78,90)	35,67 ± 2,95 (29,63, 41,50)	54,10 ± 3,81 (46,90, 61,28)	80,60 ± 1,33 (77,95, 83,17)	42,93 ± 2,96 (37,09, 48,69)
		Trans	70,28 ± 1,41 (67,40, 73,00)	30,99 ± 2,46 (26,19, 35,82)	64,59 ± 3,68 (57,46, 71,68)	71,41 ± 1,51 (68,32, 74,45)	41,84 ± 2,73 (36,22, 47,51)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	85,88 ± 1,03 (83,90, 87,90)	65,47 ± 5,34 (55,26, 75,33)	31,52 ± 3,53 (24,65, 38,61)	96,69 ± 0,63 (95,40, 97,87)	42,45 ± 3,85 (34,58, 49,80)
	Shock 4h	LR	70,20 ± 1,46 (67,30, 73,10)	12,25 ± 1,85 (8,78, 16,04)	67,25 ± 6,13 (55,07, 79,31)	70,38 ± 1,49 (67,48, 73,46)	20,69 ± 2,82 (15,38, 26,36)
		RF	93,48 ± 0,76 (92,00, 94,80)	43,39 ± 6,77 (29,82, 56,46)	41,35 ± 6,51 (28,85, 54,10)	96,68 ± 0,57 (95,60, 97,77)	42,12 ± 5,87 (30,23, 53,04)
		NB	63,55 ± 1,48 (60,70, 66,50)	10,34 ± 1,56 (7,20, 13,35)	68,98 ± 6,04 (57,14, 80,44)	63,21 ± 1,52 (60,30, 66,28)	17,95 ± 2,48 (12,84, 22,69)
		SVM	94,27 ± 0,73 (92,80, 95,70)	57,19 ± 27,02 (0,00, 100,00)	4,44 ± 2,70 (0,00, 10,34)	99,80 ± 0,15 (99,47, 100,00)	8,13 ± 4,75 (0,00, 18,18)
		RNN	73,29 ± 1,36 (70,60, 76,00)	14,90 ± 2,04 (10,89, 18,75)	76,69 ± 5,63 (65,15, 87,04)	73,08 ± 1,40 (70,34, 75,95)	24,90 ± 3,01 (18,94, 30,29)
		TCN	84,45 ± 1,12 (82,20, 86,50)	21,89 ± 3,06 (16,06, 28,05)	65,65 ± 6,09 (53,45, 78,26)	85,61 ± 1,10 (83,53, 87,67)	32,74 ± 3,89 (25,00, 40,51)
		Trans	70,33 ± 1,45 (67,30, 73,20)	13,86 ± 1,92 (10,24, 17,61)	79,04 ± 5,47 (67,92, 89,47)	69,79 ± 1,51 (66,63, 72,76)	23,54 ± 2,90 (17,93, 29,14)
		EMERGE TS-only	94,54 ± 0,71 (93,20, 95,90)	61,37 ± 13,25 (33,33, 86,69)	15,39 ± 4,67 (6,90, 24,59)	99,41 ± 0,25 (98,84, 99,79)	24,36 ± 6,62 (11,63, 36,93)
	Shock 12h	LR	76,34 ± 1,35 (73,70, 79,00)	9,79 ± 1,93 (6,28, 13,78)	64,53 ± 7,85 (48,48, 80,65)	76,81 ± 1,37 (74,10, 79,37)	16,94 ± 3,05 (11,32, 22,88)
		RF	96,43 ± 0,60 (95,20, 97,50)	53,45 ± 9,98 (32,14, 72,74)	37,61 ± 7,77 (22,72, 53,33)	98,72 ± 0,36 (97,93, 99,38)	43,78 ± 7,76 (27,58, 58,18)
		NB	41,76 ± 1,62 (38,70, 44,90)	5,48 ± 0,94 (3,72, 7,41)	89,37 ± 5,18 (77,50, 97,62)	39,91 ± 1,65 (36,77, 43,18)	10,31 ± 1,68 (7,12, 13,71)
		SVM	96,24 ± 0,61 (95,00, 97,40)	45,39 ± 32,86 (0,00, 100,00)	4,08 ± 3,40 (0,00, 11,63)	99,83 ± 0,13 (99,58, 100,00)	7,33 ± 5,88 (0,00, 20,01)
		RNN	86,91 ± 1,08 (84,80, 89,00)	16,66 ± 3,21 (10,61, 23,36)	62,16 ± 8,14 (45,45, 77,78)	87,88 ± 1,06 (85,76, 89,82)	26,17 ± 4,43 (17,61, 34,97)
		TCN	85,31 ± 1,12 (83,10, 87,40)	15,73 ± 2,91 (10,26, 21,92)	66,88 ± 7,47 (51,52, 81,25)	86,03 ± 1,13 (83,80, 88,10)	25,36 ± 4,08 (17,30, 33,51)
		Trans	76,37 ± 1,38 (73,80, 78,90)	10,88 ± 1,95 (7,00, 14,90)	73,67 ± 7,36 (58,33, 87,51)	76,48 ± 1,41 (73,75, 79,21)	18,90 ± 3,07 (12,77, 25,16)
		EMERGE TS-only	96,49 ± 0,59 (95,30, 97,60)	62,08 ± 16,22 (28,57, 91,73)	16,70 ± 6,27 (5,56, 30,31)	99,61 ± 0,20 (99,17, 99,90)	25,91 ± 8,61 (9,52, 43,48)
