SAPIENZA
UNIVERSITÀ DI ROMA

Two-Stage Patient Centric GNN for Early Prediction of Clinical Outcomes

Facoltà di Ingegneria dell'informazione, informatica e statistica

Corso di Laurea in Artificial Intelligence and Robotics

Candidate

Francesco Caracci

ID number 1915689

Thesis Advisor

Prof. Christian Napoli

Academic Year 2024/2025

Thesis defended on January 2026
in front of a Board of Examiners composed by:

Prof. Riccardo Rosati (chairman)
Prof. Luca Becchetti
Prof. Luca Iocchi
Prof. Riccardo Lazzeretti
Prof. Christian Napoli
Prof. Marco Polverini
Prof. Marco Zecchini

Two-Stage Patient Centric GNN for Early Prediction of Clinical Outcomes
Bachelor's thesis. Sapienza – University of Rome

© 2025 Francesco Caracci. All rights reserved

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: caracci.1915689@studenti.uniroma1.it

Voglio ringraziare prima di tutti mia madre Loredana e mio padre Mario. Durante tutto il percorso universitario, e in generale scolastico, ho sempre sentito il vostro supporto in ogni mia decisione; mi ritengo molto fortunato.

Grazie ai miei amici della Rotonda: Federico, Diego, Tiziano, Riccardo M., Edoardo e Riccardo D. Siete dei fratelli per me e sarò per sempre grato di essere cresciuto con voi.

Grazie a Lorenzo B., Lorenzo P., Gabriele e Antonio, amici che ho conosciuto in vari momenti della mia vita e che la hanno resa sicuramente più ricca e felice. La vostra amicizia in questi anni è stata fondamentale per me.

Grazie a Ettore, abbiamo condiviso il percorso della magistrale praticamente in simbiosi e sono sicuro che senza di te non avrei mai avuto lo stesso rendimento. Il tuo aiuto è stato fondamentale per me in questi due anni.

Grazie a Giada P., Giada T., Stefano, Fabrizio, Vincenzo, Simone e Giovanni senza il nostro gruppo questo percorso universitario sarebbe stato sicuramente più noioso e avrei saltato sicuramente più lezioni.

Grazie a nonna Elena, zia Loredana, zia Cristina, zia Catia, zio Egidio e Beatrice che in modi diversi mi sono stati vicini durante questo percorso.

Grazie alla Musica e al Cinema, le arti che hanno accompagnato i miei giorni fino a questo punto della mia vita.

Abstract

Early identification of clinical deterioration enables timely transfer to the Intensive Care Unit (ICU) and can significantly improve patient outcomes. In this work, I propose a two stage deep learning framework that transforms Electronic Health Records (EHR) from the MIMIC-IV database into a patient level graph to jointly predict multiple ICU related outcomes. In Stage 1, a Transformer based Clinical State Estimator processes each hospital admission and outputs both the probability of subsequent ICU transfer and the expected time to transfer. The resulting latent representations and predicted risks are then aggregated across admissions to construct a heterogeneous patient graph, enriched with ICU vital sign statistics for patients who eventually reach the ICU. Graph edges encode both feature space similarity and intrapatient temporal relationships. Stage 2 performs patient level outcome prediction using a hybrid GCN/GATv2/GraphSAGE model with Jumping Knowledge aggregation. The model simultaneously forecasts in-hospital mortality and ICU length of stay (LOS). On MIMIC-IV, the approach achieves strong performance, including 0.932 AUROC / 0.959 accuracy for mortality prediction and 59 h RMSE / 38 h MAE for LOS regression. An ablation study demonstrates that the multiedge graph design and multitask learning objective both contribute substantially to model improvements. The method relies solely on routinely collected hospital data, is fully end to end differentiable and supports real time execution on a single GPU—highlighting its potential for prospective deployment as a graph aware clinical decision support tool.

Contents

Chapter 1

Introduction

Unrecognized clinical deterioration in hospital represents a persistent challenge for healthcare systems. When signs of worsening physiology are not detected promptly, patients may require emergency transfer to the Intensive Care Unit (ICU). Numerous studies have highlighted that emergency transfers are consistently associated with increased mortality rates and prolonged recovery times. These consequences are not solely attributed to disease severity: even operational delays and late recognition of clinical instability may play a central role. Such evidence reinforces the central idea that the ability to anticipate deterioration has a direct and measurable impact on patient outcomes [1, 2, 3].

In most hospitals, clinicians rely on Early Warning Scores (EWS) for continuous safety surveillance. These scoring systems, while simple and affordable, primarily evaluate a limited set of vital parameters against fixed clinical thresholds. Although widely used, they capture only a small part of the complexity seen in real world trajectories and their effectiveness varies a lot depending on the specific hospital

population and implementation context [4, 5, 6, 7]. Moreover, they often operate in isolation from richer clinical information, struggling to incorporate laboratory trends and the longitudinal progression of a patient throughout multiple encounters with the hospital system. As a result, early signs of decline that would be visible when considering the broader clinical picture may go unnoticed. The growing digitalization of healthcare and particularly the widespread use of Electronic Health Records (EHR), has opened new avenues for automated predictive analytics. EHRs collect and store virtually every clinical interaction, including diagnoses, procedures, laboratory values, vital signs and administrative data. Datasets such as MIMIC-IV [8], derived from real ICU admissions, provide unprecedented research opportunities thanks to their scale and completeness. Despite this richness, EHR data also exhibit several characteristics that complicate learning: variables span different modalities and units of measurement; data points are irregularly sampled, often missing and influenced by clinical workflow; and patient histories extend across multiple hospital admissions, each with its own progression and outcomes. Conventional machine learning techniques may struggle to capture such complexity, limiting their ability to generalize effectively.

In the past decade, deep learning has shown strong potential for extracting structure from raw clinical data. Sequence models such as recurrent networks and Transformers have demonstrated excellent capabilities in capturing temporal patterns and leveraging correlations across different types of measurements. These architectures have been applied successfully to tasks such as disease forecasting, risk assessment and patient phenotyping. However, a common limitation remains: the majority of works treat each admission as an independent and isolated sample [9].

This assumption neglects two key aspects of healthcare data: first, a single patient may be hospitalized multiple times, meaning that past admissions could contain valuable cues about vulnerability, chronic progression, or response to treatment that are not considered with these methods; second, similarities across patients such as age groups or clinical pathways, encode relational knowledge that can help models generalize to new cases and ignoring these connections can reduce the predictive power by a lot. Graph Neural Networks (GNNs) address these limitations by representing clinical datasets not as disjointed tables, but as networks of interconnected entities. In a graph, nodes may represent patients and edges may capture similarity relationships or temporal continuity across hospital encounters. This provides a mechanism for propagating information along clinically meaningful pathways. When a patient is admitted again, the model can incorporate accumulated knowledge about their previous stays. When two patients share similar clinical profiles, the model may infer patterns enabling earlier recognition of risk. Graph learning embodies the principle that additional context can refine predictions by moving beyond a purely individual perspective.

This thesis develops a two stage predictive pipeline (see Fig.1.1 tailored to this patient level graph paradigm. The first stage employs a Transformer based Clinical State Estimator that processes admission level EHR inputs and produces compact latent representations together with probabilistic estimates of ICU transfer and expected time to transfer. These admission level outputs are not treated as final predictions on their own; instead, they serve as informative components of a richer, patient level representation. For each patient, embeddings from their historical admissions are aggregated via a sequence encoder to produce a summary feature,

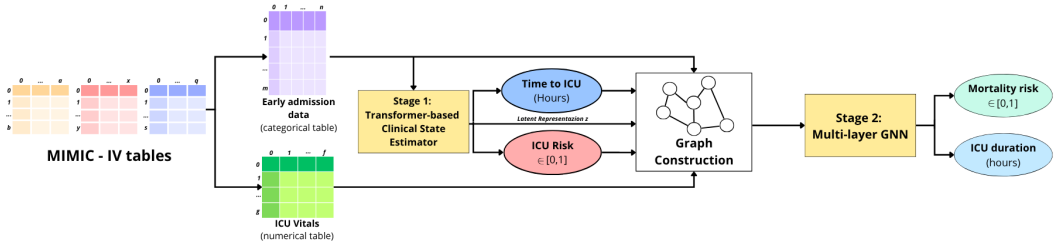

Figure 1.1. Pipeline Architecture: admission and vitals dataset aggregation; admission features are tokenized and fed to the Transformer to produce a latent representation (z) and to output ICU risk and delay; these (plus the Vitals for ICU patients) are used to enrich the graph that is processed by the GNN to predict mortality risk and ICU duration.

while the embedding from the latest admission is retained explicitly so that the downstream prediction is grounded in the present clinical state. Additional node features include aggregated vital statistics and simple summary descriptors such as the number of prior admissions or whether the patient had any previous ICU encounters. In the second stage, a heterogeneous patient graph is constructed in which each node represents a patient and node attributes encode both current and historical admission information. Edges capture population level similarity, linking patients who share comparable clinical features or risk patterns. Additionally, Stage 1 outcomes contribute to the edge design by weighting connections according to predicted likelihoods of ICU transfer, allowing the graph structure to emphasize potential clinical deterioration. This approach enables the model to consider both personal medical history and cohort level relationships. Finally, this hybrid GNN architecture, combining GCN [10], GATv2 [11] and GraphSAGE [12] layers together with Jumping Knowledge aggregation, is trained to jointly predict in hospital mortality and ICU length of stay for the most recent admission of each patient. A

deliberate design choice of this work is to focus evaluation on the latest admission. This mirrors the operational reality in which clinicians require actionable predictions for the patient currently under care, while the model benefits from auxiliary information contained in earlier encounters. Concentrating the outcome definition on the most recent stay also simplifies evaluation and supports direct comparisons with standard clinical workflows. The model's multi-task formulation further shapes the learned representations: jointly predicting mortality and length of stay encourages the network to capture aspects of both acute risk and expected resource utilization.

The proposed model will be developed and validated on the MIMIC-IV dataset. Results show that integrating transformer derived admission embeddings into a patient level graph leads to improvements over robust baselines models on the same dataset [13] and provides good multitask performance while preserving interpretability and operational constraints. Beyond performance metrics, the patient centric graph representation naturally supports explainability methods that attribute predictions both to a patient's own historical admissions and to relevant neighbors in the cohort, an important aspect when considering clinical acceptance.

Chapter 2

State of the Art

Predictive modeling in healthcare has progressively evolved from rule based systems and traditional statistical methods toward data driven approaches capable of learning complex patterns from large scale electronic health record (EHR) data. Early clinical decision support tools were typically based on handcrafted rules or linear risk scores, designed to capture a limited set of known risk factors. While interpretable and easy to deploy, such approaches struggle to accommodate the heterogeneity, temporal complexity and high dimensionality of modern clinical data. The increasing digitization of healthcare and the availability of longitudinal datasets such as MIMIC have driven the adoption of machine learning and deep learning methods that can automatically extract informative representations from raw EHR inputs. However, clinical prediction remains fundamentally challenging due to irregular sampling, missing values, multimodal data sources and strong interpatient variability. Moreover, clinically relevant outcomes such as mortality, ICU admission and length of stay are not isolated events, but emerge from a combination of patient

specific trajectories and population level patterns. Recent advances in model design have therefore focused on two complementary objectives. The first is the development of architectures capable of encoding temporal dynamics within individual patient records, capturing both short term physiological fluctuations and longer term disease progression. The second is the incorporation of relational structure, enabling models to leverage similarities and dependencies across patients, admissions and clinical contexts. At the same time, there is growing emphasis on modularity, interpretability and computational efficiency, reflecting the practical constraints of real world clinical deployment.

