Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269 Université de Strasbourg

ÉCOLE DOCTORALE MATHEMATIQUES, SCIENCES DE L'INFORMATION ET DE L'INGENIEUR – ED269

ICube UMR 7357

THÈSE présentée par :

Mamadou Ben Hamidou CISSOKO

soutenue le : 23 **Octobre 2024**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Informatique

TITRE de la thèse

Adaptive Time-Aware LSTM for Predicting and Interpreting ICU Patient Trajectories from Irregular Data

THÈSE dirigée par :

M. LACHICHE Nicolas	Professeur, Université de Strasbourg

RAPPORTEURS :

Mme BRINGAY Sandra	Professeur, Université Paul-Valery Montpellier 3
M. Guyet Thomas	Assoc Professeur, Inria Research Center of Lyon

AUTRES MEMBRES DU JURY :

Mme Sinoquet Christine	Assoc Professeur, Université de Nantes
M. Weber Jonathan	Professeur, Université de Haute-Alsace (UHA)

INVITÉS :

M. CASTELAIN Vincent	Professeur**,** Université de Strasbourg
M. SAULEAU Eric-André	Professeur**,** Université de Strasbourg

Université de Strasbourg
Laboratoire des sciences de l'Ingénieur, de l'Informatique et de l'Imagerie (UMR 7357)

Adaptive Time-Aware LSTM for Predicting and Interpreting ICU Patient Trajectories from Irregular Data

Presented by CISSOKO MAMADOU BEN HAMIDOU

PhD thesis in MATHEMATICS, INFORMATION AND ENGINEERING SCIENCES

October 23, 2024

BRINGAY SANDRA	PROFESSOR	Reporter
GUYET THOMAS	ASSOCIATE PROFESSOR	Reporter
SINOQUET CHRISTINE	ASSOCIATE PROFESSOR	Examiner
WEBER JONATHAN	PROFESSOR	Examiner
LACHICHE NICOLAS	PROFESSOR	Supervisor
CASTELAIN VINCENT	PROFESSOR	Guest
SAULEAU ERIC-ANDRÉ	PROFESSOR	Guest

Abstract

In personalized predictive medicine, accurately modeling a patient's illness and care processes is crucial due to the inherent long-term temporal dependencies. However, Electronic Health Records (EHRs) often consist of episodic and irregularly timed data, stemming from sporadic hospital admissions, which create unique patterns for each hospital stay. Consequently, constructing a personalized predictive model necessitates careful consideration of these factors to accurately capture the patient's health journey and assist in clinical decision-making.

Long Short-Term Memory (LSTM) networks are effective for handling sequential data like EHRs, but they face two significant limitations: the inability to interpret prediction results and to take into account irregular time intervals between consecutive events. To address these limitations, we introduce novel deep dynamic memory neural networks called Multi-Way Adaptive and Adaptive Multi-Way Interpretable Time-Aware LSTM (MWTA-LSTM and AMITA-LSTM) designed for irregularly collected sequential data. The primary objective of both models is to leverage medical records to memorize illness trajectories and care processes, estimate current illness states, and predict future risks, thereby providing a high level of precision and predictive power.

To enhance their capabilities, both models extend the standard LSTM in two key ways. Firstly, they incorporate frequency measurement and the most recent observation to enhance personalized predictive modeling of patient illnesses, enabling a more accurate understanding of the patient's condition. Secondly, they parameterize the cell state to handle irregular timing effectively, utilizing both elapsed times and a frequency-based decay factor. These enhancements allow the models to comprehend the impact of interventions on the course of illness, facilitating the memorization of illness courses and improving the ability to capture the temporal dynamics of healthcare data, thus accommodating variations and irregularities in event and observation timing.

The effectiveness of our proposed models is validated through empirical experiments conducted on two real-world clinical datasets and three time series datasets for forecasting. The results demonstrate the superiority of our framework over current state-of-the-art models and other robust baselines. This showcases the potential of our approach in advancing personalized predictive medicine by offering a more accurate and comprehensive method for modeling patient health trajectories, ultimately aiding in more informed clinical decision-making.

Résumé

En médecine prédictive personnalisée, modéliser avec précision la maladie et les processus de soins d'un patient est crucial en raison des dépendances temporelles à long terme inhérentes. Cependant, les dossiers de santé électroniques (DSE) se composent souvent de données épisodiques et irrégulières, issues des admissions hospitalières sporadiques, créant des schémas uniques pour chaque séjour hospitalier. Par conséquent, la construction d'un modèle prédictif personnalisé nécessite une considération attentive de ces facteurs pour capturer avec précision le parcours de santé du patient et aider à la prise de décision clinique.

Les réseaux de mémoire à long terme (LSTM) sont efficaces pour traiter les données séquentielles comme les DSE, mais ils présentent deux limitations majeures : l'incapacité à interpréter les résultats des prédictions et à prendre en compte des intervalles de temps irréguliers entre les événements consécutifs. Pour surmonter ces limitations, nous introduisons de nouveaux réseaux neuronaux à mémoire dynamique profonde appelés Multi-Way Adaptive et Adaptive Multi-Way Interpretable Time-Aware LSTM (MWTA-LSTM et AMITA-LSTM), conçus pour les données séquentielles collectées de manière irrégulière. L'objectif principal des deux modèles est de tirer parti des dossiers médicaux pour mémoriser les trajectoires de maladie et les processus de soins, estimer les états de maladie actuels et prédire les risques futurs, offrant ainsi un haut niveau de précision et de pouvoir prédictif.

Pour améliorer leurs capacités, les deux modèles étendent le LSTM standard de deux manières clés. Premièrement, ils intègrent la mesure de la fréquence et l'observation la plus récente pour améliorer la modélisation prédictive personnalisée des maladies des patients, permettant une compréhension plus précise de l'état du patient. Deuxièmement, ils paramètrent l'état de la cellule pour gérer efficacement l'espacement temporel irrégulier entre les événements, en utilisant à la fois les temps écoulés et un facteur de décroissance basé sur la fréquence. Ces améliorations permettent aux modèles de comprendre l'impact des interventions sur le cours de la maladie, facilitant la mémorisation des évolutions de la santé du patient et améliorant la capacité à capturer la dynamique temporelle des données de santé, en prenant en compte les variations et les espacements temporels irréguliers des observations.

L'efficacité de nos modèles proposés est validée par des expériences empiriques menées sur deux ensembles de données cliniques réelles et trois ensembles de données de séries temporelles pour la prévision. Les résultats démontrent la supériorité de nos modèles par rapport aux modèles actuels de référence.

Declaration

I hereby declare that this thesis, entitled "Adaptive Time-Aware LSTM for Predicting and Interpreting ICU Patient Trajectories from Irregular Data", submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (PhD) in Computer Science, at the University of Strasbourg, is my original work. It has not been submitted for any other degree or examination in any other institution.

The work presented in this thesis gave rise to the following publications:

[a] **CISSOKO Mamadou Ben H**, CASTELAIN Vincent, LACHICHE Nicolas. "Prise en compte de données séquentielles hétérogènes dans l'apprentissage profond : application aux données de soins intensifs." 23e conférence francophone Extraction et Gestion des Connaissances (EGC 2023), 979-10-96289-19-6 .

[b] **CISSOKO Mamadou Ben H**, CASTELAIN Vincent, LACHICHE Nicolas. " Modeling Temporal Dynamics in Irregular ICU Data Using MWTA-LSTM." International Conference on Artificial Intelligence (EPIA 2024).

[c] **CISSOKO Mamadou Ben H**, CASTELAIN Vincent, LACHICHE Nicolas. "Predicting Patient Health Outcomes with AMITA from Irregular Time Series." 31st International Conference on Neural Information Processing (ICONIP 2024).

[d] **CISSOKO Mamadou Ben H**, CASTELAIN Vincent, LACHICHE Nicolas. "Multi-Way adaptive Time Aware LSTM for irregularly collected sequential ICU data.", Expert Systems With Applications Journal (ESWA Impact Factor of 7.5 and CiteScore of 13.80).

[e] **CISSOKO Mamadou Ben H**, CASTELAIN Vincent, LACHICHE Nicolas. "Predicting and Interpreting Healthcare Trajectories from Irregularly Collected Sequential Patient Data Using AMITA.", Under Review(Information Sciences Journal).

I furthermore declare that I have properly acknowledged and cited all sources, including publications, articles, books, and any other materials used or consulted in the preparation of this thesis. All contributions from other individuals or organisations have been duly acknowledged.

I take full responsibility for the accuracy and originality of the content presented in this thesis. Any assistance received in terms of technical, intellectual, or financial support has been duly acknowledged. I understand that any act of plagiarism or academic dishonesty is a serious offence and may result in severe consequences.

I hereby grant the University of Strasbourg the non-exclusive right to reproduce and distribute copies of this thesis, either in print or electronic format, for scholarly purposes.

October 28, 2024

Acknowledgments

Research, and the journey of writing a thesis in particular, is a deeply collective endeavor. While the academic work itself is shaped by the guidance of mentors and collaborators, it is also influenced by countless smaller moments—the discussions over coffee, train rides to conferences, inspiration drawn from readings, insightful questions from students, and the friendships formed along the way.

As I bring this chapter of my life to a close, representing three years of intense research, I begin with the traditional yet sincere acknowledgments. I am keenly aware that this section may end up being one of the most widely read parts of my thesis, yet paradoxically, one of the least revisited! Therefore, I apologize in advance if anyone feels overlooked.

First and foremost, I am deeply thankful to the Almighty ALLAH for His boundless and immeasurable mercy, which has guided me through every step of my life. Next, I owe an enormous debt of gratitude to the exceptional individuals who have supported and guided me throughout this journey.

My deepest thanks go to Professor Nicolas Lachiche and Professor Vincent Castelain, who served as my supervisors. I consider myself incredibly fortunate to have had the opportunity to work under your guidance. I thank you both for introducing me to this fascinating field of research, for your continuous encouragement, and for your involvement and support at every stage of my work. Our discussions, your feedback, and your patience have been invaluable to my development as a researcher.

A special note of gratitude goes to Professor Nicolas Lachiche, my primary thesis supervisor. From my first internship with you in 2020 in Vietnam, through to this very moment in 2024, I owe so much to you. I still remember receiving that email from you, asking if I would like to collaborate on a research project, a turning point that opened doors to new and exciting research paths. You introduced me to engaging topics, encouraged me to explore new perspectives, and above all, instilled in me the values of scientific rigor and clarity. Through your mentorship, I learned how to articulate my ideas, present my research coherently, and approach every challenge with critical thinking.

I would also like to sincerely thank the members of my thesis jury. I am honored and humbled by your participation in this process, and to be honest, slightly in awe of your expertise. Special thanks to Christine Sinoquet and Eric-André Sauleau for your thorough and insightful reports. Your broad perspectives and practical insights have been an invaluable guide throughout this journey.

Lastly, but certainly not least, I want to express my profound gratitude to my family and friends. This thesis would not have been possible without your unwavering support and encouragement, despite the physical distance. To my mothers, Fanta, Adam, and Hawa thank you for your endless love, care, and concern throughout every stage of my life. To my entire family, and especially to my sister Fatoumata and my wife Rokia, thank you for your love and steadfast support. And to my dear friends, who have stood by me through all the adventures and challenges along the way, I am deeply grateful.

This work is as much a testament to your love and support as it is to my own efforts. Thank you all.

Dedicated to my beloved grandma Djénéba !

Contents

List of Figures

List of Tables

Part I

Introduction & ICU data's validation

Chapter 1

Introduction

" Will finds, freedom chooses. To find and to choose is to think !!! "

– Victor Hugo

This chapter presents an overview of the Thesis. First of all, the motivation behind the research developed is presented. Next, the main and specific objectives are detailed as well as the methodology proposed to achieve them. Finally, the structure of the document is presented.

1.1 Motivation

The rapid advancements in Information and Communication Technologies (ICT) have significantly transformed the healthcare sector, particularly through the conversion of medical records into Electronic Health Records (EHR). This shift has resulted in a substantial increase in data production and collection . The widespread adoption of EHR-based systems offers substantial opportunities for health research, especially in developing data-driven approaches like Machine Learning (ML). These records comprehensively document all interactions between patients and healthcare providers throughout their hospital visits or stays.