eICU	Mortality 48h	LR	75,63 ± 1,39 (73,00, 78,40)	24,66 ± 2,74 (19,48, 29,97)	59,77 ± 4,90 (49,99, 69,29)	77,58 ± 1,42 (74,86, 80,36)	34,85 ± 3,28 (28,65, 40,91)
		RF	88,71 ± 1,02 (86,80, 90,60)	42,63 ± 10,43 (22,57, 64,29)	9,30 ± 2,78 (4,30, 15,18)	98,46 ± 0,41 (97,65, 99,22)	15,16 ± 4,22 (7,35, 23,88)
		NB	81,91 ± 1,24 (79,30, 84,30)	13,27 ± 3,48 (6,67, 20,23)	11,84 ± 3,18 (5,82, 18,35)	90,52 ± 0,94 (88,59, 92,33)	12,46 ± 3,20 (6,25, 18,88)
		SVM	89,08 ± 1,01 (87,10, 91,00)	50,35 ± 10,35 (30,77, 70,59)	11,23 ± 3,03 (5,98, 17,70)	98,64 ± 0,39 (97,83, 99,43)	18,24 ± 4,49 (10,08, 27,40)
		RNN	67,53 ± 1,52 (64,50, 70,50)	20,80 ± 2,21 (16,62, 25,24)	70,17 ± 4,44 (61,16, 78,81)	67,20 ± 1,59 (64,28, 70,21)	32,04 ± 2,89 (26,58, 37,89)
		TCN	71,34 ± 1,44 (68,70, 74,00)	22,82 ± 2,42 (18,10, 27,49)	68,06 ± 4,52 (59,16, 76,58)	71,74 ± 1,52 (68,78, 74,69)	34,13 ± 3,04 (28,18, 40,00)
		Trans	79,58 ± 1,29 (76,90, 82,00)	28,27 ± 3,15 (22,32, 34,54)	56,43 ± 4,76 (47,27, 65,31)	82,42 ± 1,28 (79,87, 84,86)	37,60 ± 3,52 (30,87, 44,66)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	86,12 ± 1,10 (83,90, 88,11)	33,28 ± 4,88 (23,96, 43,53)	26,88 ± 4,16 (19,15, 35,09)	93,39 ± 0,81 (91,81, 94,92)	29,63 ± 4,11 (21,98, 37,81)
	ARF 4h	LR	40,73 ± 1,56 (37,70, 43,80)	2,04 ± 0,59 (0,95, 3,19)	76,08 ± 11,17 (50,00, 95,01)	40,15 ± 1,57 (37,14, 43,22)	3,97 ± 1,13 (1,88, 6,15)
		RF	69,55 ± 1,48 (66,60, 72,30)	4,16 ± 1,14 (2,14, 6,43)	81,30 ± 10,24 (60,00, 100,00)	69,36 ± 1,49 (66,39, 72,17)	7,90 ± 2,07 (4,15, 12,03)
		NB	71,00 ± 1,46 (68,20, 73,80)	3,93 ± 1,11 (1,90, 6,23)	72,66 ± 11,13 (50,00, 92,86)	70,98 ± 1,47 (68,26, 73,91)	7,43 ± 2,02 (3,73, 11,52)
		SVM	75,09 ± 1,36 (72,60, 77,70)	2,94 ± 1,08 (1,12, 5,20)	45,20 ± 12,61 (21,03, 70,00)	75,58 ± 1,37 (73,10, 78,29)	5,50 ± 1,97 (2,12, 9,52)
		RNN	42,91 ± 1,56 (39,80, 46,10)	2,39 ± 0,64 (1,20, 3,67)	86,48 ± 8,60 (66,67, 100,00)	42,19 ± 1,57 (39,15, 45,26)	4,64 ± 1,22 (2,36, 7,04)
		TCN	47,06 ± 1,61 (43,90, 50,30)	2,42 ± 0,67 (1,16, 3,77)	81,27 ± 10,02 (60,00, 100,00)	46,50 ± 1,61 (43,42, 49,75)	4,70 ± 1,26 (2,28, 7,23)
		Trans	43,52 ± 1,62 (40,50, 46,70)	2,34 ± 0,64 (1,19, 3,62)	83,49 ± 9,64 (62,50, 100,00)	42,86 ± 1,62 (39,84, 46,22)	4,54 ± 1,21 (2,35, 6,94)
		EMERGE TS-only	69,36 ± 1,46 (66,60, 72,20)	3,07 ± 1,00 (1,27, 5,15)	59,06 ± 12,97 (33,33, 83,36)	69,53 ± 1,47 (66,67, 72,39)	5,83 ± 1,83 (2,47, 9,62)
	ARF 12h	LR	52,38 ± 1,56 (49,40, 55,60)	1,60 ± 0,56 (0,62, 2,71)	68,58 ± 14,23 (38,46, 92,86)	52,20 ± 1,58 (49,14, 55,48)	3,13 ± 1,07 (1,23, 5,24)
		RF	95,04 ± 0,70 (93,60, 96,40)	3,19 ± 2,87 (0,00, 9,76)	11,64 ± 10,34 (0,00, 33,33)	95,98 ± 0,63 (94,72, 97,17)	4,92 ± 4,33 (0,00, 14,55)
		NB	85,38 ± 1,13 (83,10, 87,50)	2,21 ± 1,22 (0,00, 4,80)	27,70 ± 14,03 (0,00, 58,35)	86,03 ± 1,11 (83,72, 88,12)	4,06 ± 2,21 (0,00, 8,70)
		SVM	90,86 ± 0,95 (88,90, 92,70)	2,37 ± 1,69 (0,00, 6,16)	17,72 ± 12,02 (0,00, 42,88)	91,69 ± 0,93 (89,81, 93,41)	4,13 ± 2,88 (0,00, 10,42)