Two Stage Architectures

Two stage architectures have emerged as a principled response to the complexity of high dimensional predictive tasks. Rather than learning a monolithic end to end mapping from raw inputs to outcomes, these approaches decompose the problem into a representation learning phase followed by a task specific inference phase. This separation allows different components of the model to focus on distinct aspects of the problem, improving both scalability and robustness. The two stage paradigm was originally popularized in large scale advising systems, where retrieval–then–ranking pipelines are used to efficiently handle massive candidate spaces [14]. A lightweight first stage model generates rough but informative representations to identify a subset of relevant candidates, which are then refined by a more expressive second stage model. This idea has since been transferred to other fields, including tabular data analysis and healthcare prediction. In clinical applications, these architectures

are particularly appealing because they replicate the structure of many real world workflows. An initial stage can be used to summarize complex EHR data into stable latent representations or early risk estimates, while a subsequent stage performs more refined reasoning once additional context becomes available. Prior work has shown that such decompositions improve generalization and reduce sensitivity to noise, especially in settings characterized by missing data and irregular observation patterns [15, 16]. From a modeling perspective, the first stage typically learns representations that are broadly informative across tasks, capturing latent notions of patient state or disease severity. The second stage can then exploit these representations for outcome specific prediction, relational reasoning, or uncertainty modeling. This modularity also facilitates interpretability and reuse, as intermediate embeddings can be analyzed independently or transferred to related tasks.

Transformer Based EHR Modeling

Transformers have become a central architecture for sequence modeling due to their reliance on self-attention mechanisms rather than recurrent computation. This design enables efficient modeling of long range dependencies and the handling of variable length sequences, both of which are critical for EHR data. In the healthcare domain, transformer based models have been adapted to address the irregular timing and multimodal nature of clinical observations. Early approaches introduced time aware attention mechanisms that explicitly encode temporal gaps between events, allowing the model to distinguish between recent and distant measurements. For example, An et al.[17] proposed a hierarchical attention network that integrates

laboratory values, vital signs and temporal decay functions to generate patient level representations tailored to ICU prediction tasks. Other works have focused on adapting language model style architectures to structured EHR data. CEHR-BERT [18] extends the BERT framework by incorporating age and temporal embeddings into tokenized clinical events, demonstrating strong performance across a range of prediction tasks including mortality and readmission. This line of work highlights the potential of pretraining and contextualized embeddings for capturing complex clinical semantics. Transformers have also played a key role in multimodal EHR modeling. Frameworks such as TAPER [19] adopt a two stage strategy in which structured medical codes are first embedded using a transformer encoder and unstructured clinical notes are then incorporated through a text based autoencoder initialized from pretrained biomedical language models. More recent multimodal transformers integrate structured time series and free-text notes within a unified attention-based architecture, achieving improved performance while offering insights into which modalities and features drive predictions. Despite these successes, transformer based EHR models are typically limited to intrapatient modeling. Each patient or admission is treated as an independent sequence and potential relationships across patients are not explicitly represented. As a result, these models may fail to exploit population level regularities, such as recurring patterns of deterioration or shared responses to treatment, that could enhance prediction reliability.

Graph Neural Networks in Healthcare

Graph Neural Networks provide a natural framework for modeling relational data by representing entities as nodes connected through edges that encode domain specific relationships. In healthcare, this mathematical structure aligns well with the interconnected nature of clinical information, where patients, diagnoses, procedures and laboratory tests are linked through complex dependencies. Early applications of GNNs to clinical prediction focused on patient similarity graphs, where nodes correspond to patients and edges are weighted according to feature space proximity or shared clinical characteristics. Studies such as Rocheteau et al.[20] demonstrated that propagating information across similar patient trajectories can improve outcome prediction, particularly for rare events and underrepresented subpopulations. Subsequent works extended this idea to specific disease contexts, including heart failure and fairness aware risk prediction [21, 22]. More recent research has explored heterogeneous and knowledge enhanced graph structures. Instead of modeling only patient–patient similarity, these approaches incorporate multiple node types and relation semantics, including diagnoses, medications and laboratory tests. GraphCare, for example, constructs personalized knowledge graphs that combine individual EHR data with external medical knowledge, enabling the model to reason jointly over patient specific observations and broader clinical context. Such graph based frameworks offer several advantages. They provide a principled mechanism for information sharing across patients, improve robustness to data sparsity and offer opportunities for interpretability through analysis of graph structure and message passing pathways. However, many existing approaches rely on static or shallow node

features, limiting their ability to fully exploit rich temporal representations derived from raw EHR sequences.

Multi-Task Learning in Clinical Prediction

Multi-task learning (MTL) has emerged as an effective strategy for capturing shared structure across related clinical outcomes by jointly optimizing multiple prediction objectives within a common representation space. In many healthcare settings, outcomes such as in hospital mortality, ICU length of stay, discharge disposition and readmission risk are not independent events, but rather different manifestations of an underlying patient state that evolves over time. Modeling these outcomes in isolation may therefore lead to redundant or suboptimal representations that fail to capture their shared clinical determinants. By contrast, multi-task learning explicitly exploits correlations among tasks, encouraging the model to extract latent features that generalize across outcomes. From a clinical perspective, these shared representations often align with intuitive concepts such as overall disease severity, physiological reserve, or trajectory of deterioration, which influence multiple endpoints simultaneously. As a result, MTL can improve predictive accuracy while also promoting more stable and clinically meaningful internal representations. The seminal work of Harutyunyan et al.[23] established multi-task learning as a strong baseline for ICU prediction on the MIMIC-III benchmark. By jointly training recurrent neural networks to predict mortality, length of stay, physiologic decompensation and phenotype classification, the authors demonstrated consistent improvements over single-task models across all evaluated outcomes. Im-

portantly, their results highlighted the regularizing effect of multi-task supervision, which reduced overfitting and improved robustness in the presence of noisy and incomplete clinical data. This work was later extended to large scale and more heterogeneous EHR datasets by Rajkomar et al.[24], who trained deep sequence models directly on raw EHR inputs to simultaneously predict a broad set of clinical events, including mortality, prolonged hospitalization, readmission and discharge diagnoses. Their study showed that multi-task learning remains effective even at scale, enabling models to leverage vast amounts of weakly supervised data and to learn transferable patient representations across institutions and care settings. More recenttly, studies have explored multi-task learning in conjunction with attention based and transformer architectures. The work of Shickel et al. [25], for example, proposed a flexible multimodal transformer that jointly models structured data and clinical text to predict multiple ICU outcomes. Shared attention mechanisms allow the model to dynamically focus on task relevant signals while still maintaining a common representation backbone. Beyond improved performance, such architectures enhance interpretability by enabling task specific attribution analyses over shared latent features. Overall, multi-task learning provides both practical and conceptual advantages for clinical prediction. In addition to improving data efficiency and generalization, it introduces a bias that discourages task specific shortcuts and instead, promotes the learning of clinically grounded representations. These properties make MTL particularly compatible with two stage and graph based frameworks, where shared embeddings can serve as a stable substrate for subsequent relational reasoning and downstream outcome modeling.

Discussion and Positioning

Taken together, these research directions illustrate a clear and coherent trend toward modular, expressive and relational architectures for clinical prediction. Transformer based models have demonstrated strong performance in capturing temporal dynamics between patients, from irregular and heterogeneous EHR data. Graph neural networks have extended this capability by enabling population level reasoning and information sharing across clinically similar patients and multi-task learning has emerged as a powerful mechanism for encouraging generalizable representations that reflect shared clinical structure across outcomes. Despite this progress, most existing approaches address only a subset of these challenges at a time. Transformer-based models typically operate on isolated patient trajectories and do not explicitly account for interpatient relationships. GNN based methods often rely on static or shallow node features, limiting their ability to exploit rich temporal information derived from raw EHR sequences. Multi-task learning instead, is frequently embedded within homogeneous architectures, without explicit separation between representation learning and relational inference. The framework proposed in this thesis integrates these complementary ideas into a unified and principled two stage design. A transformer based first stage focuses on learning expressive admission level representations and early risk estimates from heterogeneous EHR data, capturing temporal dynamics and feature interactions at the individual level. These representations are then aggregated into patient level features and passed to a second stage graph neural network, which models relational structure across clinically similar patients and admissions. A multi-task objective jointly predicts mortality and ICU length of stay,

encouraging balanced representations that align with clinically meaningful notions of severity and progression. By explicitly separating temporal representation learning from relational outcome modeling, the proposed approach addresses key limitations of prior work while remaining computationally efficient and clinically interpretable. This modular design mirrors real world clinical workflows, in which early patient information guides monitoring and triage, while population level experience informs prognosis once critical care has commenced. As such, the framework situates this thesis at the intersection of the most promising directions in contemporary clinical machine learning and provides a solid conceptual foundation for the methodological contributions presented in the subsequent chapters.

Chapter 3

Dataset

The Medical Information Mart for Intensive Care IV (MIMIC-IV) is one of the most comprehensive and widely used publicly available databases for clinical research. It was developed and maintained by the Laboratory for Computational Physiology at the Massachusetts Institute of Technology (MIT) in collaboration with the Beth Israel Deaconess Medical Center in Boston. MIMIC-IV builds upon previous versions of the database, extending and modernizing its scope to reflect current clinical practices and electronic health record (EHR) standards. The database contains health related data associated with patients who were admitted to the Beth Israel Deaconess Medical Center between 2008 and 2022, these include hospital admissions to the emergency department, general wards and intensive care units (ICUs). In total, MIMIC-IV encompasses information from more than 65,000 unique patients with at least one ICU stay and over 200,000 emergency department visits, offering a large and heterogeneous sample of real world hospital trajectories. The data is fully anonymus in accordance with the Health Insurance Portability and Accountability

Act (HIPAA) standards, ensuring that no protected health information is included. MIMIC-IV captures the full digital footprint of a patient's hospital experience: it integrates structured and semi-structured data sources such as demographics, vital signs, laboratory results, diagnostic and procedure codes, medication prescriptions and administrations, imaging reports and outcome indicators including discharge location and mortality. Unlike earlier versions such as MIMIC-III, this release incorporates contemporary documentation practices, including ICD-10 diagnosis and procedure coding and aligns more closely with modern EHR schemas used in hospitals worldwide. The inclusion of electronically charted medication administration records, laboratory timestamps and high frequency vital measurements provides researchers with the granularity necessary for temporal and event based modeling.

The database is organized into modular components to facilitate targeted analysis and efficient data access. At a high level, MIMIC-IV is divided into two primary modules: `hosp` and `icu`. The `hosp` module is derived from the hospital-wide EHR system and contains tables that describe the entire patient stay, regardless of whether the patient was admitted to the ICU. It includes demographic information, admission and discharge details, diagnosis and procedure codes, laboratory results, microbiology tests, medication prescriptions and other administrative or clinical events recorded during hospitalization. The `icu` module instead, originates from the specialized ICU information system and contains data captured at a higher temporal resolution including continuously monitored vital signs, detailed nurse charting, fluid balance, ventilator settings and hourly laboratory measurements. The granularity of this module makes it particularly valuable for physiological modeling and the development of early warning systems for clinical deterioration. Together, these modules offer both

a hospital-wide and an ICU specific view of patient care, enabling studies that span from admission level outcomes to second by second critical care observations. Each patient in MIMIC-IV is assigned a unique identifier that allows linking information across different tables and modules, while maintaining strict anonymization. The dataset's relational structure, implemented in a standardized PostgreSQL schema, supports complex joins between entities such as patients, admissions and individual clinical events. This design provides flexibility for a wide range of research purposes, including cohort selection, longitudinal tracking and multimodal data integration. Owing to its rich content, standardized format and reproducibility, MIMIC-IV has become a cornerstone for research in machine learning for healthcare. It has been extensively used for developing and benchmarking models for mortality prediction, length of stay estimation, readmission forecasting, sepsis detection and representation learning from structured EHR data. The version used in this work, MIMIC-IV v2.2, represents the latest release at the time of the study, containing cleaned and updated records up to 2022.