EHRs encompass three types of data: structured data (e.g., patient age, admission date, measurements, and medical codes), semi-unstructured data (e.g., doctors' comments and descriptions of non-standardized conditions), and unstructured data (e.g., narrative clinical notes documenting patient conditions, family history, diagnoses, procedures, and medications). The primary function of EHRs is to provide current and pertinent patient information quickly, aiding medical practitioners in delivering higher-quality care through seamless information exchange. These centralized systems often compile extensive databases of patient records over several years, providing a valuable resource for statistical analysis and bridging the gap between medical analysis and machine learning techniques, which often require large volumes of data to perform effectively. Many research studies utilize this data for predictive analysis, gaining insights into disease progression and developing health monitoring systems that enable doctors to provide enhanced care [61].

Machine Learning (ML), a specialized area within Artificial Intelligence, focuses on constructing models that autonomously learn from data, discerning meaningful patterns without requiring explicit programming . In recent years, ML has revolutionized academia and industry, demonstrating remarkable achievements across various domains and often outperforming traditional methods . It excels in building powerful predictive models, such as neural networks, with impressive accuracy.

Inspired by advancements in other fields, ML, particularly deep learning models, has found extensive application in clinical tasks like diagnostics, prediction of clinical events, disease progression analysis, and forecasting future hospitalization [39, 49, 96]. ML enhances patient care delivery and decision-making by aiding in early diagnosis, identifying patients at risk of complications, and improving treatment outcomes. These models can characterize clinical conditions, identify relevant patterns, and forecast health status evolution [132], offering significant opportunities for diagnosing conditions like diabetes and hypertension using clinical data and supporting healthcare practitioners in knowledge extraction and clinical decision-making [178].

Furthermore, the establishment of extensive private data repositories and open-source medical databases like MIMIC III [70] and eICU [125] has empowered researchers to tackle clinical prediction problems. These efforts include risk prediction, intervention recommendation, disease progression, and patient sub-typing [5, 19, 22, 87, 124]. Some methods combine various types of input data, including tabular, textual, longitudinal, and image data, revealing that comprehensive patient representations lead to improved outcomes [133].

Despite the advantages of machine learning (ML) in clinical research, extracting insights from Electronic Health Record (EHRs) data remains challenging due to their complexity and heterogeneity. EHRs data exhibit temporal dependencies, sparsity, and high dimensionality, which can negatively impact ML model performance and complicate data processing, analysis, and visualization. Addressing these challenges necessitates the development of efficient and capable ML models.

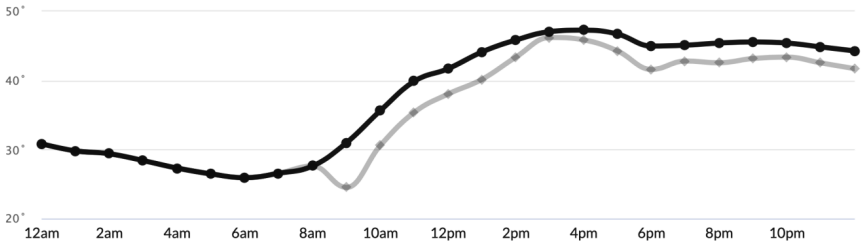
1.1.1 Machine learning solutions

Machine learning is central to the ongoing AI revolution, impacting various domains significantly. Contemporary machine learning involves analyzing vast datasets to identify patterns and improve performance. A key challenge is managing the complex and diverse structures in real-world data. Effective machine learning must accurately represent these structures in its models. For instance, in healthcare, it's crucial to understand how treatments and hospital admissions over time affect a patient's health. In graph classification, the focus is on capturing both substructures and the overall structure. Traditional methods rely on feature engineering, which can be insufficient and resource-intensive.

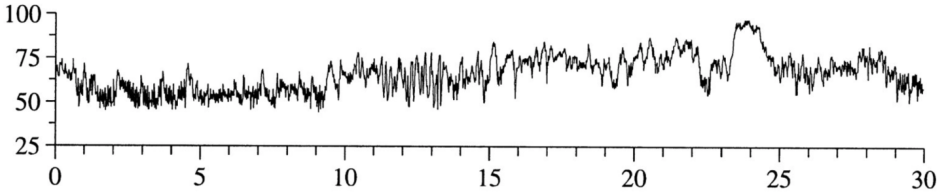
Deep learning, particularly neural networks, alleviates this by learning features directly from data [86]. Recent advancements, along with increased computational power, have led to breakthroughs in fields like computer vision, signal processing, and text modeling [3, 46, 78]. Deep neural networks, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are flexible and can model various data structures. For example, CNNs capture translation invariance in images, while RNNs handle temporal dependencies in sequences. When combined with Conditional Random Fields (CRFs), deep neural networks can model structured outputs, making them adept at capturing structural information.

This thesis focuses on RNNs [138], significant for their self-loop connections and shared parameters across time steps, making them robust for sequential data and temporal dependencies. RNNs are universal approximators, capable of modeling complex functions through iterative estimation, and can replace Feedforward Neural Networks (FNNs) for vector data modeling. They can be deepened without adding parameters, thus preventing overfitting. Recent advancements in RNNs, such as Gated RNNs [27, 63] and attention mechanisms [32, 80], have improved their ability to model long-term dependencies, enhancing their suitability for tasks like machine translation and question answering [3].

Nonetheless, RNNs may not adequately model complex data structures and relationships. Sequential models, like RNNs, often assume regularly sampled sequential data, such as word sequences or visual frames in videos. This assumption falls short in handling event data such as Electronic Health Records (EHRs), which are sequential in time but exhibit irregular events with intricate information. EHRs encapsulate a patient's hospital admissions over time, resulting in episodic, irregularly sampled data where the time intervals between admissions vary, signifying shifts in the patient's health status. Interventions, such as medical procedures and medications, can alter the trajectory of health, necessitating a distinct modeling approach compared to regular clinical observations like diseases or



Temperature time series (actual and "feels like") of a day in Pittsburgh, PA.



Heart rate time series from a bedside monitoring device (30 minutes). Source: [59]

Figure 1.1: Examples of regular time series which are generated from devices with automated observations at a predefined frequency [43].

lab tests. Therefore new models based on recurrent neural networks need to be proposed to address the above challenges.

1.1.2 Types of time series

Time series data can take the form of observations made at consistent time intervals, indicating a fixed frequency. For instance, consider the hourly temperature readings from the weather station, or the recording of an electrocardiography (ECG) signal in as shown in Fig. 1.1. In both cases, the measurements represent real-valued values. This category of time series is known as regular time series, which typically stems from a device with automated data collection set at a predefined frequency.

However, not all time series adhere to regular time intervals. Take, for example, the measurements of certain physiological variables or lab tests conducted on a patient during the initial day of their hospital admission, as illustrated in Fig. 1.2. Each observation is registered only when the relevant measurement event takes place, such as when a physician orders a lab test to evaluate the patient's underlying physiological conditions. This variant of time series is known as event time series, recording discrete events with or without associated values. Unlike regular time series, event time series may feature irregular time gaps between successive data points.

Another instance of such a time series involves multiple measurements or signals that track various facets of a patient's condition over time, as depicted in Fig. 1.2. Each entry signifies the incidence of a particular type of clinical event, be it the administration of medication, a lab test order, or a medical procedure. This type of data captures discrete events and their associated information, forming an event time series with irregular time intervals between observations.

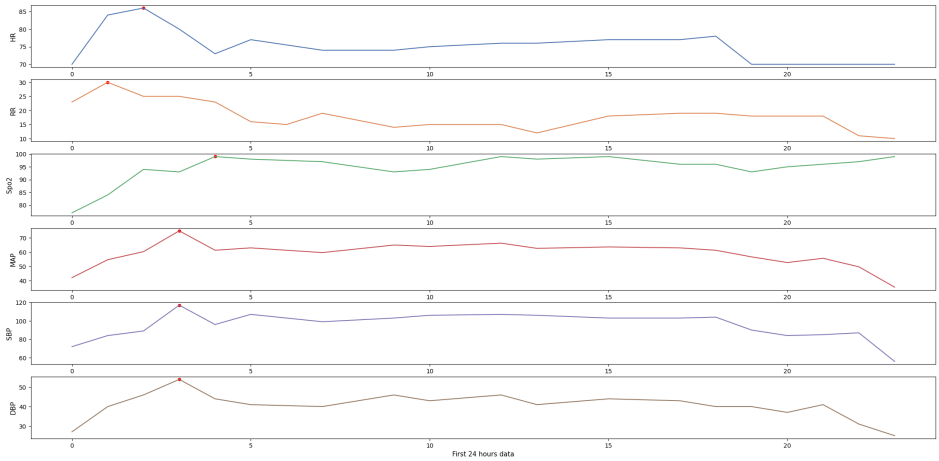


Figure 1.2: Examples of event time series which record occurrences of discrete events. Time gaps between two consecutive data points can be irregular.

1.1.3 Time series in healthcare

Healthcare is a domain that generates a wealth of intriguing yet intricate time series data. Our capacity to analyze this data and develop machine learning solutions based on it holds paramount importance, as it directly impacts patient care and, consequently, the physical and mental well-being of patients. Within healthcare settings, a diverse array of time series data is generated. This includes data originating from various clinical and health-assisting devices, such as bedside monitoring systems [76], smart healthcare solutions leveraging smart devices [14], and the Internet of Things (IoT) [53].

In this thesis, our primary focus lies in event time series data derived from Electronic Health Records (EHR), which constitute a comprehensive repository of patient-related measurements, observations, and treatments. These records provide insights into patient conditions, their management, and the dynamic nature of their healthcare journey. The core objective of this thesis revolves around addressing medical challenges, primarily in the form of prediction tasks, using this event time series data. These tasks encompass predicting patient outcomes, such as mortality , readmission or length of stay etc..

1.1.4 Tasks defined upon time series

Recent strides in data acquisition and processing technologies have facilitated the accumulation of vast time series datasets across various domains. For instance, in manufacturing, factories often employ multiple sensors to monitor the production process, yielding hundreds of thousands of sensor readings daily. In finance, time series data emerges from financial transactions and fluctuations in stock market prices worldwide. Healthcare sees the generation of substantial clinical time series data through patient electronic health records, wearable devices, and continuous monitoring equipment. Additionally, in transportation, GPS tracking systems offer real-time time series data detailing the locations of vehicles and individuals.

These extensive collections of time series data present new opportunities to address critical tasks within their respective domains. These tasks encompass monitoring dynamic systems, detecting anomalies or malfunctions, and predicting future states of systems. Below, we outline four fundamental tasks or challenges in the context of time series:

(1) Time Series Classification: This task involves assigning a label (category) to a time series, enabling us to differentiate it from similar time series. The label may signify a distinct pattern or condition associated with the time series [66], or it may represent an interpretation of a heart condition from an ECG time series signal.

(2) Time Series Clustering: Here, the objective is to group individual time series based on their similarities within a set of multiple time series. Similarity metrics can be derived from trends in observed values, patterns of event occurrences, or the types of observed events. For instance, with biomarker time series data from Parkinson's Disease patients, we can identify cohorts of patients exhibiting similar disease progression patterns [170]. These patient time series clusters can be invaluable in designing tailored treatment plans for disease subtypes with heterogeneous characteristics.

(3) Time Series Forecasting: This task aims to predict future values of a time series based on past observations and their historical trends. Accurate forecasting hinges on modeling both the overall trajectories of a time series and the dependencies between future values and recent changes in the time series. This is a crucial area of research with broad applications across domains, such as stock price prediction in the financial industry [88] and predicting future lab measurements in preventive healthcare.

(4) Event Prediction: Event prediction seeks to forecast the occurrence of the next event and associated information, such as event type and timing. While closely related to time series forecasting, event prediction focuses on discrete event occurrences and models the dependencies between them. Examples include modeling user activity events in online recommendation systems or events in law enforcement operations [148], and early warning systems in hospitals for forecasting adverse clinical events like septic shock [119].

1.2 EHR data & Challenges

EHR systems are invaluable for enhancing patient care, containing comprehensive medical information like diagnoses, treatments, and outcomes collected from various hospital sensors. This diverse data, stored without a spatial or strict sequential structure, includes symptoms, medication records, lab tests, procedures, physiological signals, and clinical notes authored by physicians.