		RNN	55,51 ± 1,58 (52,40, 58,50)	1,78 ± 0,61 (0,68, 3,06)	71,24 ± 13,76 (42,86, 100,00)	55,34 ± 1,60 (52,17, 58,32)	3,47 ± 1,16 (1,35, 5,90)
		TCN	61,96 ± 1,53 (59,10, 65,10)	1,90 ± 0,68 (0,74, 3,38)	64,81 ± 14,91 (33,33, 91,68)	61,93 ± 1,54 (58,95, 64,99)	3,68 ± 1,29 (1,46, 6,50)
		Trans	46,86 ± 1,59 (43,80, 50,10)	1,68 ± 0,54 (0,73, 2,78)	80,65 ± 12,49 (53,85, 100,00)	46,48 ± 1,60 (43,41, 49,75)	3,29 ± 1,04 (1,45, 5,40)
		EMERGE TS-only	84,80 ± 1,16 (82,40, 87,00)	2,66 ± 1,31 (0,63, 5,70)	35,01 ± 14,62 (9,07, 66,67)	85,36 ± 1,16 (82,93, 87,49)	4,91 ± 2,37 (1,18, 10,17)
	Shock 4h	LR	69,20 ± 1,44 (66,30, 72,00)	6,71 ± 1,37 (3,96, 9,21)	53,36 ± 7,92 (37,49, 68,30)	69,85 ± 1,46 (66,95, 72,73)	11,89 ± 2,30 (7,29, 16,00)
		RF	95,99 ± 0,62 (94,70, 97,20)	19,20 ± 30,76 (0,00, 100,00)	1,11 ± 1,72 (0,00, 5,56)	99,85 ± 0,12 (99,58, 100,00)	2,08 ± 3,17 (0,00, 10,26)
		NB	61,97 ± 1,52 (58,90, 64,80)	6,19 ± 1,19 (3,84, 8,51)	61,73 ± 7,88 (45,45, 76,67)	61,98 ± 1,56 (58,79, 64,90)	11,23 ± 2,05 (7,14, 15,14)
		SVM	96,02 ± 0,62 (94,80, 97,30)	8,97 ± 25,64 (0,00, 100,00)	0,35 ± 0,97 (0,00, 3,03)	99,91 ± 0,10 (99,69, 100,00)	0,68 ± 1,84 (0,00, 5,88)
		RNN	70,50 ± 1,39 (67,60, 73,10)	6,88 ± 1,44 (4,07, 9,63)	52,27 ± 8,15 (36,84, 67,44)	71,25 ± 1,39 (68,46, 73,90)	12,12 ± 2,39 (7,27, 16,67)
		TCN	80,34 ± 1,24 (77,90, 82,60)	8,15 ± 1,94 (4,24, 12,18)	39,33 ± 7,89 (24,44, 55,56)	82,01 ± 1,23 (79,56, 84,36)	13,46 ± 3,02 (7,25, 19,59)
		Trans	71,71 ± 1,43 (68,90, 74,30)	7,44 ± 1,53 (4,32, 10,37)	54,57 ± 8,24 (38,46, 70,76)	72,41 ± 1,43 (69,66, 75,10)	13,05 ± 2,53 (7,87, 17,79)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	95,34 ± 0,67 (94,00, 96,60)	14,91 ± 11,11 (0,00, 40,00)	4,11 ± 3,17 (0,00, 11,43)	99,05 ± 0,31 (98,43, 99,59)	6,32 ± 4,73 (0,00, 16,67)
	Shock 12h	LR	80,66 ± 1,26 (78,10, 83,10)	5,53 ± 1,69 (2,62, 8,95)	38,01 ± 9,38 (20,00, 56,00)	81,85 ± 1,26 (79,38, 84,25)	9,62 ± 2,79 (4,74, 15,04)
		RF	97,28 ± 0,51 (96,20, 98,20)	2,90 ± 16,63 (0,00, 100,00)	0,11 ± 0,68 (0,00, 3,03)	99,99 ± 0,02 (99,90, 100,00)	0,22 ± 1,29 (0,00, 5,88)
		NB	45,17 ± 1,63 (41,90, 48,40)	3,40 ± 0,75 (2,06, 5,02)	69,84 ± 8,82 (52,00, 87,50)	44,48 ± 1,65 (41,13, 47,74)	6,47 ± 1,38 (3,99, 9,38)
		SVM	97,28 ± 0,51 (96,20, 98,20)	0,00 ± 0,00 (0,00, 0,00)	0,00 ± 0,00 (0,00, 0,00)	100,00 ± 0,00 (100,00, 100,00)	0,00 ± 0,00 (0,00, 0,00)
		RNN	85,26 ± 1,10 (83,10, 87,40)	5,65 ± 1,97 (2,19, 9,74)	28,10 ± 8,48 (12,00, 44,00)	86,86 ± 1,08 (84,70, 88,89)	9,36 ± 3,11 (3,70, 15,64)
		TCN	82,69 ± 1,19 (80,30, 85,00)	5,69 ± 1,79 (2,40, 9,39)	34,43 ± 9,06 (16,13, 52,18)	84,05 ± 1,17 (81,78, 86,25)	9,72 ± 2,91 (4,25, 15,55)
		Trans	78,58 ± 1,30 (76,00, 81,00)	5,47 ± 1,57 (2,54, 8,78)	42,20 ± 9,60 (23,07, 60,71)	79,60 ± 1,29 (77,02, 82,03)	9,65 ± 2,64 (4,56, 15,25)
		EMERGE TS-only	94,95 ± 0,68 (93,60, 96,20)	6,68 ± 4,79 (0,00, 18,18)	6,65 ± 4,75 (0,00, 16,67)	97,42 ± 0,51 (96,39, 98,36)	6,55 ± 4,59 (0,00, 16,67)