Data Preprocessing

To prepare the dataset for the training of the model, a series of preprocessing steps was carried out with the goal of consolidating, cleaning and transforming the raw MIMIC-IV data into a structured format suitable for graph based learning. Data from both the `hosp` and `icu` modules were used, producing a consistent admission level representation. A base table was first created to capture demographic and administrative details for each hospital admission. This included patient identifiers,

Attribute	Data Type	Source	Attribute	Data Type	Source
hadm_id	Identifier (integer)	admissions	subject_id	Identifier (integer)	patients
gender	Binary (M/F)	patients	age_group	Categorical (8 bins)	patients
year	Integer (admission year)	patients	hosp_mortality	Binary (0/1)	admissions
adm_type	Multiclass (4 categories)	admissions	adm_loc	Multiclass (4 categories)	admissions
disc_loc	Multiclass (4 categories)	admissions	bicarbonate_test	Binary (normal/abnormal)	labevents
creatinine_test	Binary (normal/abnormal)	labevents	glucose_test	Binary (normal/abnormal)	labevents
ast_test	Binary (normal/abnormal)	labevents	bilirubin_test	Binary (normal/abnormal)	labevents
hematocrit_test	Binary (normal/abnormal)	labevents	was_in_icu	Binary (0/1)	icustays
mean_sysbp	Continuous (mmHg)	chartevents	min_sysbp	Continuous (mmHg)	chartevents
max_sysbp	Continuous (mmHg)	chartevents	mean_hr	Continuous (beats/min)	chartevents
min_hr	Continuous (beats/min)	chartevents	max_hr	Continuous (beats/min)	chartevents
mean_rr	Continuous (breaths/min)	chartevents	f min_rr	Continuous (breaths/min)	chartevents
max_rr	Continuous (breaths/min)	chartevents	mean_temp	Continuous (°C)	chartevents
min_temp	Continuous (°C)	chartevents	max_temp	Continuous (°C)	chartevents
mean_spo2	Continuous (%)	chartevents	min_spo2	Continuous (%)	chartevents
max_spo2	Continuous (%)	chartevents	mean_glucose	Continuous (mg/dL)	chartevents
min_glucose	Continuous (mg/dL)	chartevents	max_glucose	Continuous (mg/dL)	chartevents
hosp_to_icu_1	Continuous (hours)	icustays	icu_duration_1	Continuous (hours)	icustays

Table 3.1. Variables resulting after the preprocessing of the MIMIC-IV Dataset.

gender, age and admission metadata such as admission type (emergency or elective), source location (e.g., transfer from another facility or direct admission) and discharge disposition (home, deceased, transfer, other). A binary flag was added to indicate whether the patient was admitted to the ICU during that hospitalization. This ensured that every row corresponded to a single hospital stay with a well defined outcome label. To incorporate physiological data, a complementary table summarizing vital sign measurements recorded during ICU stays was constructed. For each admission ID, mean, minimum and maximum values for all available chart events—namely systolic blood pressure, heart rate, respiratory rate, body tempera-

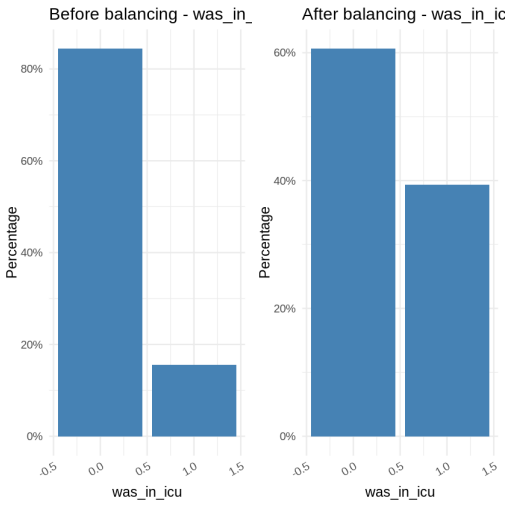

Figure 3.1. As we can see from this image the distribution of patients for the *was_in_icu* label is much more balanced and more suitable for the training process.

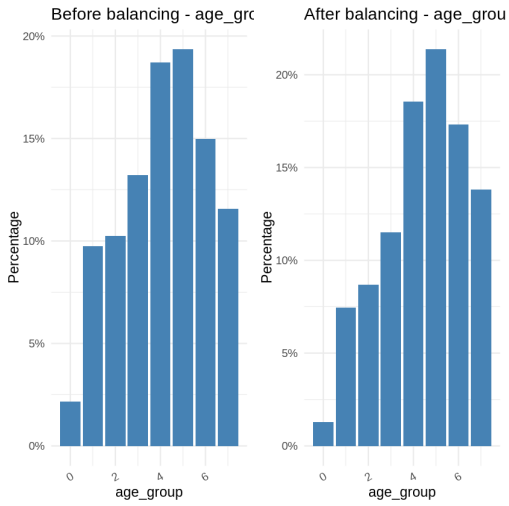

Figure 3.2. This image shows that the realism of the dataset is not been affected by the dataset preprocessing. Only the more meaningful label have been balanced.

ture, peripheral oxygen saturation (SpO_2) and point of care glucose—were computed. Each feature was z-score normalized, resulting in a standardized distribution (mean = 0, standard deviation = 1) and mitigating the effect of missing measurements and scaling discrepancies across patients. This aggregation condensed hundreds of time-stamped measurements into a compact yet clinically meaningful feature set suitable for neural processing, balancing predictive power, interpretability and computational cost. Categorical variables were normalized and encoded to improve usability. Admission related attributes were grouped into macroclasses to enhance generalization and then label-encoded. Age was discretized into eight ordered bins (0–20, 21–30, ..., 81–99). For laboratory tests, categorical indicators reflecting the overall status of each test within the first 24 hours after hospital admission were

introduced. If any measurement for a given test was recorded as “abnormal,” the corresponding flag was set to “abnormal”; if all values were normal or missing, it was set to “normal.” If no record existed at all, the flag was assigned the value “null.” This transformation allowed the model to retain clinically relevant information while reducing the dimensionality of sparse laboratory data. Two continuous outcome related variables were also computed for each admission: the time elapsed (in hours) from hospital admission to the first ICU entry (*Time to ICU*) and the total ICU stay duration (*ICU Duration*). Both values were normalized using min–max scaling to improve numerical stability and facilitate model convergence. To prevent extreme values from skewing training, admissions whose values for either of these two variables fell within the top 1% of the overall distribution were excluded. To address class imbalance, the data was subsampled to produce a more balanced yet realistic dataset. From the original 546,028 admissions, 85,242 ICU patients and 127,863 non-ICU patients were randomly selected, resulting in a final dataset of 213,105 admissions, of which approximately 40% involved an ICU stay. This balance was found to preserve the real world prevalence of critical cases while improving training stability and evaluation robustness.

Final Dataset Structure and Label Definitions

After the full preprocessing pipeline, each row of the dataset corresponds to a single hospital admission and integrates demographic information, admission metadata, laboratory indicators, aggregated ICU vital signs and outcome variables. The final structure is designed to support both stages of the modeling pipeline, providing

a coherent and machine-learning–ready representation of each clinical encounter. Every admission includes unique identifiers for the hospital stay (`hadm_id`) and the patient (`subject_id`), which serve as the basis for grouping multiple admissions belonging to the same individual. Demographic and administrative fields such as gender, discretized age group, admission year, admission type, admission source location and discharge destination describe the clinical context in which each hospital stay occurred. Laboratory information is captured through a small set of categorical indicators that encode whether key tests (including bicarbonate, creatinine, glucose, AST, bilirubin and hematocrit) were abnormal within the first 24 hours. This approach condenses sparse laboratory logs into a compact representation that preserves clinically meaningful distinctions while avoiding the need to model irregular measurement patterns directly. Physiological information is incorporated through aggregated vital signs extracted from the ICU module. For each admission, I computed mean, minimum and maximum values of systolic blood pressure, heart rate, respiratory rate, body temperature, oxygen saturation and point-of-care glucose. All these measurements were z-score normalized using statistics computed across the entire dataset, ensuring consistent scale and reducing the influence of outliers and missing measurements. Together, these physiological aggregates provide a concise yet expressive summary of a patient's clinical state during a potential ICU stay. Alongside these features, the dataset contains several outcome variables that serve as supervision for the predictive tasks. The binary `in_hosp_mortality` indicator specifies whether the patient died during the admission, while `disc_loc` reports the discharge destination as a four class categorical label. The flag `was_in_icu` identifies whether the admission included at least one ICU stay and is used to determine

when ICU related targets are applicable. Two continuous variables capture temporal aspects of critical care: `hosp_to_icu_1`, the number of hours from hospital entry to the first ICU admission and `icu_duration_1`, the total time spent in the ICU. Both were min–max normalized and trimmed by removing the top 1% of extreme values to ensure numerical stability and improve learning dynamics. These labels play different roles within the modeling pipeline: Stage 1 uses `hosp_to_icu_1` and the ICU transfer flag to learn admission level representations, while Stage 2 focuses on predicting mortality, discharge location and ICU duration for the last admission of each patient trajectory. This final structure provides a unified and information dense representation of hospital encounters, balancing clinical interpretability and computational tractability. It ensures that each admission can be understood both as an isolated event and as part of a longitudinal patient history. In the next chapter, I describe how these admission level records are transformed into patient level sequences and subsequently embedded within a graph neural architecture, outlining the full methodology used to build the two stage relational model. A summary of the variables contained in both our custom tables is shown in Table 3.1.

Chapter 4

Methodology

Graph based modeling: a conceptual introduction

Graph structured representations offer a flexible and expressive mathematical framework for modeling systems in which entities interact, co-evolve or influence one another in complex and non trivial ways. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is defined by a set of nodes and a set of edges, where each node $v \in \mathcal{V}$ carries a feature vector $\mathbf{x}_v \in \mathbb{R}^d$ and each edge (u, v) may store a scalar or vector weight describing the intensity, frequency or nature of the relationship. This formulation captures both the individuality of each element and the structure of its interactions, allowing models to interpret data not as isolated points but as part of a larger relational landscape.

This perspective is particularly attractive in clinical applications. Patient trajectories rarely develop in isolation, since they reflect combinations of demographic factors, comorbidities, disease progression patterns, physiological instability and therapeutic interventions. Groups of patients often present similar syndromes or exhibit comparable responses to treatment and conditions such as sepsis, heart failure

or respiratory failure tend to cluster within specific clinical phenotypes. Graphs allow these latent relational structures to become explicit, enabling downstream models to exploit them directly instead of relying on handcrafted features or implicit correlations captured by models trained on tabular data alone. Graph Neural Networks (GNNs) adapt deep learning to graph domains by performing iterative message passing. At layer l, each node updates its representation by aggregating information from its neighbors according to the rule

$$\mathbf{h}_v^{(l+1)} = \text{UPDATE}\Big(\mathbf{h}_v^{(l)},\, \text{AGG}\left\{\phi(\mathbf{h}_v^{(l)}, \mathbf{h}_u^{(l)}, w_{uv}) \;:\; u \in \mathcal{N}(v)\right\}\Big),$$

where ϕ models pairwise interactions and AGG is a permutation-invariant operator such as sum, mean or a learned attention weighted combination. Through successive iterations, information propagates across increasingly large neighborhoods, allowing each node to encode not only its own characteristics but also the context provided by clinically similar nodes.