In the context of event time series data within EHRs, each entry represents a significant clinical event (e.g., medication order) with detailed attributes such as (timestamp, type, item, and value). These events collectively form a complex multivariate event time series, crucial for understanding patient care dynamics, as depicted in Fig. 1.3, which illustrates the patient's clinical care history from EHRs records. The history is represented as multivariate

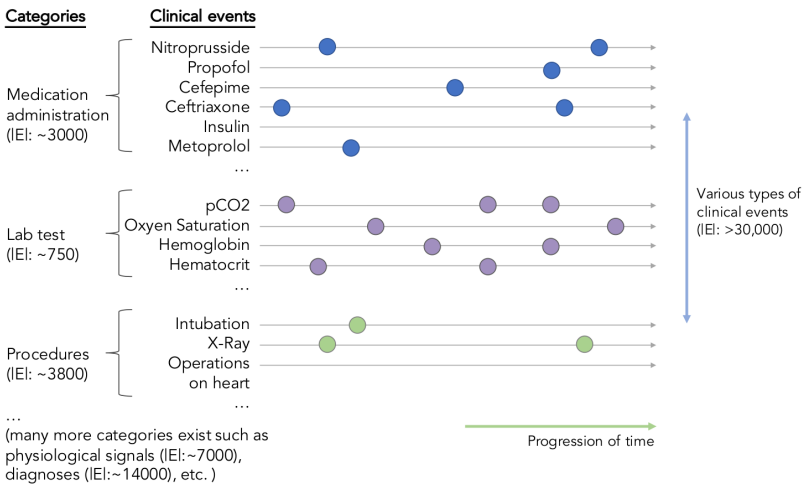


Figure 1.3: A circle on time-axis corresponds to an occurrence of a clinical event. The numbers of clinical events in each category are counted from MIMIC-III [90].

event time series. Machine learning models applied to EHR data can uncover temporal relationships between clinical events, enabling better patient management and proactive intervention to prevent adverse outcomes, thereby enhancing overall care quality [16, 69, 162].

However, constructing machine learning models from EHR data presents several challenges owing to its distinctive characteristics. Effectively modeling medical predictions requires careful consideration of the challenges involved in processing the highly diverse observational data found in real-world clinical databases. These challenges can be categorized into six(6) key areas:

(1) **High dimensionality:** EHR data is inherently high-dimensional due to the numerous types of clinical events and diagnoses that can occur during the patient's hospitalization. The vast number of event types and clinical concepts increases the complexity of modeling, as it requires from the model to learn and maintain knowledge about each event type. This problem of high dimensionality is often coined as the curse of dimensionality. Additionally, predicting future events involves learning complex dependencies between a future event and a series of preceding events, which can be challenging to enumerate and learn [173].

(2) **Missing and Irregular Observations:** Clinical events in EHR data are often not universally observed among patients due to their association with specific diseases or conditions. For instance, events like insulin administration are typically only recorded for patients with diabetes, leading to sparse data. This sparsity poses challenges for training robust models that generalize well across all event types and unseen patients. In databases like MIMIC-III[1], which catalog over 30,000 clinical event types, the average occurrences per type (e.g., medication administration 10.1, lab tests

[1]We computed the average counts from tables (inputevents-mv, labevents, procedureevents-mv)

7.3, procedures 1.5) are notably low. Additionally, irregular intervals between observations further complicate the analysis of EHR data, which primarily serves the purpose of patient information storage rather than clinical research. This routine data collection introduces variability in data completeness over time, resulting in missing information challenges [2, 40].

Moreover, EHR data often contain outliers [13] that undermine data consistency and accuracy. Poor data quality can lead to unreliable outcomes in algorithms a principle encapsulated by the adage "garbage in, garbage out" in Data Science and Machine Learning. Hence, rigorous data preprocessing is essential for ensuring the reliability of machine learning tools.

(3) **Patient Variability:** EHR data reflects the heterogeneity of patient sequences across different patient populations. Each patient may have a unique combination of clinical complications, medication regimes, and observed sequence dynamics. While average behaviors can be captured by a single model, it may struggle to represent the detailed dynamics of individual patient sequences.

(4) **Temporality:** EHR data is inherently longitudinal, with a set of patient visits creating multiple time series. The order and time between these visits are key and contain important information for understanding the health status evolution(progression of the disease) and patient trajectory over the care period and extracting appropriate clinical knowledge. Clinical events are irregularly sampled. Both the order of clinical events and the time difference between events are valuable pieces of information for learning prediction models. However, encoding temporal information into predictive modeling is a relatively new challenge and lacks established methodologies and is hampering the application of conventional ML methods.

(5) **Sizes of Target Groups (Data Bias):** In clinical machine learning applications, sufficient patient data is vital for training accurate predictive models and gaining meaningful insights. However, scarcity of records, especially for specific clinical events or conditions, frequently leads to Class Imbalance Problems (CIP) in healthcare data analytics. CIP arises when certain classes, such as rare clinical events, are significantly underrepresented compared to others [55].

Traditional prediction algorithms assume balanced class distribution in training sets. In imbalanced datasets, this assumption results in a bias towards the majority class, neglecting adequate learning of the minority class distribution [17]. This issue is particularly critical in healthcare, where accurately predicting minority class events is often more important. For instance, in clinical scenarios, a false-negative HIV test before renal dialysis can have severe consequences. To mitigate this challenge, effective prediction models prioritize achieving high precision for the minority class while maintaining reasonable precision for the majority class. This strategy may involve adjusting the classifier's decision boundary to enhance sensitivity or recall, even if it reduces precision.

(6) **Security and privacy:** Clinical data is highly sensitive and their utilization is limited by privacy restrictions, regulations, and organizational guidelines [6]. EHRs are

regulated by laws protecting patients' privacy such as the Health Insurance Portability and Accountability Act in the US and the General Data Protection Regulation in the European Union [165]. Sharing and working with clinical data requires approval from an institutional board, committee, or health organization, which is restrictive and time-consuming, limiting the access and hampering the research and the setting up of innovative studies [6].

In summary, EHR data presents unique challenges due to its irregular, diverse, and high-dimensional nature. Each patient's care history is recorded at varying intervals, influenced by health status and local practices, leading to diverse health trajectories. Integrating these various data types into a unified model requires a specialized approach, capable of extracting essential information while avoiding redundancy.

1.3 Clinical event time series prediction

In this section, we explore predicting clinical event time series, emphasizing the critical role of learning a patient's state representation for effective prediction.

Clinical event time series involve events unfolding continuously. In statistical terms, these are often modeled as temporal point processes, where basic point processes denote individual event occurrences over time, and marked point processes assign values to each event instance [67, 83]. In cases of categorical event values, they form multivariate event processes, where each category establishes its own basic point process [98]. In EHR data, clinical events like medication administration are represented within these time series, generating multivariate event sequences for each patient.

Using these multivariate event time series from EHR data, we frame the event prediction task as follows: given the complete history of events up to time t (denoted as $y_{[1:t]}$), the objective is to predict the occurrence of the next event (y_{t+1}). In continuous-time prediction, this typically involves modeling an intensity function for the point process, often using models like Hawkes processes [135, 163] or their variations [98, 112].

Alternatively, discrete-time prediction converts the continuous event time series into discrete time intervals, typically using non-overlapping moving windows (e.g., 6 hours). Each event type occurrence within a window is represented as a binary indicator value. This transformation results in a large, sparse binary matrix, where rows represent event types and columns denote segmented time intervals during hospitalization, akin to the depiction in Fig. 1.6, illustrating the representation patient's records as a sparse matrix from MIMIC-III [70]. Likewise, the prediction task can be defined based on sequences of column-vectors within the matrix, as demonstrated in Fig. 1.4.

1.3.1 Patient state representation for clinical event prediction

A key challenge in creating event prediction models from Electronic Health Records (EHRs) is condensing a patient's historical health data into a relevant format for predicting future

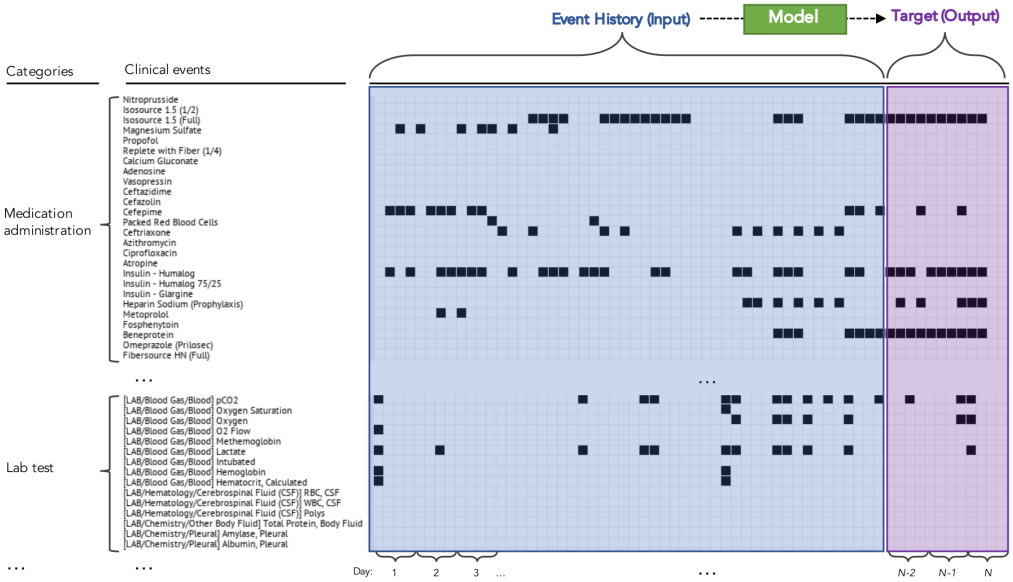


Figure 1.4: Prediction task defined over the multivariate clinical event time series introduced in Fig. 1.6. Given full event history (blue box), the goal is to predict occurrences of each events in future window (purple box) [90].

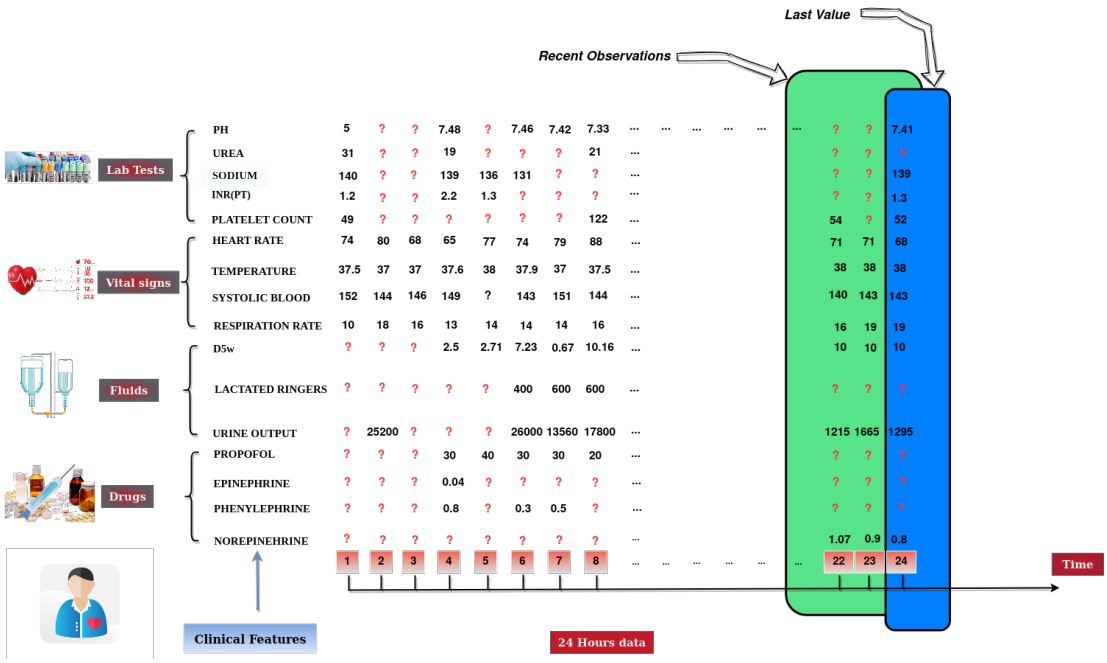


Figure 1.5: Irregular sampling of physiological variables (EHR data)