B.3 Additional Test Metrics for Models Trained on eICU

Table B.3 reports the supplementary test metrics for the models trained on eICU. The table includes the internal eICU test set and the external MIMIC-IV test set, complementing the main results presented in Section 5.3. Values are reported as bootstrap mean ± standard deviation, with the 95% confidence interval shown below.

Table B.3: Supplementary test metrics for models trained on eICU and evaluated on the internal eICU test set and the external MIMIC-IV test set. Values are reported as mean ± standard deviation, with the 95% confidence interval shown below.

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
eICU	Mortality 48h	LR	79,92 ± 1,24 (77,50, 82,20)	31,79 ± 2,96 (25,62, 37,45)	73,01 ± 3,94 (64,76, 80,70)	80,77 ± 1,31 (78,20, 83,22)	44,22 ± 3,24 (37,39, 50,14)
		RF	90,63 ± 0,92 (88,80, 92,50)	62,21 ± 5,99 (50,72, 73,69)	36,24 ± 4,83 (26,92, 46,50)	97,30 ± 0,52 (96,21, 98,31)	45,65 ± 4,88 (35,93, 55,21)
		NB	74,00 ± 1,39 (71,30, 76,60)	22,87 ± 2,65 (17,74, 28,03)	58,01 ± 4,73 (48,60, 67,59)	75,97 ± 1,46 (73,19, 78,82)	32,74 ± 3,21 (26,51, 38,99)
		SVM	80,87 ± 1,22 (78,50, 83,30)	33,16 ± 3,11 (26,86, 39,01)	73,75 ± 3,93 (65,96, 81,42)	81,75 ± 1,29 (79,25, 84,19)	45,68 ± 3,34 (38,94, 51,85)
		RNN	74,65 ± 1,32 (72,00, 77,20)	27,44 ± 2,60 (22,15, 32,35)	80,18 ± 3,77 (72,55, 87,30)	73,97 ± 1,42 (71,13, 76,62)	40,83 ± 3,11 (34,40, 46,78)
		TCN	77,51 ± 1,30 (75,00, 80,00)	29,49 ± 2,80 (24,03, 35,10)	75,96 ± 3,92 (67,96, 83,81)	77,70 ± 1,37 (75,03, 80,29)	42,42 ± 3,21 (35,87, 48,56)
		Trans	81,88 ± 1,25 (79,50, 84,30)	34,59 ± 3,28 (27,98, 40,83)	73,73 ± 4,19 (65,38, 81,73)	82,88 ± 1,29 (80,29, 85,36)	47,01 ± 3,49 (39,88, 53,53)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	91,18 ± 0,90 (89,40, 92,90)	71,19 ± 6,32 (58,82, 83,02)	32,46 ± 4,55 (23,71, 41,77)	98,39 ± 0,42 (97,49, 99,10)	44,42 ± 4,94 (34,48, 54,45)
	ARF 4h	LR	79,04 ± 1,32 (76,50, 81,60)	5,11 ± 1,53 (2,34, 8,38)	68,36 ± 11,97 (44,44, 90,91)	79,21 ± 1,34 (76,56, 81,76)	9,47 ± 2,71 (4,48, 15,25)
		RF	97,57 ± 0,46 (96,50, 98,40)	30,73 ± 9,70 (11,99, 50,00)	40,86 ± 12,34 (16,67, 66,67)	98,50 ± 0,37 (97,67, 99,19)	34,45 ± 9,72 (14,27, 53,33)
		NB	21,88 ± 1,30 (19,50, 24,60)	1,94 ± 0,49 (1,00, 2,93)	95,69 ± 5,18 (83,33, 100,00)	20,67 ± 1,27 (18,36, 23,34)	3,79 ± 0,94 (1,98, 5,69)
		SVM	85,67 ± 1,08 (83,60, 87,70)	7,26 ± 2,17 (3,31, 11,88)	67,04 ± 12,24 (42,86, 89,49)	85,98 ± 1,09 (83,89, 88,02)	13,02 ± 3,66 (6,29, 20,54)
		RNN	91,63 ± 0,85 (89,80, 93,30)	13,03 ± 3,54 (6,32, 20,00)	73,98 ± 11,37 (50,00, 93,33)	91,92 ± 0,84 (90,20, 93,54)	21,99 ± 5,36 (11,65, 32,33)
		TCN	92,85 ± 0,79 (91,30, 94,20)	13,44 ± 3,99 (5,97, 21,05)	63,25 ± 12,58 (36,36, 88,24)	93,34 ± 0,77 (91,77, 94,73)	21,98 ± 5,87 (10,66, 32,88)
		Trans	85,51 ± 1,11 (83,30, 87,60)	8,49 ± 2,29 (4,12, 13,05)	81,66 ± 9,99 (60,00, 100,00)	85,58 ± 1,12 (83,27, 87,63)	15,29 ± 3,82 (7,79, 22,86)
		EMERGE TS-only	98,38 ± 0,40 (97,60, 99,10)	48,01 ± 26,69 (0,00, 100,00)	14,18 ± 9,19 (0,00, 33,33)	99,76 ± 0,16 (99,39, 100,00)	21,02 ± 12,41 (0,00, 45,45)
	ARF 12h	LR	80,56 ± 1,23 (78,20, 82,90)	3,76 ± 1,33 (1,49, 6,70)	66,42 ± 14,83 (36,35, 100,00)	80,72 ± 1,23 (78,34, 83,13)	7,08 ± 2,41 (2,86, 12,38)
		RF	81,62 ± 1,20 (79,20, 83,80)	4,99 ± 1,54 (2,27, 8,08)	85,39 ± 11,46 (60,00, 100,00)	81,58 ± 1,20 (79,19, 83,92)	9,40 ± 2,76 (4,44, 14,88)
		NB	16,24 ± 1,16 (14,00, 18,50)	1,21 ± 0,36 (0,58, 1,96)	91,03 ± 8,81 (70,00, 100,00)	15,39 ± 1,14 (13,13, 17,61)	2,38 ± 0,71 (1,15, 3,84)
		SVM	89,79 ± 0,98 (87,90, 91,70)	6,33 ± 2,34 (2,25, 11,32)	58,81 ± 15,10 (28,57, 87,50)	90,14 ± 0,97 (88,29, 92,00)	11,34 ± 3,95 (4,20, 19,65)
		RNN	94,75 ± 0,69 (93,30, 96,00)	12,15 ± 4,33 (4,44, 21,43)	59,17 ± 15,44 (28,57, 87,50)	95,16 ± 0,67 (93,82, 96,45)	19,92 ± 6,42 (7,55, 33,33)
		TCN	97,24 ± 0,53 (96,20, 98,30)	17,85 ± 7,74 (4,35, 34,78)	40,61 ± 15,84 (12,50, 75,00)	97,88 ± 0,46 (96,96, 98,79)	24,27 ± 9,44 (6,25, 43,48)
		Trans	95,28 ± 0,66 (93,90, 96,50)	13,53 ± 4,92 (5,00, 23,91)	59,50 ± 15,42 (27,27, 87,50)	95,68 ± 0,63 (94,34, 96,86)	21,77 ± 7,09 (8,33, 36,11)
		EMERGE TS-only	98,84 ± 0,33 (98,10, 99,40)	44,09 ± 30,58 (0,00, 100,00)	14,38 ± 11,04 (0,00, 37,50)	99,80 ± 0,14 (99,49, 100,00)	20,51 ± 14,11 (0,00, 47,62)
	Shock 4h	LR	72,39 ± 1,41 (69,90, 75,10)	8,52 ± 1,63 (5,35, 11,71)	62,31 ± 7,74 (47,72, 77,15)	72,80 ± 1,42 (70,26, 75,52)	14,95 ± 2,66 (9,73, 20,12)