This relational inductive bias is especially useful in healthcare. Early warning tasks, for example predicting ICU transfer or in-hospital mortality, benefit from contextual knowledge: if two patients exhibit similar clinical profiles at admission, one patient's trajectory may provide informative priors for the other. Conversely, GNNs can learn to attenuate spurious similarities by down-weighting edges that do not correspond to clinically meaningful relationships. In this way the model incorporates both individual level and population level signals in a principled and data driven manner. In this thesis the graph abstraction is adopted with a patient centric perspective. Each node corresponds to a unique patient and all temporal information from previous admissions is embedded directly into the node features. This choice

simplifies the graph topology, reduces computational overhead and avoids the risk of temporal leakage that may arise when explicitly connecting admissions across time. It also aligns naturally with the prediction objectives considered here, which concern the most recent admission of each patient. Edges encode clinically relevant forms of proximity. Some capture phenotypic similarity, expressed through distances between patient representations derived from demographic factors, laboratory values, physiological measurements and learned admission level embeddings. Others reflect the estimated probability of ICU transfer, which allows the model to propagate risk related information across the population while maintaining a sparse graph structure.

The remainder of this chapter describes how this conceptual framework is translated into a concrete modeling pipeline. Stage 1 presents a Transformer based encoder that summarizes each admission into a latent representation. These embeddings provide the building blocks for constructing the patient level graph, where longitudinal information, demographic factors and aggregated physiological data are combined into expressive node features. Stage 2 introduces a hybrid GNN architecture that integrates temporal, individual and relational information to produce multi-task predictions for mortality, discharge location and ICU duration.

4.1 Stage 1, Clinical State Estimator

The first component of the pipeline is a model that transforms each hospital admission into a compact and informative representation capturing both demographic characteristics and the earliest clinical indicators of patient status. This

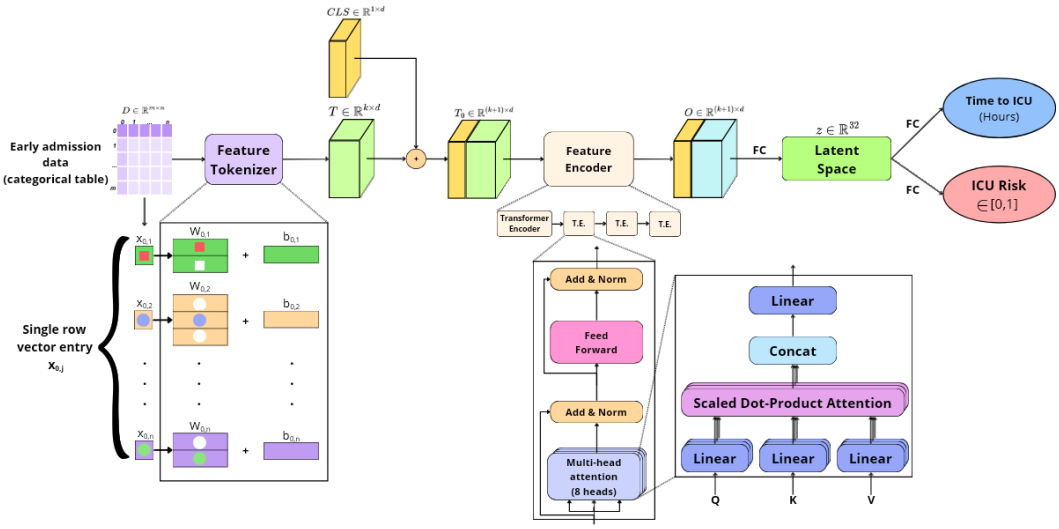

Figure 4.1. Stage 1 Architecture. Admission features are tokenized and passed through a Transformer encoder. The contextualized representation of the special token [CLS] is then projected into a latent vector z used for the prediction of ICU transfer risk and time to ICU.

transformation is crucial since all subsequent stages of the methodology rely on these latent representations as building blocks for constructing patient level histories and relational structures. The goal of this stage is therefore twofold: to provide clinically meaningful predictions at the admission level and to produce a dense embedding that encapsulates the initial severity of the encounter. Let $\mathbf{x} = (x_1, \ldots, x_{k_{\text{cat}}})$ be the categorical admission descriptors associated with a patient's encounter. These attributes include demographic factors, such as age group or gender, as well as admission related descriptors, such as the type or entry location of the hospitalization, together with structured indicators of abnormal laboratory tests recorded within the first 24 hours. Each categorical variable x_j is mapped to a trainable vector space through a dedicated embedding matrix $\mathbf{W}_j \in \mathbb{R}^{S_j \times d_e}$, where S_j denotes the number of distinct categories of the variable. Stacking these embeddings produces a sequence

$$\mathbf{T} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{k_{\text{cat}}}],$$

which serves as the input to a Transformer encoder. Subsequently, a learnable classification token `[CLS]` is is concatenated to the sequence [26]. In analogy with classification oriented Transformer architectures, this token acts as an anchor for aggregating information across all attributes and this has proven effective in NLP and, more recently, in tabular tasks [27]. While most categorical features describe simple, isolated properties, the Transformer is able to capture complex cross-feature interactions that may reflect clinically important patterns. For example, specific combinations of age groups and abnormal laboratory results may signal early organ impairment or increased physiological stress, patterns that cannot be properly captured by purely linear or additive models. The Transformer encoder employed

here consists of two layers, each equipped with multi-head self-attention with eight heads. Given query, key and value projections $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, the attention mechanism computes

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_e}}\right)\mathbf{V},$$

a formulation that allows each token to attend to all others while weighting their relative influence. This facilitates the detection of subtle interactions among laboratory and demographic features that might otherwise remain hidden. After self-attention and feed-forward refinements, the contextualized representation of the `[CLS]` token encodes the entire admission in a high dimensional latent space. This vector is then processed by a fully connected multilayer perceptron consisting of two hidden layers with non-linear activations and dropout. The final output of this network is a compact latent embedding

$$z \in \mathbb{R}^{d_z}, \qquad d_z = 32,$$

which serves as the core representation of the admission. This embedding is later used to summarize both the most recent admission and the temporal evolution of each patient's history in Stage 2. Two linear heads are built upon this embedding. The first produces a logit $\hat{r}$ that estimates the probability of the admission resulting in an ICU transfer, while the second outputs a scalar $\hat{t}$ corresponding to the predicted normalized time to ICU transfer. Formally,

$$\hat{r} = \mathbf{w}_{\text{risk}}^\top z + b_{\text{risk}}, \qquad \hat{t} = \mathbf{w}_{\text{time}}^\top z + b_{\text{time}}.$$

Training is performed using a multi-task objective that combines a binary cross-entropy loss for ICU risk and a mean-squared error for the normalized time to

ICU,

$$\mathcal{L}_{\text{stage1}} = \text{BCEWithLogits}(\hat{r}, r) + \text{MSE}(\hat{t}, t),$$

where r is the ground-truth ICU admission indicator and t is the min–max normalized time elapsed before transfer. Continuous values are normalized to avoid scale imbalances and to facilitate stable gradient-based optimization. The learning process relies on the Adam optimizer with a learning rate of 10^{-4} and incorporates early stopping based on the validation loss. In practice, this results in models that converge reliably and produce embeddings that reflect clinically meaningful patterns, including latent comorbidities, laboratory interactions and demographic factors associated with elevated risk. From a clinical standpoint, the Stage 1 embedding z can be interpreted as a compact representation of the admission's initial severity. It synthesizes heterogeneous information sources into a format that is both machine-readable and clinically informative. From a modeling perspective, it provides a stable foundation for the construction of patient histories and relational structures, which are essential components of the graph based modeling pipeline developed in the subsequent sections.

4.2 Patient level graph construction

The transition from admission level representations to a patient centric graph marks a key conceptual shift in the pipeline. While Stage 1 focuses on characterizing individual hospital encounters, Stage 2 operates at the level of entire patient trajectories. Constructing an appropriate graph representation therefore requires transforming heterogeneous, temporally distributed and partially missing clinical

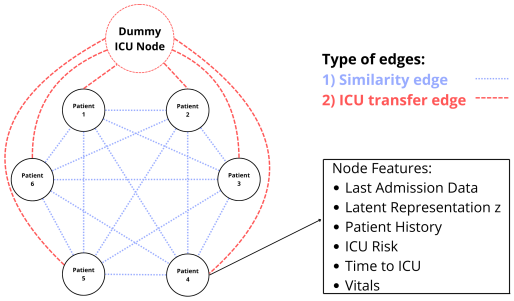

Figure 4.2. Patient centric graph construction: each node represents a patient, characterized by admission data, latent representation z, patient history, ICU risk, time to ICU and ICU Vitals (only for ICU patients). Nodes are linked to each other through weighted similarity edges and ICU transfer edges (to a dummy node).

information into a unified, relational structure where each node corresponds to a single patient. Representing one patient as one node is a deliberate modeling choice. It ensures that predictions concern the clinically relevant unit, namely the most recent admission of each patient, while avoiding the complexity induced by representing each hospitalization as a separate vertex. Such an alternative would require explicit temporal edges, careful handling of future information and a much larger graph, creating both computational and conceptual challenges. By embedding longitudinal information directly inside node features, the model benefits from temporal context without exposing itself to risks of leakage or excess structural complexity.

Each patient node is assigned a feature vector that integrates multiple sources of information. First, all encoded categorical variables associated with the last admission are included, capturing demographic descriptors, administrative details and abnormal laboratory indicators that describe the current clinical episode. Second, the latent embedding produced by Stage 1 for the final admission, denoted by z_{last}, provides a dense summary of early severity patterns and biomarker interactions

that may not be obvious in raw tabular form. Third, the entire sequence of past admissions is processed through a gated recurrent unit. Given the latent embeddings $(z_{p,1}, \ldots, z_{p,T_p})$, ordered by their temporal occurrence, the GRU produces a final hidden state $\mathbf{s}_p$ that acts as a compact representation of the patient's clinical history. This vector embeds patterns of chronicity, recurrence or physiological destabilization across multiple admissions. To supplement admission derived variables with physiological context, aggregated vital signs extracted from prior ICU stays are incorporated. High frequency measurements are averaged and normalized across admissions for each patient, providing coarse yet informative indicators of long term physiological status. Since electronic health records often contain missing or sparsely sampled measurements, missing values are imputed using dataset wide means before normalization. Although this procedure does not capture the full granularity of time-series data, it yields a stable and interpretable summary that complements higher level representations from Stage 1 and the GRU. Finally, several scalar descriptors are added, including the number of previous admissions, whether the patient has ever been in the ICU and whether the last hospitalization involved an ICU stay. These variables provide structural context regarding the patient's healthcare utilization patterns and previous exposure to critical care. The resulting node representation thus becomes

$$\mathbf{x}_p = [\text{categorical}_{\text{last}},\, z_{\text{last}},\, \mathbf{s}_p,\, \text{vitals}_{\text{agg}},\, \text{scalar descriptors}]\,,$$

a vector that encapsulates both the patient's current condition and their longitudinal background. Edges are then constructed to reflect clinically meaningful forms of similarity. The primary mechanism is a K-nearest-neighbor graph computed in

the space of node features, using Euclidean distance as the similarity metric. For each patient i, the k most similar patients are identified and an undirected edge is added for each neighbor j. The raw distances are transformed into similarity weights according to

$$w_{ij} = \frac{1}{1 + \|\mathrm{x}_i - \mathrm{x}_j\|},$$

ensuring that smaller distances correspond to stronger relational influence. This transformation normalizes edge weights to a bounded interval, stabilizing message passing during GNN propagation. In addition to similarity based edges, the graph includes a dedicated auxiliary node representing a notional ICU transfer pool. Each patient is connected to this node through a directed edge whose weight equals the ICU transfer probability predicted by Stage 1 for the patient's most recent admission. This auxiliary node contains no meaningful features and is excluded from prediction heads. Its purpose is to serve as a global aggregation point for risk related information, allowing high risk patients to exert indirect influence on others without resorting to a dense patient to patient connectivity pattern [28]. Crucially, the entire construction respects strict temporal correctness. All features, including GRU summaries, vital aggregates and Stage 1 embeddings, are computed only from admissions occurring on or before the prediction target. No future information is ever used in defining node features, edge weights or labels. This ensures that the graph represents the clinical state as it would have been known at the time of the final admission. The final graph is encoded in a `torch_geometric.data.Data` structure, containing the node feature matrix, the edge list in coordinate format, the vector of edge weights and a node mask distinguishing real patient nodes from the

auxiliary ICU transfer node. Labels for mortality, discharge destination and ICU duration are aligned with the ordering of patient nodes, enabling direct multi task learning in Stage 2.