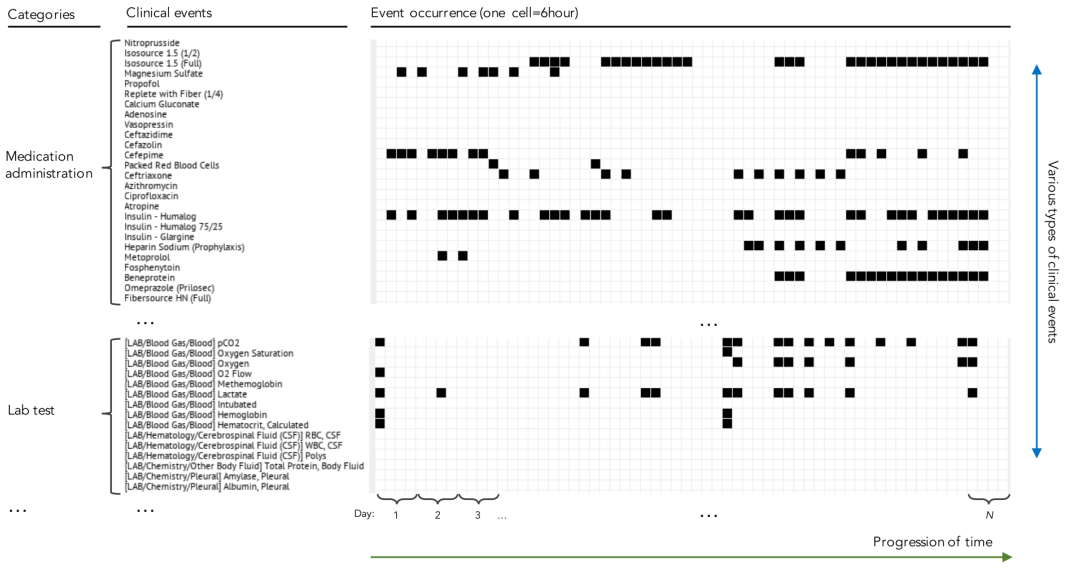


Figure 1.6: Rows correspond to different clinical events and columns correspond to time. Each cell (bin) indicates occurrence or non-occurrence of an event during a time-window (e.g., 6 hours) [90].

medical events. The "patient state" refers to this essential, concise summary of a patient's history, critical for accurate predictions. Developing efficient methods to generate these patient states is crucial for building robust clinical event prediction models. The following sections will overview existing approaches for defining patient states and prediction models for EHR-derived multivariate event time series.

1.3.1.1 Recent observations

A straightforward way to define a patient's state is by focusing on the most recent observations. For example, when given a patient's chronological event history from admission to the present, we typically consider only the observations from the last 6 hours, ignoring earlier ones as shown in Fig. 1.5. This method is effective because recent observations are more likely to predict near-future events and is computationally efficient since it processes only a small subset of data, reducing the computational load.

Additionally, this method is compatible with various standard classification algorithms like Support Vector Machines (SVM), Naive Bayes classifiers, decision trees, and neural networks. By using the most recent observations for each event type, we can create a fixed-sized vector for these algorithms to predict future events. However, there are drawbacks:

(1) This approach may prevent the model from learning long-term trends or dependencies from the past, as a patient's condition can change rapidly, making recent observations less reflective of their current state.

(2) It may miss important information about events that haven't occurred recently, focusing only on recent data and potentially overlooking crucial insights from earlier occurrences.

1.3.1.2 Last value carry forward (LOCF)

This method improves upon the previous approach by copying the most recent value for each event type and using it as the current patient state, as shown in Fig. 1.5. Despite its simplicity, this technique has proven effective and popular in handling missing data in clinical and medical studies involving longitudinal data [51, 116]. Like the previous method, it is compatible with various classification techniques for predicting future events.

However, it's important to note that even though this method improves on the issue of missing data. This approach still has limitations in modeling complex trends or long-term event dependencies, as it relies solely on the most recent observations for predictions.

1.3.1.3 Temporal templates

This method addresses the aforementioned challenges by employing pre-defined temporal templates for individual time series and their combinations. Briefly, the temporal template approach transforms complex multivariate clinical time series, whether discrete or real-valued, spanning extended time periods, into fixed-sized vector representations. The core concept here is to establish a collection of feature functions, also known as feature templates, designed to map time series that encompass clinical variables into fixed-size vectors and their amalgamations. Examples of these feature functions include event-type-specific summary statistics such as minimum, maximum, or average observations over defined time windows (e.g., the first 6, 12, 24, or 36 hours) for real-valued time series, or counts of event occurrences for discrete event time series as shown in Fig. 1.7. Due to its ability to offer a more comprehensive summary of clinical time series data, many early efforts in predicting clinical events from Electronic Health Records (EHRs) relied on the templates approach. The fixed-size feature vectors generated by these templates are then input into various classification algorithms to make predictions for the next event. This approach has proven successful in a range of EHR prediction tasks such as outlier detection challenges [54].

Nevertheless, the primary drawback of this approach lies in the need to define the temporal templates and the information they represent in advance. Additionally, the number of potential features generated through these methods can be quite substantial. One solution to mitigate the necessity of a priori template definition is to employ predictive patterns extracted directly from data using frequent data mining techniques [4].

1.3.1.4 Probabilistic latent state-space models

Recent research has shifted towards defining patient states and making predictions using diverse probabilistic latent state-space models, including hidden Markov models and linear

Features

Hours	BILIRUBIN	UREA	SYSTOLIC BLOOD	TEMPERATURE	HEART RATE	LACTATE	.....	SpO2	PLATELET COUNT
1	0.2	32	113	37	114	1.2	.	96	128
2	NM	NM	110	NM	113	NM	.	99	250
3	1.2	NM	110	38	108	1.1	.	100	NM
4	NM	15	113	NM	110	NM	.	100	344
5	1.2	16	116	NM	102	0.9	.	100	344
6	NM	NM	102	37	108	NM	.	NM	NM
7	NM	NM	103	36	104	2.4	.	99	NM
8	NM	NM	106	NM	93	NM	.	97	NM
9	NM	NM	97	36	88	2.8	.	96	270
10	0.8	16	94	NM	88	NM	.	96	NM
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
24	NM	28	115	35	97	NM		99	290

Figure 1.7: Temporal representation of a patient during the first 24 hours following his/her admission to the intensive care unit (ICU).

dynamical systems [99], Gaussian processes [82, 140], or combination thereof [100]. This approach offers greater flexibility in capturing the intricate dynamics of clinical time series by employing a shared latent state-space, defined through an autoregressive function of a prior latent state and a recent observation. This framework has the advantage of more efficiently representing correlated observations within the latent space. However, it's worth noting that a limitation of probabilistic models arises from the fact that the behavior and expressiveness of the latent state-space are constrained by a specific (predefined) probabilistic distribution, such as Gaussian, Bernoulli, or Weibull distributions, which may not perfectly align with the observed data.

1.3.1.5 Modern neural-based models

Recently, significant progress in modern latent embedding and deep learning models has led to the emergence of low-dimensional latent state representations that demonstrate strong predictive performance across various tasks. Notable examples of these advancements include the utilization of real-valued vector-based representation methods like Skipgram and CBOW for modeling patient states [36, 108]. Additionally, hidden state-space models have been developed based on recurrent neural networks (RNN) [5, 22, 24, 124]. There are also non-recurrent models, including those utilizing attention mechanisms, convolutional neural networks (CNN), and Transformers [23, 94, 147].

These modern approaches do not rely on assuming a specific probabilistic distribution for generating the hidden state space. Instead, they adopt a data-driven strategy to learn the mapping from input to hidden state, and ultimately to output, employing a series of linear

transformations (matrix multiplications) and non-linear activation functions like sigmoid or tangent hyperbolic. As a result, these models tend to be more flexible (as they do not assume a specific distribution form) and are better equipped to capture non-linearities inherent in complex EHR-derived time series data compared to the probabilistic latent-space approaches.

In this thesis, we further develop and investigate models based on modern neural-based temporal methods, particularly Long Short-Term Memory (LSTM) models [63].

1.4 Research goals & Hypotheses

To effectively tackle the challenges outlined in Section 1.2, it is imperative to develop specialized machine learning models capable of navigating the unique complexities inherent in EHR data. Challenges such as high dimensionality, irregular observations, patient variability, temporality, data biases, and stringent security regulations present significant obstacles for traditional machine learning approaches.

The research goals presented in this section focus on creating innovative algorithms and frameworks, specifically multimodal deep learning models, that not only capture the intricate dynamics within EHR data but also integrate temporal information to more accurately represent patients' health trajectories over time. This research aims to bridge the gap between the complex, unstructured nature of EHR data and the demand for precise, actionable clinical predictions, ultimately enhancing patient care and outcomes.

Moreover, this research intends to validate these machine learning models through real-world clinical applications using two EHR datasets: MIMIC III [70] and eICU [125]. This includes developing strategies to handle time irregularities, manage the heterogeneity of clinical data, and ensure the models' robustness against challenges like missing data and patient variability. In summary, the research goals are directly linked to the challenges posed by EHR data. By crafting advanced machine learning methods tailored to these challenges, this research aspires to enable more accurate predictions of patient outcomes, thereby contributing to improved healthcare delivery and patient management.

1.4.1 Hypotheses

The research questions in this thesis are presented from a clinical and data science perspective. From a clinical perspective, the thesis aims to answer:

- **What is the probability of a given outcome for a patient at a specific time given data available at that time?**

In this thesis, we define outcome risk as mortality risk and length of stay of patients at a given time. Both tasks are considered as a proxy by authorities (like the Centre for Medicare and Medicaid Services in the US) to measure care quality and hospital reimbursements [77].

To achieve the presented objectives, this work addresses the following research questions:

- How can heterogeneous clinical data be represented for EHR-driven prediction models?
  - A patient record in an EHR is described by a sequence of visits and each visit includes demographics and clinical information of the patient. Demographics include patient age, gender, visit type, etc. The clinical information includes concepts like diagnoses, procedures, medications, labs, and more. Here, the goal is to combine and represent the information about each visit in a machine-friendly way to enable the learning of predictive models. Put differently, we learn numerical representations of patient visits.
- How can temporal information be incorporated in EHR driven prediction models?
  - Temporal dynamics in EHRs refer to the sequence of irregular recording of clinical features in this thesis. In other words, we move from visit to patient representation now. Of note, we do not consider the actual time gaps between visits at this stage but rather the time gaps between events for each clinical feature. The goal, here, is to capture the order of patient events in the EHR and include this information when building prediction models.
- How to model irregular time observed in patients' health trajectories?
- Is it possible to design a generic framework for irregular medical time series modeling?
- How does the multimodal architecture impact the performance of clinical prediction tasks?

These research questions explore the different ways to represent the patient's timeline using neural networks and measure their impact on real-world medical data, considering the irregular recording of events.

1.5 Contributions

The significance of this thesis is organised around two central lines of work: (i) introducing novel RNN architectures to model different structured data types and (ii) applying such models to a wide range of practical problems in healthcare, time series forecasting. In particular, our contributions are:

(1) In the initial phase of this thesis, we conduct an in-depth examination centered around pivotal variables. Our objective is to assess their prognostic significance and identify predictive factors crucial for effective treatment in ICU mortality scenarios. This investigation serves to validate the findings outlined in the work by Hugerot and al [127].

(2) The second part of this thesis work addresses the modeling of irregular timestamps in the clinical event time series. The main resulting contribution is the implementation of a generic framework for irregular event time series and evaluate the effect of different imputation techniques on the model's performance. The framework processes numerical and categorical medical events and supports the patient's metadata. Besides, it gathers state-of-the-art sequential deep learning models and time representation techniques. Using this framework, we conducted an empirical study of mortality and length of stay based on two real-world datasets and make a comparative study with current state of the art temporal neural-based approaches.

(3) Thirdly, we propose a novel deep learning model called MWTA operating on Electronic Health Records (EHRs), in which a patient's health history is represented as a sequence of events for each clinical feature. We extend the Long Short-Term Memory unit, a variant of RNNs to handle irregularly timed events by modelling the elapsed time between two consecutive records in the memory forgetting. MWTA also explicitly models the interaction between disease progression and interventions (e.g., medical treatments), in which the interventions change the course of the illness and shape future medical risks. We demonstrate that MWTA is effective for different learning tasks on datasets.

(4) Fourthly, we propose a novel adaptive pooling strategy called (ASP) that specifically targets and resolves outlier issues that can potentially occur during the analysis of EHR data. The incorporation of this feature enhances the robustness and reliability of our framework.

(5) Fifthly, We introduce a novel framework called Adaptive Multi-Way Interpretable Time-Aware LSTM (AMITA), which effectively handles timing irregularities, and is capable of capturing the complex interactions between different clinical features at different stages. AMITA extends the standard LSTM $gates$ (forget, input, and output) by using the time interval between events, the frequency of measurements and the contextual information from the patient's history. This extension enables the adjustment of the memory cell, diminishing the impact of previous memory as the elapsed time between events increases while retaining the enduring impact of earlier events. Additionally, we incorporate two gate mechanisms to effectively reflect intervention effects on current illness states.