		RF	95,69 ± 0,66 (94,40, 97,00)	39,86 ± 11,81 (16,67, 62,50)	19,95 ± 6,63 (7,14, 33,33)	98,77 ± 0,37 (98,02, 99,48)	26,23 ± 7,87 (10,52, 40,63)
		NB	17,00 ± 1,19 (14,70, 19,20)	4,32 ± 0,69 (2,92, 5,61)	95,66 ± 3,29 (88,00, 100,00)	13,80 ± 1,10 (11,69, 16,01)	8,25 ± 1,27 (5,67, 10,60)
		SVM	76,78 ± 1,33 (74,40, 79,40)	10,39 ± 1,91 (6,61, 14,04)	64,86 ± 7,78 (50,00, 80,00)	77,27 ± 1,35 (74,71, 79,92)	17,85 ± 3,01 (11,90, 23,53)
		RNN	68,57 ± 1,47 (65,90, 71,60)	8,83 ± 1,51 (5,92, 11,97)	75,58 ± 6,82 (61,76, 88,01)	68,29 ± 1,52 (65,45, 71,38)	15,77 ± 2,48 (10,95, 20,76)
		TCN	77,49 ± 1,27 (75,00, 79,90)	11,10 ± 1,99 (7,11, 15,04)	67,98 ± 7,71 (52,36, 81,58)	77,87 ± 1,30 (75,36, 80,34)	19,02 ± 3,11 (12,50, 25,00)
		Trans	75,99 ± 1,33 (73,50, 78,50)	10,54 ± 1,88 (6,93, 14,03)	68,74 ± 7,54 (53,65, 82,06)	76,29 ± 1,36 (73,69, 78,82)	18,22 ± 2,98 (12,41, 23,57)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	96,10 ± 0,61 (94,90, 97,30)	50,05 ± 26,96 (0,00, 100,00)	6,06 ± 3,88 (0,00, 15,00)	99,76 ± 0,16 (99,38, 100,00)	10,61 ± 6,46 (0,00, 24,39)
	Shock 12h	LR	74,16 ± 1,31 (71,70, 76,80)	6,37 ± 1,44 (3,66, 9,20)	62,10 ± 9,48 (44,00, 80,77)	74,50 ± 1,32 (71,90, 77,19)	11,51 ± 2,47 (6,80, 16,31)
		RF	96,79 ± 0,55 (95,70, 97,80)	32,72 ± 13,56 (8,33, 60,04)	16,77 ± 7,28 (3,70, 33,33)	99,03 ± 0,32 (98,35, 99,59)	21,73 ± 8,70 (5,41, 40,00)
		NB	14,46 ± 1,12 (12,30, 16,50)	2,94 ± 0,57 (1,92, 4,11)	94,94 ± 4,44 (85,29, 100,00)	12,21 ± 1,06 (10,16, 14,26)	5,69 ± 1,07 (3,74, 7,89)
		SVM	84,06 ± 1,16 (81,80, 86,20)	9,24 ± 2,22 (5,20, 13,58)	55,13 ± 9,81 (35,99, 75,02)	84,87 ± 1,14 (82,56, 87,06)	15,76 ± 3,52 (9,20, 22,44)
		RNN	82,89 ± 1,16 (80,70, 85,20)	9,17 ± 2,16 (5,08, 13,54)	59,37 ± 9,62 (40,00, 77,78)	83,55 ± 1,16 (81,26, 85,85)	15,82 ± 3,46 (9,18, 22,71)
		TCN	80,52 ± 1,22 (78,10, 82,90)	8,36 ± 1,96 (4,66, 12,17)	61,80 ± 9,61 (43,47, 80,00)	81,04 ± 1,21 (78,64, 83,54)	14,67 ± 3,21 (8,53, 20,93)
		Trans	82,82 ± 1,17 (80,60, 85,10)	9,05 ± 2,15 (5,29, 13,51)	58,82 ± 9,78 (40,00, 78,26)	83,49 ± 1,18 (81,26, 85,73)	15,63 ± 3,44 (9,47, 22,63)
		EMERGE TS-only	96,10 ± 0,61 (94,90, 97,30)	50,05 ± 26,96 (0,00, 100,00)	6,06 ± 3,88 (0,00, 15,00)	99,76 ± 0,16 (99,38, 100,00)	10,61 ± 6,46 (0,00, 24,39)
MIMIC-IV	Mortality 48h	LR	85,08 ± 1,11 (82,80, 87,20)	36,13 ± 4,07 (27,92, 44,00)	47,94 ± 4,78 (38,32, 56,61)	89,63 ± 1,01 (87,61, 91,59)	41,11 ± 3,92 (33,33, 48,40)
		RF	89,55 ± 1,00 (87,60, 91,50)	65,08 ± 12,78 (37,47, 88,89)	9,04 ± 2,77 (3,92, 14,55)	99,41 ± 0,26 (98,87, 99,89)	15,77 ± 4,48 (7,27, 24,81)
		NB	51,96 ± 1,59 (49,00, 55,10)	6,37 ± 1,19 (4,02, 8,76)	24,86 ± 3,99 (17,11, 32,97)	55,27 ± 1,67 (52,04, 58,44)	10,13 ± 1,79 (6,49, 13,63)
		SVM	83,96 ± 1,17 (81,60, 86,30)	35,36 ± 3,75 (28,16, 42,65)	56,86 ± 4,72 (47,41, 66,10)	87,27 ± 1,15 (84,92, 89,40)	43,52 ± 3,78 (35,97, 50,52)
		RNN	73,66 ± 1,39 (70,90, 76,40)	25,68 ± 2,49 (20,75, 30,70)	74,75 ± 4,21 (65,74, 82,61)	73,52 ± 1,47 (70,72, 76,51)	38,17 ± 3,04 (32,02, 44,06)
		TCN	77,13 ± 1,33 (74,70, 79,80)	28,01 ± 2,74 (22,78, 33,33)	69,88 ± 4,34 (60,63, 78,22)	78,02 ± 1,39 (75,39, 80,77)	39,93 ± 3,17 (33,71, 45,93)
		Trans	84,70 ± 1,11 (82,60, 86,90)	37,35 ± 3,74 (29,87, 44,85)	59,52 ± 4,80 (50,00, 68,75)	87,78 ± 1,11 (85,65, 89,94)	45,80 ± 3,70 (38,35, 52,83)
		EMERGE TS-only	88,92 ± 1,01 (87,10, 90,90)	48,63 ± 6,56 (35,59, 60,94)	28,23 ± 4,38 (19,49, 37,26)	96,35 ± 0,65 (95,08, 97,53)	35,58 ± 4,79 (26,09, 44,79)
	ARF 4h	LR	60,91 ± 1,53 (57,70, 63,80)	32,85 ± 2,47 (28,15, 37,68)	46,85 ± 3,19 (40,82, 52,92)	65,92 ± 1,72 (62,55, 69,14)	38,58 ± 2,51 (33,81, 43,36)
		RF	76,18 ± 1,37 (73,40, 78,70)	76,78 ± 6,13 (64,70, 88,24)	13,28 ± 2,08 (9,56, 17,60)	98,57 ± 0,42 (97,73, 99,32)	22,59 ± 3,16 (16,88, 29,10)
		NB	36,84 ± 1,53 (33,80, 39,90)	22,83 ± 1,61 (19,74, 25,92)	59,09 ± 3,00 (53,31, 64,80)	28,92 ± 1,67 (25,88, 32,28)	32,91 ± 1,97 (29,11, 36,71)
		SVM	66,62 ± 1,49 (63,50, 69,30)	35,19 ± 3,05 (29,36, 40,87)	32,32 ± 2,94 (26,62, 38,35)	78,82 ± 1,49 (75,84, 81,77)	33,65 ± 2,73 (28,16, 39,00)
		RNN	72,15 ± 1,44 (69,50, 75,00)	45,82 ± 3,56 (38,37, 52,66)	33,48 ± 2,95 (27,71, 39,48)	85,91 ± 1,27 (83,42, 88,43)	38,64 ± 2,93 (32,61, 44,40)
		TCN	73,90 ± 1,41 (71,20, 76,60)	50,60 ± 4,53 (41,66, 59,09)	23,50 ± 2,62 (18,58, 28,68)	91,84 ± 0,99 (89,93, 93,67)	32,04 ± 3,10 (25,84, 38,05)
		Trans	71,05 ± 1,42 (68,20, 73,80)	44,71 ± 3,08 (38,95, 50,59)	43,51 ± 3,03 (37,69, 49,43)	80,85 ± 1,46 (78,06, 83,60)	44,05 ± 2,68 (38,67, 49,33)