4.3 Stage 2, Patient level GNN

Stage 2 integrates the information encoded in the patient level graph with the node features derived from longitudinal histories and admission level embeddings. The goal is to learn a joint representation that incorporates individual clinical characteristics, temporal dependencies and relational context. This unified representation is then used to predict three clinically meaningful outcomes for each patient: in-hospital mortality, discharge destination and ICU duration. Before message passing begins, each patient's sequence of past admissions is encoded by the GRU described earlier, producing a temporal summary vector $\mathbf{s}_p$, this vector is then concatenated with the other node features and projected into a shared latent space. The resulting representation serves as the input to the GNN layers that propagate information across the graph. The first propagation step is implemented through a graph convolutional layer based on the GCN formulation. Given a normalized adjacency matrix $\widehat{A}$ and degree matrix $\widehat{D}$ that incorporate self loops, node features are transformed according to

$$H^{(1)} = \sigma\big(\widehat{D}^{-1/2}\widehat{A}\widehat{D}^{-1/2}XW^{(0)}\big).$$

This layer introduces controlled smoothing across neighborhoods, stabilizes the representation space and prepares the features for subsequent adaptive weighting mechanisms.

The second layer is a GATv2 module, which replaces uniform aggregation with learned attention coefficients. For a pair of adjacent nodes (i, j), an attention weight is computed as

$$\alpha_{ij} = \text{softmax}_j\Big(a^\top \sigma(W[h_i \| h_j])\Big),$$

allowing the model to highlight clinically relevant neighbors while diminishing the influence of others. Because attention is computed locally for each node, the model can differentiate between patients who share broad similarity in some attributes but diverge in key clinical factors. A third layer based on the GraphSAGE architecture aggregates representations from various neighborhoods, encouraging the model to capture intermediate range relational patterns rather than relying solely on immediate neighbors. The outputs of the three propagation layers are combined using a Jumping Knowledge mechanism. By concatenating the representations $h^{(1)}$, $h^{(2)}$ and $h^{(3)}$, the model benefits from both low level and high level structural information, improving its ability to discriminate between clinically subtle cases. An additional refinement is introduced through dynamic edge reweighting. For each edge (i, j), a lightweight multilayer perceptron processes the representations of the connected nodes together with the original similarity weight. The resulting scalar

$$\tilde{w}_{ij} = \sigma\Big(\text{MLP}_{\text{edge}}([h_i, h_j, w_{ij}^{(0)}])\Big)$$

acts as a context aware coefficient that modulates the strength of message passing. This mechanism allows the model to amplify clinically meaningful relations or attenuate spurious ones, effectively enabling the graph structure to be partially learned rather than fixed. The final representation produced by the Jumping Knowledge module is fed into three task specific heads. Mortality is modeled via a

sigmoid logit, discharge destination through a multiclass classifier and ICU duration through a scalar regressor. The regressor is applied only to patients with valid ICU duration labels, ensuring that the model does not attempt to fit undefined values. The overall objective is the sum of binary cross-entropy for mortality, multiclass cross-entropy for discharge destination and masked mean-squared error for ICU duration. The combination of temporal encoders, multiple graph convolution operators and dynamic edge weighting equips the model with the flexibility to interpret complex patient trajectories while leveraging relational structure at multiple scales. By integrating heterogeneous information sources into a unified framework, Stage 2 forms the core predictive component of the methodology, enabling the model to reason jointly over individual histories and populationlevel patterns.

Chapter 5

Results

This chapter presents the experimental evaluation of the two stage predictive pipeline, organised into three parts: the analysis of the training methodology, a comparison of the proposed models with classical and deep learning baselines for the ICU prediction tasks (Stage 1) and a comprehensive assessment of the patient level graph model (Stage 2), both under different preprocessing configurations and in comparison with prior literature on MIMIC-III/IV. The evaluation of the proposed model required first identifying the right configuration for the patient level graph model in Stage 2. Before selecting the final approach, two methodological choices had to be addressed: determining the optimal neighbourhood size k for the similarity edge, a key component in the graph construction and establishing whether the model should be trained on batches, each forming its own local graph, or instead on a single global graph built once over the entire patient population. Both decisions influence the relational structure available to the GNN and therefore have a direct impact on predictive performance.

K Value Selection

The first step consisted in identifying a stable and informative neighbourhood size for the construction of the graph. Since the properties of the graph differ substantially depending on whether it is built on mini batches or on the full patient population, the selection of k had to be performed separately for the two training regimes. For the full graph model, several values were tested: $k = 5$, $k = 15$ and $k = 30$. Very small neighbourhoods, such as $k = 5$, tended to produce an insufficiently connected graph, limiting the propagation of information across clinically related patients. Increasing k improved performance up to a point, but larger values such as $k = 30$ often resulted in overly dense graphs making the training process much slower and producing results that do not justify the more time spent. Among the tested values, $k = 15$ provided the most reliable behaviour and the best compromise between results and train duration and was therefore adopted as the final configuration for the full graph experiments. A similar analysis was then carried out for batch graph training. Here the optimal value differed, because batch wise graph construction already introduces a level of stochasticity and implicit regularisation. Dense neighbourhoods tended to reduce this beneficial variability, while sparse local graphs helped maintain sharper distinctions among patients. Among the values tested ($k = 5$, $k = 15$, $k = 30$), the most consistent improvements were obtained with $k = 5$, which encouraged more diverse local structures across batches and resulted in a more effective learning process. A summary of the results can be observed in the Table 5.1

Training Mode	k	Mort. AUC	Mort. Acc	ICU RMSE [h]	ICU MAE [h]	Disc. Acc
Batch Graph	5	0.935	0.959	59.27	38.61	0.757
Batch Graph	15	0.911	9.952	59.92	39.84	0.740
Batch Graph	30	0.925	0.943	60.85	38.91	0.742
Full Graph	5	0.874	0.943	66.91	45.07	0.734
Full Graph	15	0.892	0.949	66.35	44.29	0.741
Full Graph	30	0.880	0.946	69.09	46.07	0.741

Table 5.1. Performance comparison across neighbourhood sizes k for batch wise and full graph training.

Batch Graph vs. Full Graph Training

Once the neighbourhood size was fixed, attention shifted to the comparison between mini batch graph training and full graph optimisation. In the first case, each batch of patients induces its own local graph, built exclusively from the nodes appearing in that batch; the resulting adjacency varies from one iteration to the next. In the second case, all patients are connected through a single, large graph whose structure remains unchanged during the entire optimisation process. Although full graph training might appear theoretically advantageous, as it preserves a coherent relational structure, its behaviour in this setting revealed a number of limitations. The patient population under study is characterised by high clinical heterogeneity, with substantial variability in comorbidities, disease trajectories and admission patterns. Training on a fixed global graph forces the model to repeatedly propagate information through the same neighbourhood configuration, which may not be uniformly representative of the entire cohort. This static structure encourages oversmoothing, especially when k is moderately large, causing node representations

to drift toward similar values and reducing the model's ability to separate subtle but clinically meaningful differences. Mini batch graph training, by contrast, introduces a mild but beneficial form of stochasticity. Since the graph is reconstructed at each iteration, the model is continually exposed to different local neighbourhoods. This variability acts as a regularising mechanism, similar to edge dropout or data augmentation, preventing the GNN from overfitting to specific patient to patient connections. Moreover, the optimisation process is better conditioned: gradients computed over diverse and smaller subgraphs tend to explore the loss landscape more effectively, leading to faster convergence and improved generalisation. The two approaches were then evaluated under their respective optimal neighbourhood sizes, that is, $k = 15$ for the full graph model and $k = 5$ for the batch based alternative. The results, summarised in Table 5.2, clearly indicate that mini batch graph training achieves consistently higher performance across all prediction tasks. Mortality classification, ICU-duration regression and discharge prediction all improved when the model was trained on local batch graphs instead of on a single global structure. This outcome suggests that, in heterogeneous real world clinical datasets, static full graph training may impose constraints that reduce the expressive capacity of the learned representations. In contrast, the controlled variability introduced by batch wise graph construction helps the GNN avoid oversmoothing and adapt more effectively to the underlying diversity of the population. For these reasons, the mini batch configuration with $k = 5$ was adopted as the definitive training strategy for Stage 2.

Training Mode	Mort. AUC	Mort. Acc	ICU RMSE [h]	ICU MAE [h]	Disc. Acc
Batch Graph ($k = 5$)	0.935	0.959	59.27	38.61	0.757
Full Graph ($k = 15$)	0.892	0.949	66.35	44.29	0.741

Table 5.2. Comparison between batch graph and full graph training regimes for Stage 2 using their respective optimal neighbourhood sizes.

5.1 Evaluation Metrics

The performance of the proposed two stage pipeline is assessed using a set of standard metrics for classification and regression tasks. These metrics are chosen to reflect both predictive accuracy and clinical relevance, allowing for a comprehensive evaluation of the model across heterogeneous outcomes.

For the binary classification tasks: ICU risk prediction in Stage 1 and mortality or discharge outcomes in Stage 2, model performance is evaluated using the *Accuracy* and the *Area Under the Receiver Operating Characteristic Curve* (AUC). Accuracy is defined by the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively. Accuracy gives an intuitive measure of overall correctness but may be sensitive to class imbalance, which is common in clinical datasets. To solve this problem, AUC is employed as a threshold independent metric that measures the model's ability to correctly rank positive cases above negative ones. Formally, AUC corresponds to the probability that a randomly selected positive instance receives a higher predicted risk score than a randomly

selected negative instance:

$$\text{AUC} = P\big(s(\mathbf{x}^+) > s(\mathbf{x}^-)\big),$$

where $s(\mathbf{x})$ denotes the model's predicted score. In clinical risk stratification, AUC is particularly relevant as it reflects discriminative ability independently of a specific operating point, aligning with real world scenarios in which decision thresholds may vary across institutions or patient populations.

For the continuous outcomes: time to ICU transfer in Stage 1 and ICU length of stay in Stage 2, performance is evaluated using *Root Mean Square Error* (RMSE) and *Mean Absolute Error* (MAE). Given true targets y_i and predictions $\hat{y}_i$, these metrics are defined by the formula:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \qquad \text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|.$$

RMSE penalises large errors more strongly due to the quadratic term, making it sensitive to substantial miscalculation that may correspond to clinically critical failures. MAE, on the other hand, provides a linear and more robust measure of average prediction error, expressed in hours, which facilitates direct clinical interpretation. Reporting both metrics allows for a balanced assessment of overall accuracy and robustness to outliers.

The selected metrics combined are able to capture various aspects of the model performance that are relevant in a clinical decision-support context. Accuracy offers an immediate sense of correctness, while AUC evaluates the quality of patient risk ranking, which is essential for prioritisation and early warning systems. Similarly, MAE reflects the typical temporal deviation between predicted and actual outcomes,

whereas RMSE highlights the presence of large and potentially unsafe errors. Together, these metrics provide a comprehensive and clinically meaningful evaluation framework, enabling consistent comparison across different model architectures and between the two stages of the pipeline. Their widespread adoption in the clinical machine learning literature further facilitates comparison with related works and supports the interpretability and reproducibility of the reported results.