- In healthcare, patient data often arrives irregularly due to varying clinical protocols and condition monitoring. We've enhanced the AMITA forget gate mechanism to account for the timing and frequency of events and the patient's historical context. This adaptation allows the model to adjust to each patient's unique temporal patterns, promoting focus on critical data while minimizing the impact of less relevant or sporadic measurements.

(6) Finally, We provide a high level of interpretability and how each clinical feature behaves within the patient's health course using aggregate functions over patient's medical history data.

This thesis is organized into two parts. The first part provides the related work that

motivated our contributions and is organized in two chapters. In the first chapter, we establish a survey on time-aware deep learning models for representing irregular clinical time series. A survey on neural based architectures specially Recurrent Neural Networks is detailed in the second chapter. On the other hand, the second part exposes our four main contributions. The first two chapters describe the implementation of MWTA-LSTM and AMITA architecture for irregular event time series that aims to combine the different levels of temporality and types of input data for building an accurate representation of the patient. The third chapter includes the studies that validate our proposed methods and use of real-world clinical databases and time series data, discussion of the results. Lastly, the fourth part concludes the thesis work, and introduces recommendations for future work.

Chapter 2

ICU- EHR data

"...the quality of my care and my confidence in its outcomes would never be better than the quality of the information behind them. The information in the room [ICU] that morning was detailed and exhaustive but was impossible to organize and interpret within the time required."

– William A. Knaus, *founding partner of APACHE*

In ICUs, physicians must monitor a vast array of variables, including physiological data, laboratory results, and device parameters. During a typical morning round, a physician faces an overwhelming amount of recorded data for each patient. This high-dimensional data surpasses the human capacity to process and address all observations effectively. As Miller noted in his study, even experienced physicians struggle to systematically solve problems involving more than seven variables [75].

The Intensive Care Unit (ICU) admits the most severely ill patients, providing specialized care like mechanical ventilation that cannot be administered elsewhere in the hospital. The NHS [110] recommends a one-to-one staff-to-patient ratio in the ICU to ensure continuous monitoring and immediate intervention, which studies show reduces mortality, hospital stay length, and complications [71, 128].

The ICU is data-rich, almost to an overwhelming extent. William A. Knaus, creator of the APACHE [75] system for predicting patient mortality, recounted his early experience in the ICU, where detailed but disorganized information made timely decision-making difficult. This experience highlighted the need for empirical models to assess patient severity, leading to the development of illness severity scores that are now used to estimate mortality probabilities such SAPS-II, APACHE and IGS II scores [41, 85, 75].

The digitalization of healthcare has led to the widespread adoption of Electronic Health Records (EHRs), which generate vast amounts of clinical data. By 2019, 90% of U.S. hospitals and 85% of Spanish health services had adopted EHR systems [29, 58, 60], driven by financial incentives like the HITECH [1] Act in the U.S. and the Innovative Medicines Initiative in the EU. EHRs are crucial for managing medical records, appointments, billing, and lab tests.

EHRs store comprehensive patient information, including demographics, medical history, diagnoses, and clinical notes. This data provides a complete picture of a patient's health and is invaluable for research. Researchers use EHR data to develop models for predicting patient outcomes, such as mortality and hospital readmission, and to identify early clinical events [25, 124, 143]. EHRs also support the shift toward personalized medicine by enabling proactive care and resource optimization. However, analyzing EHR data presents significant challenges in extracting clinical knowledge.

2.1 MIMIC & eICU databases

2.1.1 Ethics approval

This study was based on an analysis of publicly available, anonymized databases that already had institutional review board (IRB) approval. The Institutional Review Boards (IRB) of the Beth Israel Deaconess Medical Center in Boston and the Massachusetts Institute of Technology in Cambridge have granted their approval for the collection, processing, and release of data for the MIMIC-III database. The eICU research committee has also granted approval for data collection, processing, and release for the eICU database, which does not require Institutional Review Board approval. In full compliance with European Union data protection laws, all data were processed using the computational infrastructure available at Strasbourg University, thus, no additional approval was required.

In summary, the data contained in the records for both databases have been deidentified using data cleansing and date shifting. This was performed in accordance with the Health Insurance Portability and Accountability Act (HIPAA) [34]. Eighteen identifying data elements had to be removed to de-identify EHR data. The elements removed/shifted from the database include fields like name, dates, addresses, and telephone numbers. The database

is accessed through an application process. The applicant has to complete a training course on biomedical research and research conduct.

2.1.2 MIMIC III

The Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) III database[1] [70] is a publicly available ICU database sourced from the Beth Israel Deaconess medical center in Boston, Massachusetts. The data collected pertains to all aspects of a given patient's ICU stay, including medication, laboratory tests, bedside monitoring, chart recordings, clinical notes, discharge summaries, and patient outcomes. In compliance with the United States Health Insurance Portability and Accountability Act of 1996 (HIPAA), all Private Health Information (PHI) was removed from the dataset before release using bespoke open source software [118]. Only the guardians of the MIMIC II database possess the ability to map each individual back to the original patient's identifiable information. The MIMIC III clinical database has undergone many iterative updates, and the version used in this work is MIMIC III v2.6. The data is stored in a relational database whose schema is available from PhysioNet, the primary host of the MIMIC III database

MIMIC-III [70] is a publicly available database generated from de-identified real-world EHR data and contains medical records of about 46k critical care patients admitted in Beth Israel Deaconess Medical Center between 2001 and 2012. The database contains rich information about various medical events during the patient's stay. The features could be related to vital signs, medications, laboratory measurements, observations and notes written by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, and survival data. The database covers the health information of 38,597 adult patients and 49,785 hospital admissions. This dataset exhibits the typical challenges of any large-scale clinical data, including varying-length sequences, skewed distributions, missing values, episodic visits, and highly variable time intervals between successive visits.

2.1.3 eICU

The eICU Collaborative Research Database[2] [125] is a multi-center intensive care unit database with high granularity data for over 200,000 admissions to ICUs monitored by eICU programs across the United States. The eICU database comprises 200,859 patient unit encounters for 139,367 unique patients admitted between 2014 and 2015 to 208 hospitals located throughout the US.

The eICU is sourced from the eICU telehealth program, which is a telemedicine initiative providing continuous real-time monitoring and remote support to ICU clinicians. The database contains comprehensive records, including demographics, physiological readings from bedside monitors, diagnoses (International Classification of Diseases, Ninth Revision codes), and other clinical data collected during routine medical care.

[1] https://mimic.physionet.org
[2] https://eicu-crd.mit.edu/

2.2 Norepinephrine & Lactate correlation

To thoroughly gain a comprehensive understanding of Electronic Health Record (EHR) datasets, with a focus on the MIMIC-III database [70], we performed an in-depth analysis of key variables to evaluate their prognostic value for ICU mortality, particularly in patients suffering from acute circulatory failure. This analysis aims to corroborate and expand upon the findings of previous research by Hugerot et al. [127], providing a more nuanced perspective on the factors influencing ICU outcomes in these high-risk patients.

Acute circulatory failure, commonly known as shock, is a primary reason for admission to intensive care units (ICUs). Shock is characterized by a mismatch between arterial oxygen transport and tissue oxygen requirements, leading to cellular and tissue-level activation of anaerobic metabolism. This metabolic shift results in organ dysfunction, which can perpetuate the state of shock [159] and contribute to further organ failures . Prompt recognition and treatment of shock are critical to prevent multiple organ failure and, ultimately, death .

The objective of this study is to assess the prognostic impact on ICU mortality of the maximum dosage of norepinephrine administered within the first 24 hours of ICU admission, regardless of the indication. Additionally, this study aims to identify the factors associated with the success or futility of treatment in patients who required the highest doses of norepinephrine.

2.2.1 Materials & methods

In the Materials and Methods section, we provide a detailed account of the comprehensive strategies and methods utilized for data collection and analysis. The section initiates with an explicit delineation of the study design, highlighting the nature of the research conducted to furnish a robust framework for the investigation.

Subsequently, we delve into the criteria for participant selection, delineating both the inclusion and exclusion parameters. This approach ensures clarity and transparency regarding the demographic and clinical characteristics of the study cohort. Following this, we present a thorough breakdown of the data collection methodologies, elaborating on the selection and application of various tools, instruments, or questionnaires. The justification for these choices is also discussed, emphasizing their relevance and contribution to the research's reproducibility.

Moreover, The analytical strategies employed are thoroughly described, encompassing statistical tests, software used for data analysis, and the criteria for result interpretation. This allows for a clear understanding of how conclusions were drawn from the collected data.

2.2.1.1 Inclusion & exclusion criteria

In order to evaluate the impact on the prognosis of patients of the maximum dosage of Norepinephrine received during the first 24 hours after admission, we conducted a retrospective study with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 [70].

We used two sets of inclusion criteria to select the patients to prepare the cohort used in this study. First, we identified all the adult patients by using the age recorded at the time of ICU admission. Following previous studies [70], in our work, all the patients whose age was >15 years at the time of ICU admission is considered as an adult. The second criterion is to select all the patients for whom norepinephrine vasopressor support was prescribed and validated by a nurse within the first 24 hours of admission were included in the study. Excluded patients were those under 15 years old even if they received the norepinephrine within the first 24 hours of admission.

2.2.1.2 Data collection

For each of the included patients, most numerical, clinical, biological, and therapeutic parameters such as blood pressure, PaO2/FiO2 ratio, lactate, and the dosage of vasopressor support by norepinephrine support were collected including the vital signs, blood test results, blood gas analysis, (quantity of fluids, numbers of packed red blood cells units and or plasma, mechanical ventilation, need for vasopressors/inotropes), comorbidities, selected diagnoses, and intra-resuscitation and intra-hospital patient outcomes and Glasgow coma scale (GCS) score. The GCS score was one of the only numerical data collected by reading the patient's medical record. When sedation was initiated on the ward, the minimum GCS value recorded was the minimum GCS value on our ward before orotracheal intubation. When the patient was admitted and remained sedated during the first 24 hours in the intensive care unit, the minimum GCS score used was the GCS score prior to orotracheal intubation. If this was not recorded in the medical record, we considered the data as missing.

2.2.1.3 Statistical analysis

The statistical analysis of the data consisted of a descriptive and an inferential part. The descriptive statistical analysis of the quantitative variables was done by giving for each variable the position parameters (mean, median, minimum, maximum, 1st and 3rd quartiles) as well as the dispersion parameters (variance, standard deviation, range, interquartile range). The Gaussian character of the data was tested by the Shapiro-Wilk test. The description of the qualitative variables was done by giving the numbers and proportions of each modality in the sample and subgroups of the sample. sample. Inferential analysis for categorical variables was done either with a Chi-square test or with a Chi2 test or a Fisher's exact test, depending on the numbers.

For each variable, where relevant, the odds ratio and its 95% confidence interval were

estimated using the appropriate method (maximum likelihood estimation or small sample adjustment). Comparisons of quantitative variables between groups were performed either by a Student's t test when the variable of interest was Gaussian, or by a non-parametric Mann-Whitney test when it was not. For the study, we also produced Kaplan-Meier type survival curves according to different characteristics of our patients, and compared these curves by a Log-rank test (Mantel-Cox).