Continued on next page

Test database	Task	Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
		EMERGE TS-only	74,48 ± 1,40 (71,80, 77,20)	74,08 ± 11,67 (50,00, 95,01)	4,27 ± 1,24 (2,07, 6,83)	99,46 ± 0,28 (98,80, 99,87)	8,05 ± 2,24 (3,98, 12,58)
	ARF 12h	LR	63,33 ± 1,50 (60,20, 66,30)	21,99 ± 2,12 (18,15, 26,08)	47,53 ± 3,85 (40,00, 55,00)	66,47 ± 1,60 (63,37, 69,65)	30,03 ± 2,55 (25,38, 35,19)
		RF	73,29 ± 1,41 (70,50, 76,00)	30,68 ± 2,77 (25,00, 36,19)	48,42 ± 3,76 (40,69, 55,78)	78,24 ± 1,45 (75,30, 80,91)	37,51 ± 2,90 (31,04, 43,04)
		NB	37,21 ± 1,52 (34,30, 40,20)	13,99 ± 1,35 (11,40, 16,62)	54,09 ± 3,82 (46,15, 61,32)	33,85 ± 1,68 (30,46, 37,00)	22,21 ± 1,94 (18,50, 26,00)
		SVM	78,00 ± 1,33 (75,50, 80,60)	28,15 ± 3,98 (20,74, 36,14)	21,00 ± 2,99 (14,96, 26,47)	89,33 ± 1,08 (87,20, 91,38)	24,00 ± 3,21 (17,48, 30,10)
		RNN	76,28 ± 1,32 (73,70, 78,90)	30,53 ± 3,31 (24,10, 36,65)	33,75 ± 3,62 (26,59, 40,91)	84,73 ± 1,21 (82,35, 87,02)	32,00 ± 3,18 (25,54, 38,18)
		TCN	83,01 ± 1,19 (80,80, 85,40)	46,58 ± 6,58 (34,00, 59,09)	16,68 ± 2,92 (11,19, 22,52)	96,20 ± 0,66 (94,89, 97,40)	24,48 ± 3,83 (17,06, 31,75)
		Trans	81,66 ± 1,23 (79,20, 84,10)	42,93 ± 4,28 (34,31, 51,15)	32,01 ± 3,50 (25,29, 39,01)	91,54 ± 0,97 (89,64, 93,38)	36,60 ± 3,46 (29,76, 43,26)
		EMERGE TS-only	83,09 ± 1,13 (80,90, 85,30)	36,63 ± 14,23 (10,00, 66,67)	2,68 ± 1,21 (0,59, 5,26)	99,08 ± 0,32 (98,43, 99,64)	4,97 ± 2,19 (1,12, 9,73)
	Shock 4h	LR	71,75 ± 1,46 (68,90, 74,70)	9,77 ± 1,75 (6,29, 13,41)	47,08 ± 6,42 (34,69, 60,00)	73,26 ± 1,49 (70,34, 76,08)	16,14 ± 2,67 (10,88, 21,79)
		RF	94,07 ± 0,75 (92,60, 95,50)	37,75 ± 22,59 (0,00, 85,71)	3,59 ± 2,43 (0,00, 9,09)	99,64 ± 0,20 (99,15, 100,00)	6,47 ± 4,25 (0,00, 16,13)
		NB	35,11 ± 1,52 (32,10, 38,40)	3,47 ± 0,71 (2,15, 4,96)	38,08 ± 6,32 (25,45, 50,79)	34,93 ± 1,57 (31,74, 38,18)	6,35 ± 1,26 (3,98, 8,92)
		SVM	78,07 ± 1,31 (75,30, 80,60)	12,33 ± 2,24 (8,20, 16,82)	45,61 ± 6,46 (33,33, 57,89)	80,06 ± 1,33 (77,28, 82,66)	19,36 ± 3,20 (13,48, 25,64)
		RNN	66,64 ± 1,47 (63,70, 69,40)	10,31 ± 1,64 (7,24, 13,55)	61,80 ± 6,32 (48,44, 74,14)	66,94 ± 1,51 (64,01, 69,86)	17,64 ± 2,58 (12,65, 22,63)
		TCN	79,92 ± 1,28 (77,40, 82,40)	11,66 ± 2,38 (7,29, 16,41)	37,48 ± 6,38 (24,59, 50,70)	82,53 ± 1,22 (80,00, 84,79)	17,73 ± 3,34 (11,37, 24,41)
		Trans	76,94 ± 1,32 (74,40, 79,60)	11,94 ± 2,21 (7,66, 16,53)	46,80 ± 6,87 (32,79, 59,58)	78,79 ± 1,29 (76,23, 81,13)	18,98 ± 3,24 (12,54, 25,49)
		EMERGE TS-only	94,15 ± 0,74 (92,70, 95,60)	32,03 ± 36,32 (0,00, 100,00)	1,28 ± 1,51 (0,00, 5,17)	99,86 ± 0,13 (99,57, 100,00)	2,43 ± 2,83 (0,00, 9,68)
	Shock 12h	LR	65,31 ± 1,51 (62,40, 68,20)	6,26 ± 1,31 (3,79, 8,92)	58,95 ± 7,88 (43,33, 74,37)	65,56 ± 1,54 (62,67, 68,52)	11,29 ± 2,23 (7,02, 15,72)
		RF	95,90 ± 0,64 (94,60, 97,10)	10,25 ± 16,35 (0,00, 50,00)	1,27 ± 1,93 (0,00, 6,46)	99,59 ± 0,21 (99,16, 99,90)	2,23 ± 3,35 (0,00, 11,12)
		NB	38,03 ± 1,44 (35,30, 40,80)	1,98 ± 0,57 (0,99, 3,21)	31,94 ± 7,53 (17,49, 46,43)	38,26 ± 1,47 (35,34, 41,10)	3,72 ± 1,04 (1,87, 5,91)
		SVM	81,97 ± 1,20 (79,50, 84,30)	8,20 ± 2,10 (4,23, 12,43)	37,31 ± 7,86 (22,72, 52,78)	83,71 ± 1,17 (81,38, 85,97)	13,38 ± 3,20 (7,22, 20,00)
		RNN	72,19 ± 1,39 (69,40, 74,80)	7,75 ± 1,62 (4,79, 10,93)	58,84 ± 8,10 (43,58, 75,00)	72,71 ± 1,41 (69,89, 75,39)	13,66 ± 2,65 (8,72, 18,75)
		TCN	82,70 ± 1,19 (80,30, 85,00)	7,48 ± 2,10 (3,68, 11,85)	31,73 ± 7,62 (17,39, 47,06)	84,68 ± 1,16 (82,43, 86,92)	12,04 ± 3,19 (6,19, 18,66)
		Trans	75,64 ± 1,34 (72,90, 78,20)	7,77 ± 1,68 (4,93, 11,25)	50,62 ± 8,10 (35,71, 67,66)	76,62 ± 1,33 (73,86, 79,20)	13,44 ± 2,71 (8,60, 18,99)
		EMERGE TS-only	94,15 ± 0,74 (92,70, 95,60)	32,03 ± 36,32 (0,00, 100,00)	1,28 ± 1,51 (0,00, 5,17)	99,86 ± 0,13 (99,57, 100,00)	2,43 ± 2,83 (0,00, 9,68)