5.2 Stage 1 Results Comparison

Stage 1 focuses on two prediction tasks: ICU risk classification, using accuracy and AUC metrics and time to ICU regression for patients eventually transferred to intensive care, using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Tables 5.3 and 5.4 summarise the performance obtained by a range of classical machine learning models and deep neural architectures. To facilitate interpretation of the regression results, error metrics are additionally reported relative to the scale of the target variable, whose average time to ICU is 34.49 hours with a standard deviation of 101.33 hours, which shows a highly dispersed distribution. The classical machine learning models used as baselines are Logistic Regression and Random Forests, as they represent standard approaches in clinical risk modeling. Logistic Regression estimates the probability of ICU admission by modeling the conditional distribution

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b),$$

where $\mathbf{x} \in \mathbb{R}^d$ denotes the vector of patient level covariates, $\mathbf{w}$ and b are learned

parameters and $\sigma(\cdot)$ is the logistic sigmoid function. This formulation implies a linear decision boundary in feature space and assumes additive contributions of individual predictors. Despite its limited expressive power, the model achieves an AUC of approximately 0.75 and an accuracy around 0.70, indicating that early clinical features already contain a meaningful linear signal for ICU risk stratification. However, for time to ICU regression, both RMSE and MAE remain relatively high, suggesting that linear additive assumptions are insufficient to accurately model the timing of clinical deterioration.

Random Forests extend this setting by learning an ensemble of T decision trees $\{f_t(\mathbf{x})\}_{t=1}^{T}$, each trained on a bootstrapped subset of the data and using random feature selection at each split. For classification, predictions are obtained via majority voting,

$$\hat{y} = \text{mode}\{f_1(\mathbf{x}), \ldots, f_T(\mathbf{x})\},$$

while regression outputs are averaged across trees. This architecture enables the modeling of non linear relationships and higher order feature interactions without explicit specification. Nevertheless, performance remains comparable to that of Logistic Regression, with MAE values close to 45 hours and RMSE values substantially larger than the mean time to ICU. These results suggest that while non linear effects are present, tree-based ensembles alone struggle to capture the complex and heterogeneous temporal dynamics underlying ICU transfers, particularly in the presence of rare late deterioration events.

another model used to weight the contribution of learned feature representations,

a Multi-Layer Perceptron (MLP) was considered as a neural baseline. Given an input vector $\mathbf{x}$, the MLP computes a sequence of hidden representations

$$\mathbf{h}^{(l)} = \phi\Big(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\Big),$$

where $\phi(\cdot)$ denotes a nonlinear activation function and $\mathbf{h}^{(0)} = \mathbf{x}$. In the single task configuration, separate networks are trained for ICU risk classification and time to ICU regression, each optimising its own objective function. This approach yields modest but consistent improvements over classical methods, confirming the benefit of nonlinear feature transformations while still treating risk and timing as largely independent problems.

A more structured comparison is provided by the multitask MLP, in which a shared backbone learns a common representation $\mathbf{h}_{\text{shared}}$, followed by task-specific output heads. The model jointly minimises a combined loss of the form

$$\mathcal{L} = \lambda_{\text{cls}}\,\mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_{\text{reg}}\,\mathcal{L}_{\text{reg}}(t, \hat{t}),$$

where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for ICU risk classification, $\mathcal{L}_{\text{reg}}$ denotes a regression loss (MAE or RMSE) and $\lambda_{\text{cls}}, \lambda_{\text{reg}}$ balance the two objectives. By enforcing shared hidden representations, this formulation encourages the model to exploit latent factors common to both outcomes, such as overall disease severity or early physiological deterioration. This is reflected in a reduction of MAE to approximately 40 hours, although RMSE remains close to the intrinsic variability of the target.

Finally, the FT-Transformer provides the strongest performance across both

Model	AUC	Accuracy
Logistic Regression	0.751	0.701
Random Forest	0.756	0.707
MLP (single-task)	0.757	0.707
MLP (multi-task)	0.762	0.710
FT-Transformer (multi-task)	0.770	0.710

Table 5.3. ICU Risk prediction: performance summary across classical ML and DL approaches.

tasks. Unlike standard MLPs, this architecture is specifically designed for tabular data and explicitly models interactions among heterogeneous categorical and numerical features through self-attention mechanisms. By operating on embedded feature tokens, the Transformer can dynamically weight clinically relevant feature combinations, rather than relying on fixed compositions learned implicitly through stacked layers. This capability is particularly advantageous in the present setting, where ICU transfer timing depends on complex interactions between admission characteristics, physiological summaries and laboratory indicators. Although RMSE values remain influenced by extreme late transfers, the FT-Transformer achieves the lowest MAE (35 hours), indicating a meaningful improvement in typical prediction error. Together with the observed gains in classification performance, these results support the use of attention based architectures as a more suitable inductive bias for early ICU risk and timing prediction in heterogeneous clinical tabular data.

Model	RMSE [h]	MAE [h]	RMSE / Mean	RMSE / σ_t
Linear Regression	90	45	2.61	0.89
Random Forest	92	45	2.67	0.91
MLP (single-task)	92	45	2.67	0.91
MLP (multi-task)	90	40	2.61	0.89
FT-Transformer (multi-task)	92	35	2.67	0.91

Table 5.4. Time to ICU regression performance across different approaches. The average time to ICU is 34.49 hours, with standard deviation $\sigma_t = 101.33$ hours.

5.3 Final Results Comparison

Stage 2 operates at the patient level and combines mortality prediction, ICU duration regression and discharge location classification. Table 5.5 summarises the results of the complete model under three configurations: using all admissions, removing outliers and assigning ICU vitals based on ground-truth ICU stays. Removing outliers leads to a small decrease in AUC (from 0.854 to 0.840), while accuracy remains stable. The impact is more substantial for ICU duration regression, where RMSE is halved (from 120 to 63 hours) and MAE improves markedly (72 to 42 hours). When using ground-truth ICU vitals, overall performance reaches its best values, with mortality AUC 0.935 and discharge accuracy 0.754.

It is now possible to contextualize the results of the final model, shown in the Table 5.6, through a direct comparison with representative architectures proposed in the recent literature for mortality and ICU outcome prediction on MIMIC-III and MIMIC-IV. These works span a large range of modeling choices, ranging from linear baselines to deep sequential and graph based approaches and therefore provide

Metric	Outliers	No Outliers	GT Vitals
Mortality AUC	0.854	0.840	0.935
Mortality Accuracy	0.956	0.957	0.959
ICU RMSE [h]	120.96	62.19	59.27
ICU MAE [h]	72.35	42.00	38.61
Discharge Acc	0.722	0.722	0.757

Table 5.5. Performance Summary Across Experimental Conditions: before and after outlier removal and using ground-truth ICU vitals.

a meaningful reference for assessing the contribution of the proposed two stage framework. Early clinical risk models, such as those introduced by Harutyunyan et al. (2017), rely on logistic regression or LSTM-based architectures operating on fixed length time windows. Logistic Regression achieves an AUC of 0.84, demonstrating that aggregate physiological summaries already encode a substantial amount of prognostic information. However, its linear formulation can limit the ability to capture complex interactions, which are particularly relevant in critically ill populations. The corresponding LSTM improves performance only marginally (AUC = 0.85), suggesting that temporal modeling alone is insufficient when patient trajectories exhibit high interindividual variability and heterogeneous clinical pathways. The Bui et al. (2024) work instead uses XGBoost: a tree based and gradient boosting method that further improves discrimination (AUC = 0.87) by modeling nonlinear feature interactions. While effective for tabular data, these models remain fundamentally patient isolated and lack an explicit mechanism for sharing information across

clinically similar admissions. As a result, they do not directly address population level structure or cohort effects, which are known to influence outcomes such as mortality and prolonged ICU stays. Sequential neural architectures based on gated recurrent units, as explored by van de Water et al. (2025), achieve comparable performance (AUC = 0.87), highlighting the benefits of modeling temporal dependencies in longitudinal ICU data. Nonetheless, these approaches still treat each patient independently and rely heavily on the quality and completeness of the observed time series, making them sensitive to missing data and irregular sampling. A first explicit integration of relational reasoning is provided by Rocheteau et al. (2021), who combine LSTMs with graph neural networks to propagate information across similar patient trajectories. The resulting improvement (AUC = 0.86) and competitive ICU duration errors confirm the clinical relevance of population level context, however in this architecture temporal encoding and graph propagation have a really strong relation, which may limit flexibility and interpretability when extending the model to multiple downstream tasks. The strongest baseline prior to this work is the Transformer+GNN architecture proposed by Daphne et al. (2025) which achieves an AUC of 0.90 on MIMIC-IV. By leveraging self-attention mechanisms, this model effectively captures long range temporal dependencies before applying graph based aggregation. While powerful, this single stage design requires learning temporal, relational and outcome specific representations simultaneously, potentially increasing optimisation complexity and susceptibility to noise in early representations.

In contrast, the proposed approach explicitly decomposes the learning process into two stages. Stage 1 focuses on constructing clinically meaningful patient representations that summarise early admission information, while Stage 2 operates

exclusively at the patient level, refining these representations through graph based message passing once ICU admission is known. This separation allows Stage 2 to leverage relational information without re-encoding raw temporal signals, resulting in more stable and discriminative representations. This architectural choice is reflected in the observed performance gains: the mortality AUC of 0.94 achieved by the proposed Transformer+GNN framework exceeds all previously reported results on MIMIC-III/IV, including those obtained with tightly coupled Transformer–GNN modelsand, at the same time, ICU duration regression errors remain comparable to the best values reported in the literature, indicating that improved discrimination does not come at the expense of temporal accuracy. From a clinical perspective, these results suggest that patient outcomes in the ICU are best explained by a combination of individual physiological trajectories and shared patterns across the patient population. By explicitly separating early risk encoding from relational outcome modeling, the proposed framework aligns more closely with the clinical workflow, in which early admission information guides monitoring decisions, while population level experience informs prognosis once ICU care has commenced. Overall, the comparison demonstrates that the proposed two stage Transformer+GNN architecture not only advances the state of the art in predictive performance but also offers a more modular, interpretable and clinically grounded approach to modeling complex ICU outcomes.

Work	Model	AUC
Rocheteau et al. (2021)	LSTM	0.83
	LSTM+GNN	0.86
Harutyunyan et al. (2017)	LogReg	0.84
	LSTM	0.85
Bui et al. (2024)	XGBoost	0.87
	LSTM	0.83
van de Water et al. (2025)	GRU	0.87
Daphne et al. (2025)	Transformer+GNN	0.90
This Work	Transformer+GNN	**0.94**

Table 5.6. Performance comparison with related works on MIMIC-III/IV for mortality risk prediction.