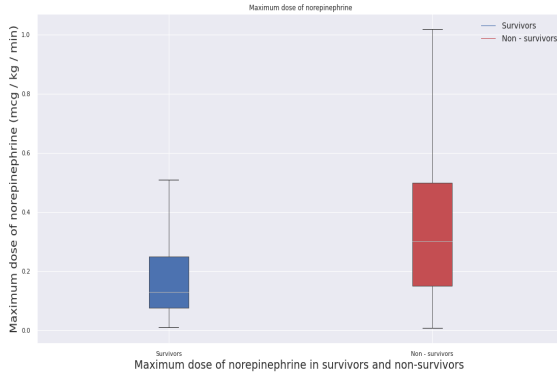
2.2.2 Cohort

Table 2.1: General description of the cohort

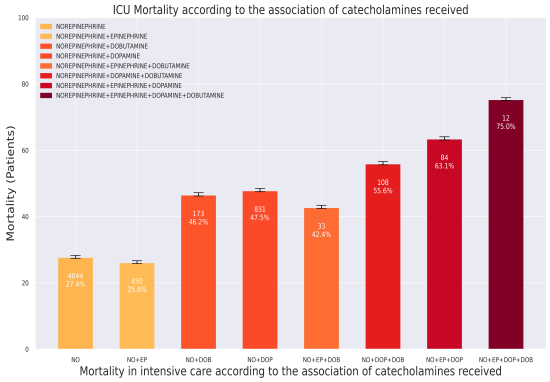
		Missing	Overall	Non-Survivors	Survivors	P-Value
Number of Patients			5735	1834	3901	
Survival ICU, mean (SD)		2010	248.1 (537.4)	6.3 (8.6)	482.5 (676.2)	<0.001
LOS, mean (SD)		2010	7.6 (9.2)	6.1 (8.6)	8.3 (9.4)	<0.001
AGE, mean (SD)		0	67.3 (15.6)	68.8 (16.0)	66.6 (15.3)	<0.001
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.2 [0.1,0.3]	0.3 [0.2,0.5]	0.1 [0.1,0.2]	<0.001
Epinephrine, mean (SD)		5156	0.2 (0.9)	0.3 (0.5)	0.1 (1.0)	0.021
Dopamine, mean (SD)		4700	16.0 (87.4)	20.7 (123.2)	11.3 (8.5)	0.082
Dobutamine, mean (SD)		5409	6.7 (5.0)	7.6 (5.9)	5.7 (3.6)	0.001
Lactate(mmol/L), mean (SD)		0	4.8 (4.0)	7.1 (5.3)	3.6 (2.6)	<0.001
Heart Rate, mean (SD)		1	118.4 (24.1)	125.2 (25.9)	115.3 (22.6)	<0.001
Systemic Blood Press(mmHg), mean (SD)		6	150.4 (25.9)	147.4 (28.5)	151.8 (24.4)	<0.001
Mean Blood Press, mean (SD)		0	44.6 (14.2)	38.5 (15.4)	47.5 (12.7)	<0.001
Temperature(°C), mean (SD)		151	37.8 (1.1)	37.6 (1.3)	37.9 (0.9)	<0.001
Urine Output, mean (SD)		404	2981.7 (2979.7)	1824.9 (3802.9)	3493.5 (2357.5)	<0.001
Urea(mmol/l), mean (SD)		2	15.3 (10.0)	17.7 (10.8)	14.2 (9.4)	<0.001
Bilirubin(mg/dL), mean (SD)		857	2.5 (5.2)	3.9 (7.3)	1.8 (3.5)	<0.001
PaO2/FiO2, mean (SD)		990	173.0 (187.1)	161.0 (201.4)	179.8 (178.2)	0.001
GCS Score, mean (SD)		243	7.3 (4.5)	6.1 (3.9)	7.9 (4.7)	<0.001
ICU Death, n (%)	0	0	3901 (68.0)		3901 (100.0)	<0.001
	1		1834 (32.0)	1834 (100.0)		
GENDER, n (%)	F	0	2466 (43.0)	803 (43.8)	1663 (42.6)	0.427
	M		3269 (57.0)	1031 (56.2)	2238 (57.4)	
IGS II, mean (SD)		0	54.3 (16.7)	63.8 (14.7)	49.7 (15.7)	<0.001
SOFA, mean (SD)		0	10.7 (3.7)	12.5 (3.7)	9.8 (3.3)	<0.001

A total of 5230 patients, corresponding to 5735 distinct ICU stays, met the inclusion criteria for norepinephrine prescription within 24 hours of ICU admission. The patient characteristics are detailed in Table 2.1.

The average patient age was 67 ± 15.6 years, with a male predominance (sex ratio 1.3). The main comorbidities were arterial hypertension (52%), chronic heart failure (23%), and chronic respiratory disease (21%). Chronic renal failure and chronic liver disease affected 14% and 13% of patients, respectively. Additionally, 18% had active solid cancer, 13% had hematological malignancies, and 20% were immunocompromised. The mean SOFA score on admission was 10.7 ± 3.7. Invasive mechanical ventilation and extra-renal purification were required in 72% of patients. The lowest mean arterial pressure in the first 24 hours was 44.6 ± 14 mmHg, and the average arterial lactate was 4.8 ± 4.0 mmol/l. The mean IGS II score was 54.3 ± 16.7. The average maximum norepinephrine dose in the first 24 hours was 0.2 µg/kg/min, with an interquartile range of [0.1, 0.3] µg/kg/min. Additionally, 6% of patients received dobutamine, and 11% received continuous intravenous adrenaline.



(a) Max dose of norepinephrine in Survivors VS Non-survivors.



(b) ICU mortality according to the combination of catecholamines received.

Figure 2.1: Vasopressors in Survivors vs Non-survivors.

Patients received an average vascular filling of 16 ± 19 ml/kg and a total intake (excluding labile blood products) of 38 ± 26 ml/kg over 24 hours. Twenty percent of patients were transfused with packed red blood cells. The mean ICU stay was 7.6 ± 9.2 days. ICU mortality was 32%, in-hospital mortality was 39%, and 8% of patients died within the first 24 hours in the ICU.

2.2.2.1 Inferential analysis of the cohort

We conducted an inferential analysis of the cohort to identify key factors linked to ICU mortality, as detailed in Table 2.1. Age, chronic heart failure, chronic renal failure, chronic liver disease, active cancer, haematological malignancy, and immunosuppression were all associated with increased mortality. There was no statistical association between ICU mortality and year of admission, gender, hypertension, or chronic respiratory disease.

All severity parameters typically linked to ICU mortality—such as arterial hypotension, GCS score, lactate, PaO2/FiO2 ratio, use of renal replacement therapy, SOFA score, and IGS II—were highly significant predictors of mortality in our cohort ($p < 0.001$).

Regarding admission reasons, cardiogenic shock, cardiorespiratory arrest, and obstructive shock were linked to higher mortality in univariate analysis. Conversely, norepinephrine use for sedation-induced hypotension, toxicology admission, and hypovolaemia were significantly associated with survival. The use of vasopressor support for other indications was also associated with increased mortality.

2.2.2.2 Analysis of norepinephrine dosages

To estimate the prognostic impact of norepinephrine dosage, the population was divided into four subgroups based on the quartiles of maximum norepinephrine dosage in the first

24 hours. The first quartile included patients receiving less than 0.09 µg/kg/min, the second 0.09-0.17 µg/kg/min, the third 0.17-0.31 µg/kg/min, and the fourth more than 0.31 µg/kg/min.

We compared various variables across these quartiles, as shown in Table 2.2. The quartiles were similar in age, year of admission, and comorbidities, except for progressive cancers, which were more common in the fourth quartile. The highest dosage quartile had significantly lower systolic, diastolic, and mean blood pressure, lower daily diuresis, more acidic pH, higher lactate, and lower PaO2/FiO2 ratio (p<0.001). Both SOFA and IGS II scores were higher in this group, as indicated in Table 2.2.

Significant differences were found in management elements across quartiles. Patients in the highest dosage quartile had higher filling and total daily volumes, more frequent use of packed red blood cells and higher transfusion volumes, and more frequent use of colloids and hemisuccinate hydrocortisone (p<0.001). Additionally, continuous intravenous epinephrine, dopamine, and dobutamine use was significantly greater in the highest dosage quartiles.

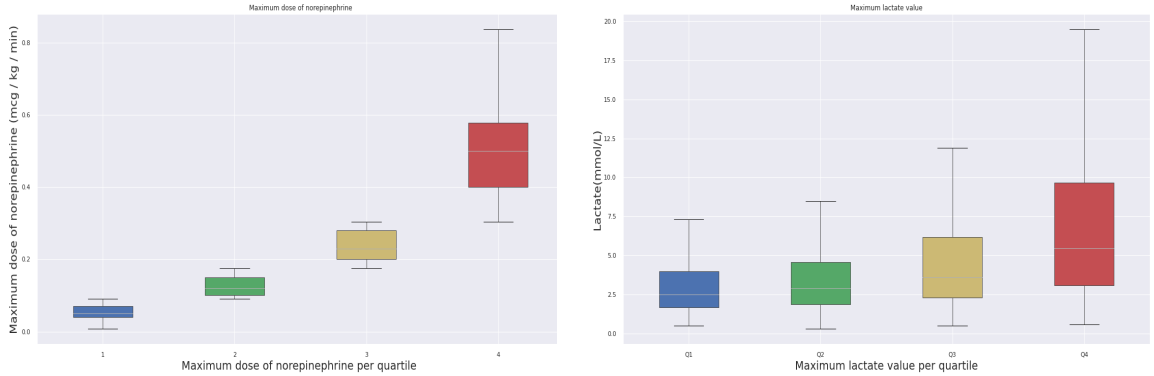
Table 2.2: Description of the different quartiles

		Missing	Overall	QUARTILE 1	QUARTILE 2	QUARTILE 3	QUARTILE 4	P-Value
Number of Patients			5735	1429	1430	1442	1434	
Survival ICU, mean (SD)		2010	248.1 (537.4)	370.5 (610.2)	300.6 (568.4)	229.3 (531.7)	133.3 (424.1)	<0.001
LOS, mean (SD)		0	7.6 (9.2)	6.6 (7.6)	7.6 (8.8)	8.8 (10.5)	7.5 (9.5)	<0.001
AGE, mean (SD)		0	67.3 (15.6)	67.8 (15.3)	68.1 (15.2)	67.2 (15.5)	66.2 (16.1)	0.006
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.2 [0.1,0.3]	0.1 [0.0,0.1]	0.1 [0.1,0.1]	0.2 [0.2,0.3]	0.5 [0.4,0.6]	<0.001
Epinephrine, mean (SD)		5156	0.2 (0.9)	0.0 (0.0)	0.1 (0.3)	0.1 (0.2)	0.3 (1.4)	0.012
Dopamine, mean (SD)		4700	16.0 (87.4)	9.6 (6.0)	11.4 (9.4)	13.1 (7.1)	22.5 (137.7)	0.268
Dobutamine, mean (SD)		5409	6.7 (5.0)	4.2 (2.2)	6.2 (4.9)	6.9 (4.3)	7.5 (5.7)	0.001
Lactate(mmol/L), mean (SD)		0	4.8 (4.0)	3.2 (2.4)	3.8 (3.0)	4.9 (3.9)	7.0 (5.1)	<0.001
Heart Rate, mean (SD)		1	118.4 (24.1)	111.4 (21.8)	115.7 (23.2)	121.0 (24.6)	125.7 (24.4)	<0.001
Systemic Blood Press(mmHg), mean (SD)		6	150.4 (25.9)	151.6 (24.3)	151.2 (25.4)	149.6 (25.4)	149.0 (28.1)	0.019
Mean Blood Press, mean (SD)		0	44.6 (14.2)	48.6 (12.7)	46.9 (12.8)	44.2 (13.4)	38.8 (15.8)	<0.001
Temperature(°C), mean (SD)		151	37.8 (1.1)	37.8 (0.9)	37.8 (0.9)	37.9 (1.1)	37.8 (1.3)	<0.001
Urine Output, mean (SD)		404	2981.7 (2979.7)	3324.1 (2144.1)	3165.8 (2216.2)	3010.5 (4167.5)	2410.4 (2829.8)	<0.001
Urea(mmol/l), mean (SD)		2	15.3 (10.0)	14.2 (9.6)	14.6 (9.4)	16.3 (10.6)	16.3 (10.2)	<0.001
Bilirubin(mg/dL), mean (SD)		857	2.5 (5.2)	1.9 (4.7)	1.9 (3.8)	2.7 (5.9)	3.3 (5.7)	<0.001
Bicarbonate(mEq/l), mean (SD)		16	18.3 (5.2)	20.3 (4.8)	19.3 (4.7)	17.8 (5.1)	15.7 (5.0)	<0.001
PaO2/FiO2, mean (SD)		990	173.0 (187.1)	192.2 (164.2)	173.8 (157.4)	175.1 (226.9)	155.1 (184.6)	<0.001
GCS Score, mean (SD)		243	7.3 (4.5)	8.4 (4.7)	7.9 (4.7)	7.0 (4.3)	6.0 (4.0)	<0.001
ICU Death, n (%)	0	0	3901 (68.0)	1210 (84.7)	1118 (78.2)	947 (65.7)	626 (43.7)	<0.001
	1		1834 (32.0)	219 (15.3)	312 (21.8)	495 (34.3)	808 (56.3)	
GENDER, n (%)	F	0	2466 (43.0)	613 (42.9)	590 (41.3)	656 (45.5)	607 (42.3)	0.128
	M		3269 (57.0)	816 (57.1)	840 (58.7)	786 (54.5)	827 (57.7)	
IGS II, mean (SD)		0	54.3 (16.7)	47.3 (15.7)	51.0 (16.2)	56.5 (15.7)	62.2 (15.3)	<0.001
SOFA, mean (SD)		0	10.7 (3.7)	8.7 (3.3)	10.2 (3.4)	11.3 (3.5)	12.5 (3.6)	<0.001