Appendix C

Model Hyperparameter Configurations

This appendix reports the hyperparameter configurations used during model development. The parameter C denotes the inverse regularisation strength used in LR and SVM, while `l1_ratio` controls the balance between L1 and L2 regularisation in the elastic-net penalty. For RF, the parameters define the number and structure of the decision trees, the number of features considered at each split, and the fraction of samples used during bootstrap aggregation. In NB, `var_smoothing` controls numerical stability by adding a small value to the variance estimates. For the neural-network models, the parameters define the recurrent cell type, hidden dimensions, convolutional channels, Transformer embedding size, attention heads, dropout, batch size, learning rate, and output aggregation rule used during training.

C.1 Optuna Search Spaces and Default Values

Table C.1 presents the hyperparameter search spaces explored by Optuna and the default values used in the reference configurations.

Table C.1: Hyperparameter search spaces explored by Optuna and default values used in the reference configurations.

Model	Parameter	Optuna search space	Default value
LR	`C`	log-uniform, $[0.01, 5.0]$	0.1254
	`l1_ratio`	continuous, $[0.0, 1.0]$	1.0
RF	`n_estimators`	integer, $[50, 200]$	177
	`max_depth`	integer, $[3, 30]$	14
	`min_samples_split`	integer, $[2, 11]$	4
	`min_samples_leaf`	integer, $[1, 11]$	9
	`max_features`	continuous, $[0.2, 1.0]$	0.2364
	`max_samples`	continuous, $[0.2, 1.0]$	0.5648
SVM	`C`	log-uniform, $[0.01, 10.0]$	1.0
	`kernel`	categorical, {`linear`, `rbf`}	`rbf`

Continued on next page

Model	Parameter	Optuna search space	Default value
	gamma	categorical, {scale, auto}	scale
NB	var_smoothing	log-uniform, $[10^{-12}, 10^{-7}]$	10^{-9}
RNN	rnn_type	categorical, {rnn, lstm, gru}	lstm
	hidden_dim	categorical, {64, 128, 256, 512}	512
	layer_dim	integer, $[1, 4]$	3
	idrop	continuous, $[0.0, 0.35]$	0
	dropout	continuous, $[0.05, 0.50]$	0.2
	bs	categorical, {8, 16, 32, 64}	16
	lr	log-uniform, $[10^{-5}, 5 \times 10^{-3}]$	10^{-3}
	loss_rule	categorical, {last, mean}	last
TCN	num_channels	categorical, {[64,64], [128,128], [256,256], [64,64,64], [128,128,128], [256,256,256], [256,256,256,256]}	[256,256,256,256]
	kernel_size	categorical, {2, 3, 5, 7}	3
	dropout	continuous, $[0.05, 0.50]$	0.2
	bs	categorical, {8, 16, 32, 64}	16
	lr	log-uniform, $[10^{-5}, 5 \times 10^{-3}]$	10^{-3}
	loss_rule	categorical, {last, mean}	last
Transformer	d_model	categorical, {64, 128, 256}	256
	n_head	categorical, {2, 4, 8}	8
	dim_ff_mul	categorical, {2, 4, 8}	4
	num_enc_layer	integer, $[1, 4]$	2
	dropout	continuous, $[0.05, 0.50]$	0.2
	bs	categorical, {8, 16, 32, 64}	16
	lr	log-uniform, $[10^{-5}, 5 \times 10^{-3}]$	10^{-3}
	loss_rule	categorical, {last, mean}	last
	warmup	categorical, {False, True}	False
	lr_factor	categorical, {0.5, 1, 2, 4, 8} if warmup enabled	0.1
	lr_steps	categorical, {500, 1000, 2000, 4000} if warmup enabled	2000