5.4 Interpretability Analysis

Interpretability is a fundamental requirement in clinical decision support systems, where model predictions must remain transparent and clinically verifiable before being used in practice. Although the two stage pipeline developed in this work achieves competitive performance, the output of the second stage results from the interaction of several components, including categorical embeddings, latent clinical representations produced by Stage 1, temporal history encoded by the GRU sequence module and relational structure derived from similarity-based edges and ICU-transfer links. For this reason, an interpretability analysis was conducted to better understand how individual node features and graph connections influence the

final predictions of the GNN. A gradient based perturbation method was employed to generate local explanations for a selected test patient. The node feature vector of the target patient was set to require gradients and the mortality risk logit was backpropagated with respect to these features. The absolute value of the resulting gradients indicates the degree to which each feature contributes to the prediction: a large magnitude reflects a strong influence on the predicted mortality probability. This procedure provides an importance ranking that includes the categorical attributes of the last admission, the Stage 1 latent clinical state, the temporal embedding summarising the patient's admission history and the normalised averages of vital signs. Alongside feature level attributions, the relational structure of the graph was also examined. After computing the predictions, all edges incident to the selected patient node were identified and their learned weights were extracted and ranked. This makes it possible to determine which neighbouring patients, or the ICU-transfer dummy hub, play the most influential role in the final prediction. High weight similarity edges typically correspond to clinically similar patients whose trajectories offer informative relational evidence, while strong connections to the ICU dummy node generally reflect patterns suggestive of severe clinical deterioration. Tables 5.7 and 5.8 report an example of feature importance scores and top incident edges obtained using this method on a random patient. This analysis reveals that the most influential inputs for the selected patient are the aggregated vital signs, particularly the systolic blood pressure statistics, which dominate the gradient based attribution. The three systolic blood pressure features (mean, minimum and maximum) occupy the top positions, indicating that the model strongly relies on cardiovascular stability when assigning high mortality risk. This result is much

clinically coherent, as hypotension and systolic variability patterns are known markers of acute deterioration. Following the vital-sign aggregates, several components of the Stage 1 latent clinical state appear prominently. These latent dimensions summarise categorical and historical information from past admissions, suggesting that the transformer based encoder successfully captures meaningful patterns that influence mortality beyond raw features alone. The presence of multiple temporal embedding dimensions (e.g., seq_emb_dim_17 and seq_emb_dim_5) further confirms that the patient's longitudinal trajectory contributes to the final decision. A smaller but still detectable influence is also observed in examination related categorical variables such as *glucose_test* and *bilirubin_test*, which aligns with the tendency of metabolic and hepatic markers to affect ICU related outcomes. The inspection of graph connections (Table 5.8) shows that the most influential edges correspond to patients with very similar clinical profiles, reflected by their relatively high learned weights (ranging from approximately 0.30 to 0.38). These bidirectional connections indicate that the GNN aggregates information from a small neighbourhood of closely related patients rather than from the wider graph, this suggests that relational information plays a complementary but fundamental role: it reinforces patterns already present in the patient's intrinsic features rather than overriding them. Overall, the interpretability results confirm that the mortality prediction arises from a combination of physiologically meaningful signals (dominated by vital signs), latent representations of patient history and a limited but coherent contribution from the most similar neighbours in the graph. This provides an intuitive and clinically plausible explanation of the model's behaviour, reinforcing the reliability of the predictions for downstream decision support applications.

Feature	Importance
avg_mean_sysbp	2.896
avg_min_sysbp	2.153
avg_max_sysbp	1.369
avg_mean_temp	0.748
stage1_latent_dim_28	0.723
stage1_latent_dim_20	0.665
seq_emb_dim_17	0.553
stage1_latent_dim_21	0.542
stage1_latent_dim_5	0.495
cat:glucose_test	0.432
seq_emb_dim_3	0.432
cat:bilirubin_test	0.418
cat:adm_type	0.392
seq_emb_dim_5	0.367
num_adm	0.359

Table 5.7. Feature importance scores for the selected patient.

Edge (src → dst)	Weight
0 → 11	0.3815
11 → 0	0.3815
0 → 26	0.3437
26 → 0	0.3437
0 → 13	0.3318
13 → 0	0.3318
31 → 0	0.3318
0 → 31	0.3318
22 → 0	0.3056
0 → 22	0.3056

Table 5.8. Top incident edges for the selected patient ranked by learned weight.

Chapter 6

Ablation Studies

The architecture presented in the previous chapters combines multiple interacting components, each designed to capture a complementary aspect of patient representation. Since the final model integrates admission level embeddings, longitudinal history, graph connectivity and heterogeneous message passing, it is interesting to quantify the actual contribution of each of these elements. Ablation studies serve precisely this purpose, providing a way to analyse how the predictive performance of the model changes when individual components are removed. Ablation experiments were carried out by retraining the model under the same conditions as the best performing model, so by batch and with $k = 5$), after selectively removing single modules or relational mechanisms. This procedure isolates the marginal contribution of each architectural block, preventing confounding interactions with the training routine or dataset partitioning. The following sections discuss the outcome of these experiments, beginning with the latent representation produced by Stage 1, moving to the temporal encoder, then to the graph connectivity and finally to the depth of

the message passing layers used in Stage 2.

6.1 Stage 1 Latent Representation

The first ablation concerns the latent vector generated by the Clinical State Estimator of Stage 1. This embedding summarises admission level information through a multi-head attention mechanism, giving the GNN an informative representation of the patient's most recent clinical state. To determine whether this contribution is essential, the Stage 1 latent vector was entirely removed from the node features. Without this embedding, the patient representation reduces to only the categorical data, the patient history provided by the GRU encoder, aggregated vitals and a few scalar descriptors, all of which are directly derived from the tabular dataset. The comparison with the full model shows a consistent degradation across all metrics. Mortality AUC decreases, classification accuracy drops and the regression tasks become less precise, with higher RMSE and MAE. This behaviour confirms that the latent representation is not redundant but rather provides a structured encoding that cannot be easily reconstructed from the remaining features. The results are reported in Table 6.1 and reinforce the motivation discussed in the Methodology for adopting a two stage architecture in which Stage 1 first learns a compact admission level representation that can then be reused by Stage 2.

6.2 Longitudinal History and the GRU Encoder

A second set of ablations examines the GRU module that encodes temporal evolution across multiple admissions. The rationale for this component, as discussed

Model Variant	Mort. AUC	Mort. Accuracy	RMSE [h]	MAE [h]	Disc. Acc
Full Model	0.935	0.959	62.46	38.61	0.754
Without Stage 1 Latent	0.930	0.949	65.96	39.31	0.735

Table 6.1. Ablation on the Stage 1 latent representation. Removing the latent embedding consistently reduces overall performance.

Model Variant	Mort. AUC	Mort. Accuracy	RMSE [h]	MAE [h]	Disc. Acc
Full Model	0.935	0.959	59.27	38.61	0.757
Without GRU History	0.912	0.928	63.12	39.63	0.754

Table 6.2. Ablation on the GRU temporal encoder. Removing longitudinal history reduces the model's ability to capture clinical deterioration over time.

in the Methodology chapter, is that clinical trajectories often unfold over several encounters and cannot be fully captured by a single snapshot of the last admission. To test this assumption, the GRU was disabled and replaced with a null vector, forcing the GNN to rely exclusively on the latest available admission together with static aggregated vitals. The results indicate that the removal of longitudinal information leads to a measurable decline in predictive performance. The effect is particularly evident in the mortality task, where temporal dynamics such as gradual instability or progressive deterioration are clinically meaningful. Regression metrics are also negatively affected, although to a slightly lesser extent, reflecting the fact that ICU duration is influenced both by acute severity at admission and by longer term patterns. Table 6.2 summarises these outcomes, which are consistent with the trends observed in Chapter 5 and with recent literature emphasising the importance of modeling temporal context in EHR based prediction.

6.3 Graph Connectivity and Relational Information

The graph used in Stage 2 has *patients* as nodes and two types of edges. The first one are similarity edges, obtained by constructing a K-nearest neighbour graph in the space of patient level features. These kind of edges encode cohort level proximity, favouring message passing among clinically similar patients. The second type consists of ICU transfer pooling edges: each patient node is connected to a single dummy ICU node with a weight proportional to the Stage 1 ICU risk prediction. This auxiliary node does not receive a label and is excluded from the loss through the node mask, but it serves as a global hub for disseminating risk related information across the cohort. To understand the contribution of each family of edge, ablation experiments were performed by selectively disabling one at a time while leaving the other intact. This procedure isolates the individual effect of each connectivity mechanism without collapsing the graph into a non relational model. The results, shown in Table 6.3, indicate a clear degradation when similarity edges are removed. This confirms that neighbourhood structure carries meaningful information, particularly for mortality prediction and discharge location. Removing ICU risk edges also produce a decline, suggesting that the dummy ICU node acts as an effective global aggregator that stabilises classification outputs. Overall, each family of edges provides a complementary contribution and the best performance is obtained when both are present, consistent with the trends discussed in Chapter 5.

Model Variant	Mort. AUC	Mort. Accuracy	RMSE [h]	MAE [h]	Disc. Acc
Full Model	0.935	0.959	59.27	38.61	0.757
No Similarity Edges	0.936	0.939	65.96	39.63	0.744
No ICU risk Edges	0.929	0.949	62.76	38.98	0.752

Table 6.3. Ablation of graph connectivity components. Removing either similarity edges or ICU risk pooling edges leads to a measurable reduction in performance, confirming the complementary role of both relational mechanisms.

6.4 Depth of Message Passing Layers

The final ablations concern the heterogeneity and depth of the message passing architecture in Stage 2. The proposed model uses a sequence of three convolutional layers, namely a GCN, a GATv2 and a GraphSAGE layer, combined through Jumping Knowledge to obtain a multiscale node representation. This design aims to exploit the stability of GCN, the adaptive weighting capabilities of attention and the inductive bias of GraphSAGE for extrapolating to unseen nodes. To investigate whether this complexity is justified, several variants were trained in which different subsets of these layers were activated. In the first group of experiments, a single layer at a time was kept active while the other two were removed, yielding three "single layer" configurations: only GCN, only GATv2 and only GraphSAGE. In the second group, pairs of layers were retained while the remaining one was removed, yielding three additional "two layer" configurations: GCN+GATv2, GCN+GraphSAGE and GATv2+GraphSAGE. Together with the full three layer architecture, these settings allow an analysis of how performance changes as the model moves from minimal to increasingly rich message passing depth. Table 6.4 summarises the results. Single

layer models perform consistently worse than the full configuration across all metrics, indicating that no individual operator is sufficient on its own to match the behaviour of the complete architecture. Among them, the GATv2-only variant tends to perform better on mortality AUC, reflecting the importance of attention-based reweighting of neighbours, while GCN-only and GraphSAGE-only configurations lag further behind. The two layer models bridge part of this gap and show that combinations such as GCN+GATv2 or GATv2+GraphSAGE already recover a substantial fraction of the performance of the full model, but still fall short of it. This pattern suggests that the three layer stack provides complementary views of the graph and that the Jumping Knowledge aggregation is effectively exploiting these multiple resolutions, in line with the improvements observed in Chapter 5.