2.2.3 Prognostic impact of norepinephrine

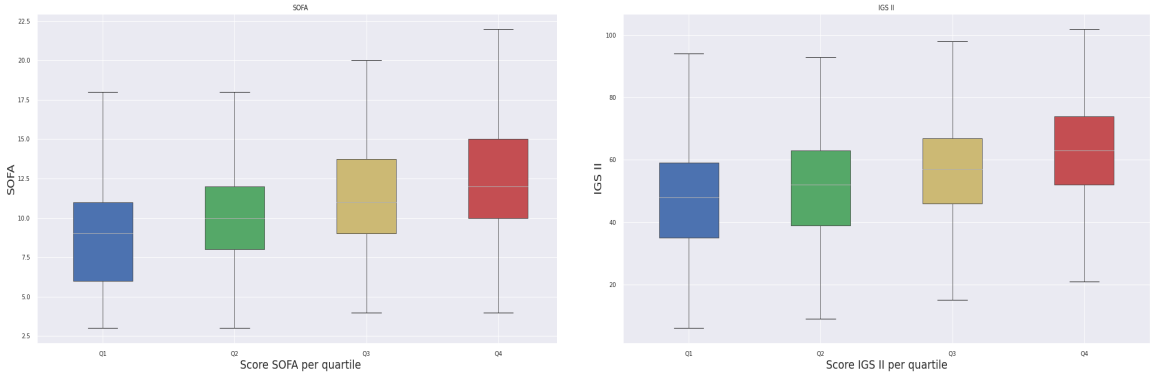
The 4^{th} quartile was very significantly (p<0.001) associated with excess mortality at D1 (16%), D7 (45%), ICU (62%) and in-hospital (69%), with a 6-fold increase in the risk of ICU death in the 4^{th} quartile.

The association between the quartile of patients receiving more than 0.31 µg/kg/min



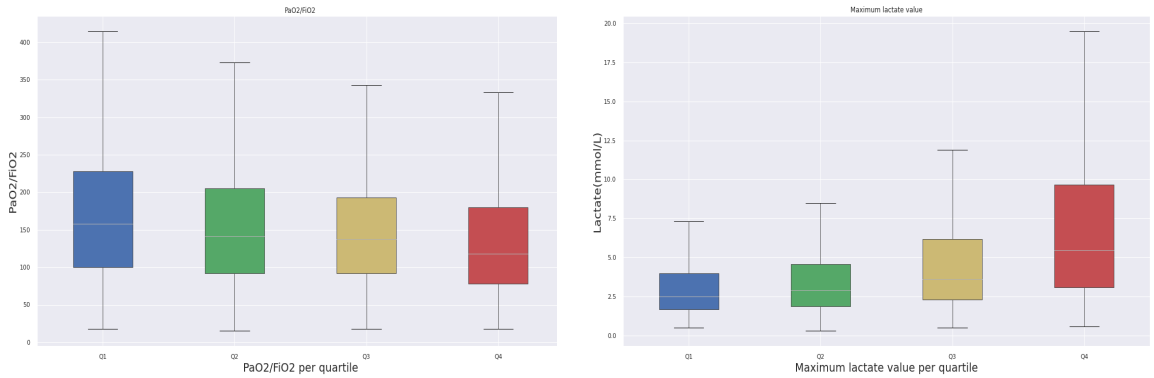
(a) Maximum of norepinephrine. (b) Maximum lactate value.

Figure 2.2: Critical clinical features assessment(Norepinephrine & Lactate) by quartile.



(a) score SOFA by quartile (b) score IGS by quartile

Figure 2.3: Critical clinical features assessment on severity scores parameters by quartile.



(a) Minimum PaO2/FiO2 ratio. (b) Maximum lactate value.

Figure 2.4: Assessment of oxygen therapy parameters by quartile.

and mortality was confirmed by multivariate analysis using a multiple logistic regression model, which took into account the main severity factors associated with ICU mortality and/or found in our study, namely the PaO2/FiO2 ratio, the use of lactate, SOFA score, the occurrence of cardiopulmonary arrest, and the use of cardio-respiratory arrest, and the use of adrenaline, with a p-value < 0.001. We've also notice that except the PaO2/FiO2 ratio, all others elements were independently associated with excess patient mortality.

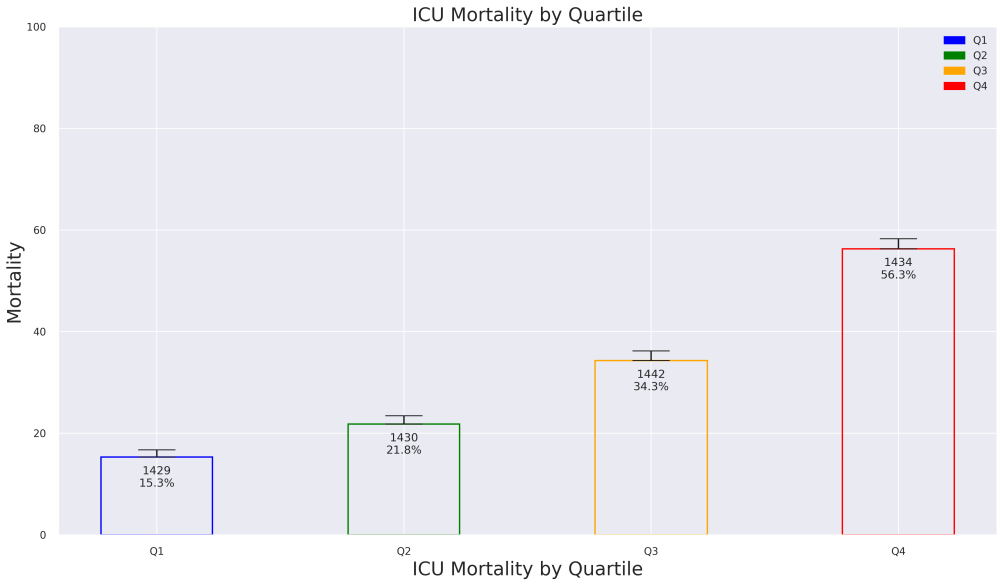


Figure 2.5: ICU Mortality for the different quartiles

2.2.3.1 High doses of norepinephrine and predictors of successful resuscitation

Given the significant and independent excess mortality in this group, Patients receiving more than 0.31 µg/kg/min of norepinephrine exhibited significant and independent excess mortality, classifying them as refractory shock states in our study, as mentioned earlier in Table 2.2. This group, demographically and comorbidly similar to the general population, had a mean maximum norepinephrine dose of 0.5 µg/kg/min. Norepinephrine support began on average three hours after admission and lasted for 19 ± 5 of the first 24 hours. One-third of these patients received additional catecholamines, with 57% receiving a mix of epinephrine, norepinephrine, dopamine, and dobutamine within the first 24 hours.

These patients had numerous negative prognostic indicators, including a lowest mean arterial pressure of 38.8 mmHg, a mean maximum lactate level of 7.0 ± 5.2 mmol/L, and a lowest PaO2/FiO2 ratio averaging 154.9 ± 185.0 mmHg. Additionally, 70% required extra-renal purification, with mean IGS II and SOFA scores at 24 hours of 62.2 ± 15.4 and 12.4 ± 3.6, respectively.

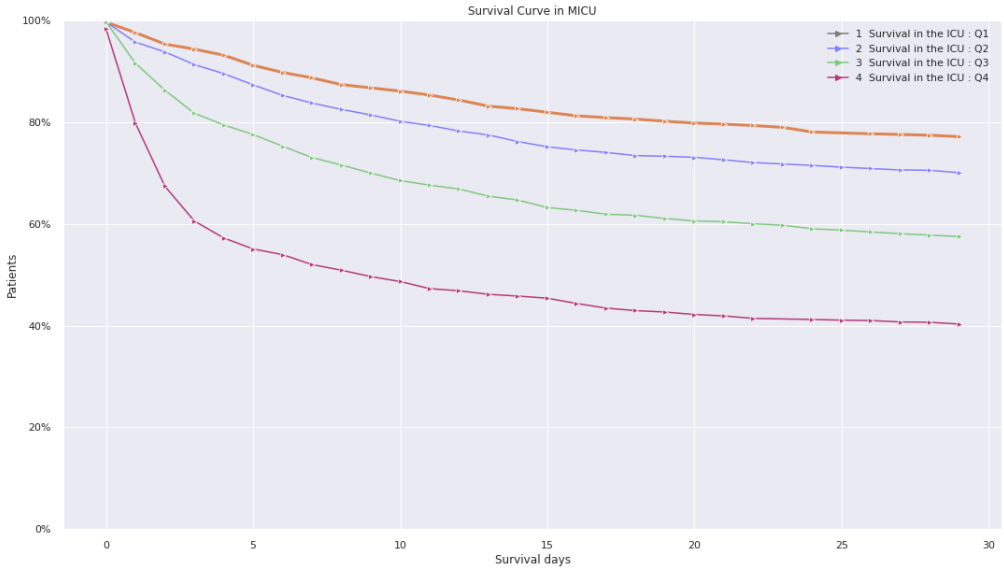


Figure 2.6: Survival in the ICU for the different quartiles

2.2.3.2 Factors for treatment success and futility

Table 2.3: Description of patients in the 4^{th} quartile

		Missing	Overall	Non-Survivors	Survivors	P-Value
Number of Patients			1434	808	626	
Survival ICU, mean (SD)		331	133.3 (424.1)	4.8 (7.4)	485.3 (710.2)	<0.001
AGE, mean (SD)		0	66.2 (16.1)	67.6 (16.1)	64.3 (16.0)	<0.001
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.5]	<0.001
Epinephrine, mean (SD)		1242	0.3 (1.4)	0.3 (0.6)	0.4 (2.2)	0.769
Dopamine, mean (SD)		1020	22.5 (137.7)	26.7 (165.5)	13.1 (9.1)	0.167
Dobutamine, mean (SD)		1291	7.5 (5.7)	8.0 (6.1)	6.5 (4.7)	0.123
Lactate(mmol/L), mean (SD)		0	7.0 (5.1)	8.8 (5.6)	4.7 (3.3)	<0.001
Heart Rate, mean (SD)		1	125.7 (24.4)	128.2 (25.1)	122.4 (23.1)	<0.001
Systemic Blood Press(mmHg), mean (SD)		4	149.0 (28.1)	146.1 (30.4)	152.8 (24.5)	<0.001
Mean Blood Press(mmHg), mean (SD)		0	38.8 (15.8)	34.5 (15.6)	44.3 (14.3)	<0.001
Temperature(°C), mean (SD)		39	37.8 (1.3)	37.5 (1.5)	38.1 (0.9)	<0.001
Urine output, mean (SD)		133	2410.4 (2829.8)	1474.3 (2214.8)	3467.6 (3068.2)	<0.001
Urea(mmol/l), mean (SD)		2	16.3 (10.2)	16.9 (10.0)	15.4 (10.4)	0.006
Bilirubin(mg/dL),, mean (SD)		168	3.3 (5.7)	4.2 (6.9)	2.1 (3.4)	<0.001
PaO2/FiO2 ratio (mmHg), mean (SD)		109	155.1 (184.6)	142.6 (170.5)	172.1 (201.3)	0.005
GCS Score, mean (SD)		94	6.0 (4.0)	5.6 (3.8)	6.6 (4.2)	<0.001
ICU Death, n (%)	0	0	626 (43.7)		626 (100.0)	<0.001
	1		808 (56.3)	808 (100.0)		
GENDER, n (%)	F	0	607 (42.3)	336 (41.6)	271 (43.3)	0.552
	M		827 (57.7)	472 (58.4)	355 (56.7)	
IGS II, mean (SD)		0	62.2 (15.3)	66.8 (14.7)	56.2 (13.9)	<0.001
SOFA, mean (SD)		0	12.5 (3.6)	13.3 (3.5)	11.3 (3.3)	<0.001

We aimed to identify factors linked to excess mortality within this specific group. For the broader population, univariate analysis showed a significant association between maximum norepinephrine dose and mortality, as detailed in Table 2.3. Lower blood pressure, pH,

and PaO2/FiO2 ratio, along with higher lactate, SOFA score, and IGS II score, were strongly associated with mortality (p<0.001). The use of extra-renal purification and occurrence of cardiorespiratory arrest also indicated poor prognosis. Among comorbidities, only chronic heart failure and chronic renal failure were linked to excess mortality.