C.2 Optuna-selected Hyperparameters

Table C.2 reports the task-specific hyperparameters selected by Optuna for each training database and model. Each row corresponds to one model parameter, while columns correspond to the prediction tasks.

Table C.2: Task-specific hyperparameters selected by Optuna for each training database and model.

Training database	Model	Parameter	Mortality 48h	ARF 4h	ARF 12h	Shock 4h	Shock 12h
MIMIC-IV	LR	C	1.96518e-2	5.79809e-2	2.93756e-2	1.50940e-2	1.96518e-2
		l1_ratio	0.412864	0.643946	0.447740	0.703950	0.845616
		n_estimators	89	141	86	171	100
	RF						

Continued on next page

Training database	Model	Parameter	Mortality 48h	ARF 4h	ARF 12h	Shock 4h	Shock 12h
		max_depth	21	27	17	14	18
		min_samples_split	4	8	6	7	9
		min_samples_leaf	9	6	5	11	11
		max_features	0.736502	0.244288	0.679937	0.889245	0.201015
		max_samples	0.808091	0.517526	0.752198	0.898092	0.798395
		C	0.553379	3.00301	1.42801	0.226280	0.229977
	SVM	kernel	rbf	rbf	rbf	rbf	rbf
		gamma	scale	auto	auto	scale	scale
	NB	var_smoothing	9.92687e-8	4.33697e-8	9.85886e-8	9.93190e-8	7.89722e-8
		rnn_type	gru	gru	gru	lstm	gru
		hidden_dim	256	256	512	64	256
		layer_dim	1	4	2	2	2
	RNN	idrop	0.309839	0.300931	0.245101	0.324087	0.296068
		dropout	0.184922	0.431938	0.297694	0.411697	0.447467
		bs	32	64	32	64	64
		lr	1.68804e-4	5.21930e-4	2.91188e-4	8.51741e-4	2.42156e-3
		loss_rule	last	mean	last	mean	mean
		num_channels	128,128, 128	256,256, 256,256	64,64, 64	64,64, 64	64,64, 64
		kernel_size	7	5	7	3	5
	TCN	dropout	0.078252	0.485453	0.439598	0.096697	0.310928
		bs	8	16	8	16	8
		lr	3.72897e-5	1.27743e-4	6.73467e-5	7.98962e-4	4.80704e-3
		loss_rule	last	mean	mean	mean	mean
		d_model	128	128	128	128	64
		n_head	4	2	8	4	2
		dim_ff_mul	4	4	2	4	4
		num_enc_layer	2	2	3	2	1
		dropout	0.247049	0.293764	0.279137	0.327601	0.245002
		bs	32	8	64	8	64
	Trans	lr	3.59846e-4	2.93307e-4	1.13080e-3	6.30046e-4	1.41481e-4
		loss_rule	last	last	mean	last	last
		warmup	No	No	No	No	Yes
		lr_factor	–	–	–	–	4
		lr_steps					1000
eICU	LR	C	1.56763e-2	1.37766e-2	1.00189e-2	1.32905e-2	1.00107e-2
		l1_ratio	0.712548	0.411312	0.950393	0.333098	0.864819
		n_estimators	194	190	190	163	197
		max_depth	29	23	5	21	22
	RF	min_samples_split	11	11	4	7	9
		min_samples_leaf	10	11	2	9	11
		max_features	0.951662	0.234722	0.271778	0.369628	0.200532
		max_samples	0.868587	0.956712	0.355555	0.938255	0.910589
		C	0.657134	0.094659	0.095950	0.382752	0.161139
	SVM	kernel	rbf	rbf	rbf	rbf	rbf
		gamma	auto	scale	scale	auto	scale
	NB	var_smoothing	9.85886e-8	9.92687e-8	9.94986e-8	9.93190e-8	9.38000e-8

Continued on next page

Training database	Model	Parameter	Mortality 48h	ARF 4h	ARF 12h	Shock 4h	Shock 12h
	RNN	rnn_type	gru	gru	gru	lstm	gru
		hidden_dim	256	256	256	64	256
		layer_dim	3	2	1	1	3
		idrop	0.252669	0.348885	0.255993	0.339355	0.171158
		dropout	0.456083	0.366786	0.267069	0.248069	0.363994
		bs	32	8	8	8	32
		lr	9.95309e-5	5.17712e-4	8.22228e-4	7.02626e-4	4.84507e-4
		loss_rule	last	last	mean	mean	last
	TCN	num_channels	256, 256	64,64, 64	256,256, 256,256	64, 64	64,64, 64
		kernel_size	3	2	2	3	2
		dropout	0.168022	0.385097	0.424495	0.386547	0.469606
		bs	64	64	64	32	16
		lr	1.27192e-5	3.39198e-4	1.04623e-4	1.17961e-4	3.30202e-5
		loss_rule	last	mean	mean	mean	mean
	Trans	d_model	128	64	128	64	256
		n_head	8	8	4	4	2
		dim_ff_mul	2	4	2	2	4
		num_enc_layer	1	1	1	2	1
		dropout	0.102658	0.191291	0.157995	0.330727	0.056417
		bs	32	64	64	64	8
		lr	8.29413e-5	1.43043e-4	2.19589e-3	1.23779e-4	1.17337e-5
		loss_rule	last	mean	last	last	last
		warmup	No	Yes	Yes	No	No
		lr_factor	–	4	1	–	–
		lr_steps		4000	2000		