6.5 Discussion

The ablation studies presented in this chapter provide a detailed empirical validation of the design choices proposed for the two stage architecture. Rather than relying on a single dominant component, the model derives its performance from the coordinated interaction of multiple modules, each addressing a distinct aspect of clinical representation learning. The systematic degradation observed when individual components are removed confirms that the final architecture is not overparameterised, but instead very balanced. The removal of the Stage 1 latent representation leads to a consistent decline across all tasks, highlighting the importance of learning a compact, attention-based embedding of admission level information. Clinically, this result supports the intuition that early physiological

patterns and admission context encode subtle indicators of severity that cannot be fully recovered from raw aggregated features alone. By explicitly decoupling the learning of these representations from downstream relational reasoning, the two stage design enables Stage 2 to operate on a denoised and more rich patient state, improving robustness and generalisation. In a similar way, the ablation of the GRU based encoder demonstrates the value of modeling patient history across multiple admissions. The observed reduction in mortality discrimination and regression accuracy suggests that outcomes such as death or prolonged ICU stay are influenced not only by acute presentation, but also by temporal trends reflecting chronic instability or progressive deterioration. From a clinical perspective, this aligns with real world decision making, where prior hospitalisations and evolving trajectories play a central role in risk assessment. The GRU therefore acts as a memory mechanism that contextualises the most recent admission within a broader temporal narrative. Graph connectivity ablations further clarify the complementary roles of population level similarity and ICU risk pooling edges. Disabling similarity edges weakens the model's ability to exploit cohort level structure, particularly for mortality and discharge prediction, indicating that clinically similar patients provide valuable contextual signals beyond individual features. In contrast, removing ICU risk edges reduces stability and performance across tasks, suggesting that the dummy ICU node functions as an effective global aggregator of severity information. Together, these results confirm that relational inductive biases significantly enhance prediction by embedding each patient within a clinically coherent population graph. Finally, the ablation on message passing depth provides insight into the necessity of heterogeneous and multiscale graph processing. Single layer variants fail to

match the performance of the full model, indicating that no individual operator (whether convolutional, attention-based, or inductive) is sufficient in isolation. Two layer configurations partially recover performance, but consistently fall short of the three layer stack combined with Jumping Knowledge. This pattern suggests that different layers capture complementary structural properties of the patient graph and that aggregating representations across depths allows the model to balance local neighborhood effects with more global relational context. Taken together, these ablations reinforce the central claim of this thesis: accurate and clinically meaningful prediction in complex EHR settings requires the integration of temporal modeling, relational reasoning and multiresolution representation learning. The performance gains observed in the full model emerge from the interaction of these components rather than from any single architectural choice. As such, the ablation results not only justify the proposed design, but also provide interpretative transparency, clarifying how and why each module contributes to the final outcomes. This strengthens the validity of the methodology and supports its applicability to real world clinical decision support scenarios.

Active Layers	Mort. AUC	Mort. Accuracy	RMSE [h]	MAE [h]	Disc. Acc
GCN + GATv2 + GraphSAGE (Full)	**0.935**	**0.959**	**59.27**	**38.61**	0.757
GCN only	0.894	0.941	61.76	40.65	0.740
GATv2 only	0.901	0.951	60.98	39.06	0.740
GraphSAGE only	0.904	0.952	60.01	39.95	0.739
GCN + GATv2	0.896	0.955	59.86	38.87	0.739
GCN + GraphSAGE	0.898	0.927	60.38	39.12	0.741
GATv2 + GraphSAGE	0.904	0.959	59.92	38.84	0.738

Table 6.4. Ablation on message passing depth. The heterogeneous three layer architecture outperforms both single layer and two layer variants.

Chapter 7

Conclusion

This thesis presented a two stage modeling framework for the prediction of ICU related outcomes in hospitalised patients, combining admission level sequence modeling with patient level graph neural networks. The proposed architecture was motivated by the observation that clinical decision making unfolds progressively: early information available at admission provides an initial estimate of risk, which is then refined as patient history accumulates and as contextual information from similar patients becomes available. By explicitly mirroring this process, the framework fills the gap between raw electronic health record data and clinically meaningful patient level reasoning. At the methodological level, the model integrates heterogeneous sources of information, including demographic characteristics, categorical admission descriptors, physiological summaries, longitudinal admission histories encoded via a GRU based sequence model and relational context derived from a population graph. Each component was designed to address a specific limitation of traditional clinical prediction models: the inability to capture non-linear feature interactions,

the neglect of temporal trajectories across admissions and the lack of population level contextualisation. The resulting architecture demonstrates how these challenges can be addressed within a unified and computationally efficient design. Empirical evaluation across multiple prediction tasks supports the effectiveness of this approach. The Stage 1 Clinical State Estimator achieved competitive performance in ICU transfer risk classification and time to ICU regression, confirming that a compact Transformer based model can extract informative latent representations from heterogeneous admission level features. These representations not only produced accurate predictions on their own but also served as robust building blocks for the next phase. In Stage 2, the incorporation of patient level trajectories and graph based message passing led to substantial gains in mortality prediction, ICU length of stay regression and discharge destination classification. Comparisons with established baselines and recent state of the art approaches on both MIMIC-III and MIMIC-IV datasets indicate that the proposed pipeline achieves performance that is at least comparable to and in several cases exceeds, previously reported results.

Ablation studies provided deeper insight into the internal dynamics of the model and validated the architectural choices made in this work. The admission level latent representation, the GRU encoder and the heterogeneous graph structure were all shown to contribute meaningfully to predictive performance. In particular, similarity based edges enabled local information sharing among clinically comparable patients, while ICU risk edges acted as a global stabilising mechanism that aggregated severity related signals across the cohort. These findings confirm that relational inductive biases are not merely auxiliary components, but play a central role in enabling robust and clinically coherent predictions. Beyond predictive accuracy, this thesis placed

strong emphasis on interpretability and clinical plausibility. The explainability analysis revealed that the model's predictions were driven by physiologically meaningful factors, such as blood pressure dynamics, temperature variability and latent severity dimensions that align with established medical knowledge. This alignment is critical for fostering clinician trust and for supporting the use of such models in decision support settings, where unintuitive predictions may prevent the adoption regardless of the performance. From a clinical perspective, the two stage architecture offers several practical advantages: its modular structure allows Stage 1 models to be deployed early in the admission workflow, providing preliminary risk estimates that can be updated and refined as more information becomes available. Stage 2 naturally integrates patient history and population level context, reflecting how clinicians reassess risk over time rather than making isolated decisions. Moreover, the trust on structured EHR variables ensures compatibility with existing hospital information systems, while the graph based formulation gives a flexible mechanism for the integration of new patients without retraining the entire model. SUrely, a real world deployment would require a prospective validation, a continuous monitoring for dataset shift and a careful evaluation of fairness and subgroup performance to ensure a reliable reproducibility across diverse patient populations.

Despite its strengths, this work also highlights several limitations that open avenues for future research. Graph construction currently depends on a fixed neighbourhood size governed by a manually selected parameter k, which may not fully capture the heterogeneity of patient similarity across clinical contexts. Adaptive or learned graph construction strategies could provide a more refined representation of interpatient relationships. Another possible improvement is to use, instead of the

GRU encoder, more expressive temporal models, such as hierarchical or continuous time architectures, that can better reflect the irregular and event driven nature of clinical data. Extending the framework to incorporate additional modalities, including longitudinal laboratory trends, clinical notes, or image features could also represents another possible and promising direction for the improvement of clinical fidelity. Future works may also explore a more explicit multitask formulation for both the stages of the pipeline jointly optimising complementary clinical endpoints to further regularise representation learning. Finally, evaluating how confident the model's predictions are, how well it performs when clinical data change over time and how reliably it transfers across different hospitals will be key to enabling safe and trustworthy deployment in real world clinical practice. In conclusion, this thesis demonstrates that combining admission level representation learning with patient level graph reasoning provides a principled and effective approach to modeling ICU related risks in large scale EHR data. By integrating temporal, relational and multitask perspectives within a coherent two stage framework, the proposed model advances the state of clinical predictive modeling and offers a flexible foundation for future developments aimed at improving patient outcomes through data driven, interpretable and clinically actionable insights.

Bibliography

[1] Lucienne TQ Cardoso, Cintia MC Grion, Tiemi Matsuo, Elza HT Anami, Ivanil AM Kauss, Ludmila Seko, and Ana M Bonametti. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Critical care*, 15(1):R28, 2011.

[2] Matthew M Churpek, Blair Wendlandt, Frank J Zadravecz, Richa Adhikari, Christopher Winslow, and Dana P Edelson. Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. *Journal of hospital medicine*, 11(11):757–762, 2016.

[3] Panagiotis Kiekkas, Anastasios Tzenalis, Vasiliki Gklava, Nikolaos Stefanopoulos, Gregorios Voyagis, and Diamanto Aretha. Delayed admission to the intensive care unit and mortality of critically ill adults: Systematic review and meta-analysis. *BioMed research international*, 2022(1):4083494, 2022.

[4] Armando D Bedoya, Meredith E Clement, Matthew Phelan, Rebecca C Steorts, Cara O'brien, and Benjamin A Goldstein. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical care medicine*, 47(1):49–55, 2019.

[5] Andrew Hao Sen Fang, Wan Tin Lim, and Tharmmambal Balakrishnan. Early warning score validation methodologies and performance metrics: a systematic review. *BMC medical informatics and decision making*, 20(1):111, 2020.

[6] Sonieya Nagarajah, Monika K Krzyzanowska, and Tracy Murphy. Early warning scores and their application in the inpatient oncology settings. *JCO Oncology Practice*, 18(6):465–473, 2022.

[7] David Chi-Wai Wong, Timothy Bonnici, Stephen Gerry, Jacqueline Birks, and Peter J Watkinson. Effect of digital early warning scores on hospital vital sign observation protocol adherence: Stepped-wedge evaluation. *Journal of Medical Internet Research*, 26:e46691, 2024.

[8] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[9] Clauirton A Siebra, Mascha Kurpicz-Briki, and Katarzyna Wac. Transformers in health: a systematic review on architectures for longitudinal data analysis. *Artificial Intelligence Review*, 57(2):32, 2024.

[10] TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[11] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro

Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[12] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[13] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[14] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 191–198, New York, NY, USA, 2016. Association for Computing Machinery.

[15] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. *arXiv preprint arXiv:2005.11650*, 2020.

[16] Jing Li, Botong Wu, Xinwei Sun, and Yizhou Wang. Causal hidden markov model for time series disease forecasting. *arXiv preprint arXiv:2103.16391*, 2021.

[17] Ying An, Yang Liu, Xianlai Chen, and Yu Sheng. Tertian: Clinical endpoint prediction in icu via time-aware transformer-based hierarchical attention network. *Computational Intelligence and Neuroscience*, 2022(1):4207940, 2022.

[18] Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. Cehr-bert: Incorporating temporal information from structured ehr data to improve prediction tasks. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, and *et al.*, editors, *Proceedings of Machine Learning for Health (ML4H) 2021*, volume 158 of *Proceedings of Machine Learning Research*, pages 239–260. PMLR, 2021.

[19] Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Taper: Time-aware patient ehr representation. *IEEE journal of biomedical and health informatics*, 24(11):3268–3275, 2020.

[20] E. Rocheteau, C. Tong, P. Veličković, N. Lane, and P. Liò. Predicting patient outcomes with graph representation learning. *arXiv preprint arXiv:2106.08159*, 2021.

[21] Heloisa O. Boll, Ali Amirahmadi, Amira Soliman, Stefan Byttner, and Mariana R. Mendoza. Graph neural networks for heart failure prediction on an ehr-based patient similarity graph. *Institute of Informatics, Universidade Federal do Rio Grande do Sul*, 2023.

[22] Christos Maroudis, K. Karathanasopoulou, C. C. Stylianides, G. Dimitrakopoulos, and A. S. Panayides. Fairness-aware graph neural networks for icu length of stay prediction in iot-enabled environments. *Under review or conference proceedings*, 2024.

[23] Hrayr Harutyunyan, Hayk Khachatrian, Devon C. Kale, Aram Galstyan, and

Russell Greiner. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.

[24] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Peter J. Liu, Xiaobing Liu, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Gavin E. Duggan, Gerardo Flores, Michaela Hardt, Jamie Irvine, Quoc Le, Kurt Litsch, Jake Marcus, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael Howell, Claire Cui, Greg Corrado, and Jeff Dean. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1:18, 2018.

[25] Benjamin Shickel, Patrick J. Tighe, Azra Bihorac, and Parisa Rashidi. Multi-task prediction of clinical outcomes in the intensive care unit using flexible multimodal transformers. *arXiv preprint arXiv:2111.05431*, 2021.

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[27] Yury Gorishniy, Ivan Rubachev, Victor Khrulkov, and Dmitry Vetrov. Ft-transformer: A self-attention architecture for tabular data. In *International Conference on Learning Representations*, 2022.

[28] Xin Liu, Jiayang Cheng, Yangqiu Song, and Xin Jiang. Boosting graph structure

learning with dummy nodes. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.