In addition to the maximum norepinephrine dose and its administration duration, we evaluated the prognostic impact of norepinephrine dosage kinetics at 24 hours. Among the 1434 patients receiving more than 0.5 µg/kg/min, 896 (62%) had a reduced dose at 24 hours, while 108 (7%) had an increasing dose at 24 hours or at death. A decrease in norepinephrine dosage at 24 hours was strongly associated with patient survival, with surviving patients showing a 46.0% decrease compared to 85.2% in non-survivors (p<0.001).

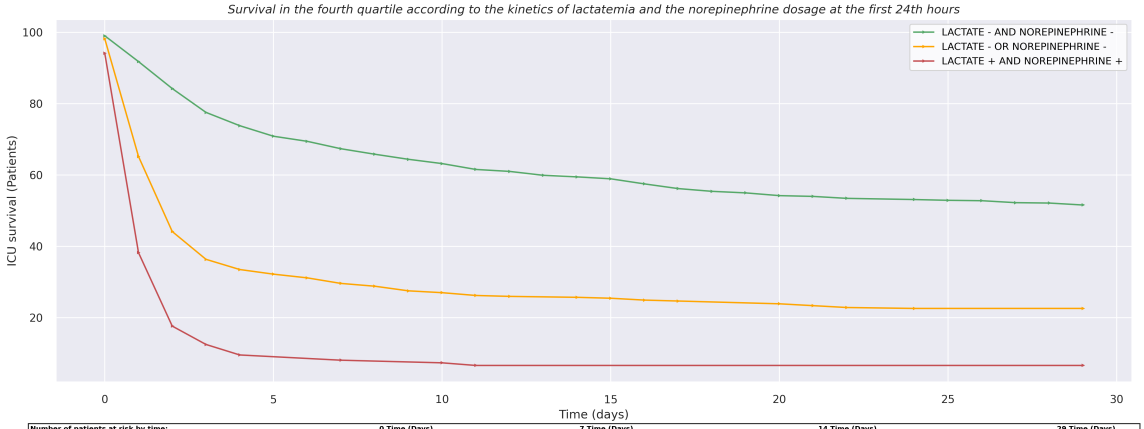
We also found a significant association (p<0.001) between decreasing norepinephrine and lactate levels and survival. Based on these findings, we constructed survival curves for fourth-quartile patients, categorized into four groups: decreased lactate and norepinephrine at 24 hours, decreased lactate and increased norepinephrine, increased lactate and decreased norepinephrine, and increased lactate and norepinephrine. Survival analysis revealed that patients with decreased lactate and increased norepinephrine, and those with increased lactate and decreased norepinephrine had similar prognoses, leading to their combination into one group: patients with decreased lactate or norepinephrine. Mortality rates at D1, D7 in ICU, and in-hospital differed significantly between the three groups. The group of 896 patients with increased lactate and norepinephrine at 24 hours had exceptionally high mortality, ranging from 50% at 24 hours to 92% in-hospital.

Table 2.4: Description of the different kinetics

		Missing	Overall	NOP+ AND LACTATE+	NOP- AND LACTATE-	NOP- OR LACTATE-	P-Value
Number of Patients			1434	136	913	385	
Survival ICU, mean (SD)		331	133.3 (424.1)	25.7 (199.8)	181.2 (480.9)	83.9 (354.4)	<0.001
AGE, mean (SD)		0	66.2 (16.1)	68.4 (16.1)	65.9 (16.3)	65.9 (15.6)	0.251
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.001
Epinephrine, mean (SD)		1242	0.3 (1.4)	0.2 (0.5)	0.3 (1.8)	0.4 (0.7)	0.940
Dopamine, mean (SD)		1020	22.5 (137.7)	18.6 (7.7)	13.6 (8.0)	38.0 (238.9)	0.264
Dobutamine, mean (SD)		1291	7.5 (5.7)	10.5 (7.1)	7.0 (5.4)	7.1 (5.5)	0.048
Lactate(mmol/L), mean (SD)		0	7.0 (5.1)	9.7 (6.0)	6.2 (4.4)	7.9 (6.0)	<0.001
Systemic Blood Press(mmHg), mean (SD)		4	149.0 (28.1)	141.9 (29.8)	151.9 (26.8)	144.8 (29.6)	<0.001
Mean Blood Press(mmHg), mean (SD)		0	38.8 (15.8)	32.9 (14.5)	41.2 (15.2)	35.1 (16.5)	<0.001
Temperature(°C), mean (SD)		39	37.8 (1.3)	37.3 (1.8)	38.0 (1.1)	37.5 (1.4)	<0.001
Urea(mmol/l), mean (SD)		2	16.3 (10.2)	16.4 (10.5)	16.3 (10.4)	16.1 (9.5)	0.942
Bilirubin(mg/dL)„ mean (SD)		168	3.3 (5.7)	2.9 (5.1)	3.1 (5.3)	3.8 (6.9)	0.118
PaO2/FiO2 ratio (mmHg), mean (SD)		109	155.1 (184.6)	154.6 (261.8)	153.2 (145.6)	159.7 (231.8)	0.857
GCS Score, mean (SD)		94	6.0 (4.0)	6.4 (4.5)	5.8 (3.8)	6.4 (4.3)	0.050
ICU Death, n (%)	0	0	626 (43.7)	9 (6.6)	519 (56.8)	98 (25.5)	<0.001
	1		808 (56.3)	127 (93.4)	394 (43.2)	287 (74.5)	
GENDER, n (%)	F	0	607 (42.3)	60 (44.1)	390 (42.7)	157 (40.8)	0.736
	M		827 (57.7)	76 (55.9)	523 (57.3)	228 (59.2)	
IGS II, mean (SD)		0	62.2 (15.3)	63.2 (16.6)	61.9 (15.1)	62.6 (15.4)	0.564
SOFA, mean (SD)		0	12.5 (3.6)	12.4 (3.4)	12.5 (3.5)	12.5 (3.8)	0.956

2.2.4 Conclusion

Following its established safety and demonstrated superiority in tolerability over dopamine and adrenaline, norepinephrine has emerged as a foremost recommendation by both French and international academic societies for managing various hemodynamic failures. Its dose-dependent vasoconstrictor effect indicates a correlation between increasing dosage require-



Number of patients at risk by time:		0 Time (Days)	7 Time (Days)	14 Time (Days)	29 Time (Days)
LACTATE - AND NOREPINEPHRINE -	(913 PATIENTS) :	904	615	543	471
LACTATE - OR NOREPINEPHRINE -	(385 PATIENTS) :	378	114	99	87
LACTATE + AND NOREPINEPHRINE +	(136 PATIENTS) :	128	11	9	9

Fourth Quartile		Overall	NOREPINEPHRINE + AND LACTATE +	NOREPINEPHRINE - AND LACTATE -	NOREPINEPHRINE - OR LACTATE -	P-Value
(n/N)		1434	136	913	385	
NAD max (mcg/kg/min), median [Q1,Q3]		0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	<0.001
AGE, mean (SD)		66.2 (16.1)	68.4 (16.1)	65.9 (16.3)	65.9 (15.6)	0.251
LACTATE(mmol/L), mean (SD)		7.0 (5.1)	9.7 (6.0)	6.2 (4.4)	7.9 (6.0)	<0.001
GENDER, n (%)	F	607 (42.3)	60 (44.1)	390 (42.7)	157 (40.8)	0.736
	M	827 (57.7)	76 (55.9)	523 (57.3)	228 (59.2)	
ICU_DEATH, n (%)	0	626 (43.7)	9 (6.6)	519 (56.8)	98 (25.5)	<0.001
	1	808 (56.3)	127 (93.4)	394 (43.2)	287 (74.5)	

Figure 2.7: ICU Mortality for different kinetics of norepinephrine and lactate.

ments, the exacerbation of circulatory failure, and heightened patient mortality rates. Notably, small-scale retrospective studies in septic shock have highlighted an alarming rise in mortality rates seven days post-administration for patients who were administered doses of norepinephrine exceeding 0.6 µg/kg/min within the initial 24 hours. Identifying the threshold dose beyond which norepinephrine administration may become counterproductive, or even detrimental, remains an unresolved challenge within the medical community.

Our investigation, derived from a retrospective analysis of patients administered norepinephrine within the first 24 hours of ICU admission, our study sought to assess the prognostic implications of the maximum dose administered during this period and to identify predictive factors influencing treatment outcomes, thereby corroborating the findings of Hugerot and al [127].

Conclusively, our findings suggest a discernible link between the maximal dose of norepinephrine administered within the initial 24 hours and an increased mortality rate beyond a threshold of 0.4 µg/kg/min. Particularly, patients receiving doses higher than 0.85 µg/kg/min displayed a notable surge in mortality, prompting a reassessment of treatment efficacy under such conditions. This observation is particularly crucial when the norepinephrine dose kinetics and lactate levels at the 24-hour mark are unfavorable, or when concurrent factors associated with higher mortality risk, like the use of adrenaline, need for extrarenal clearance, or cardiac arrest events, are observed.

Part II

State of the art

Chapter 3

Background & Related work

Begin at the beginning, the King said gravely, "and go on till you come to the end: then stop."

Lewis Carroll - Alice in Wonderland

In this section, we first define the multivariate event time series, their representation of time irregularity in highly variable and episodic medical time series. After that, we review existing approaches relevant to multivariate time series modeling. Finally, we review existing deep-learning based methods used for modeling sequential data, exposes the differences between these methods and gives an overview of their application in medical domain. We also describe the published works that have considered the time-irregularity when applying

these methods to highly variable time-series.

3.1 Irregularly-sampled EHR time series data

Irregularly-sampled time series are characterized by non-uniform time intervals between successive measurements. Such time series naturally occur in many domains. In biological studies, for example, observational time series data from free-living animals inevitably lack data points due to movement of the subjects, weak transmitter reception, or poor weather and lighting conditions that hinder observations. Even laboratory studies often suffer from instrumental drop-outs or interference caused by electrical noise [136].

The irregular and non-uniform recording of a patient's health status in electronic health records is a characteristic feature of routine care. This variability arises from a multitude of factors, each playing a distinct role in shaping the data's distribution over time. One key determinant is the nature of the physiological signal under observation. Different signals may necessitate distinct monitoring intervals. Additionally, the method used for measurements plays a crucial role. For example, continuous monitoring devices versus intermittent manual measurements can lead to significant differences in data distribution.

Moreover, the involvement of various healthcare providers further contributes to this irregularity. Different caregivers may have diverse recording practices, influenced by their professional training, expertise, and specific care protocols or thresholds for triggering measurements. Additionally, the severity of a patient's condition is a significant factor. In critical situations, healthcare providers tend to monitor vital signs more frequently and intensively to ensure timely intervention and appropriate care.

An illustrative example can be found in the monitoring of heart rate. In critical scenarios, such as during an acute medical event, the heart rate through an ECG is likely to be recorded at much shorter intervals compared to routine check-ups. This heightened frequency of measurement is essential for promptly identifying any alarming changes in the patient's condition and taking immediate action.

The diversity in data recording patterns highlights the intricacy of dealing with electronic health records. It underscores the necessity of robust analytical techniques capable of accommodating such irregularities in order to distill meaningful insights and facilitate well-informed clinical decision-making, as demonstrated in the study [109] which provides valuable insights into the difficulties and significance of comprehending and proficiently leveraging such data within healthcare contexts.

Consequently, modeling this form of data presents considerable challenges. This stems from the fact that data instances aren't naturally confined to a fixed-dimensional feature space due to the irregularity of sampling. Additionally, the limited number of observations recorded for each event type results in a highly sparse dataset. It's worth noting that this pattern of sporadic and uneven data sampling isn't exclusive to healthcare. It's prevalent in other fields like climate science, ecology and astronomy [139].

As evidence, after conducting hourly data sampling for a set of 103 variables used in our study, we discovered that the absence of encompassing both laboratory and vital signs records exceeded 93% for all patients. This phenomenon is explicable by the fact that each patient might undergo only a few medical tests based on their specific requirements and

these tests are usually scheduled sporadically over a considerable duration , as demonstrated in the following Fig. 3.1 and Fig. 3.2.

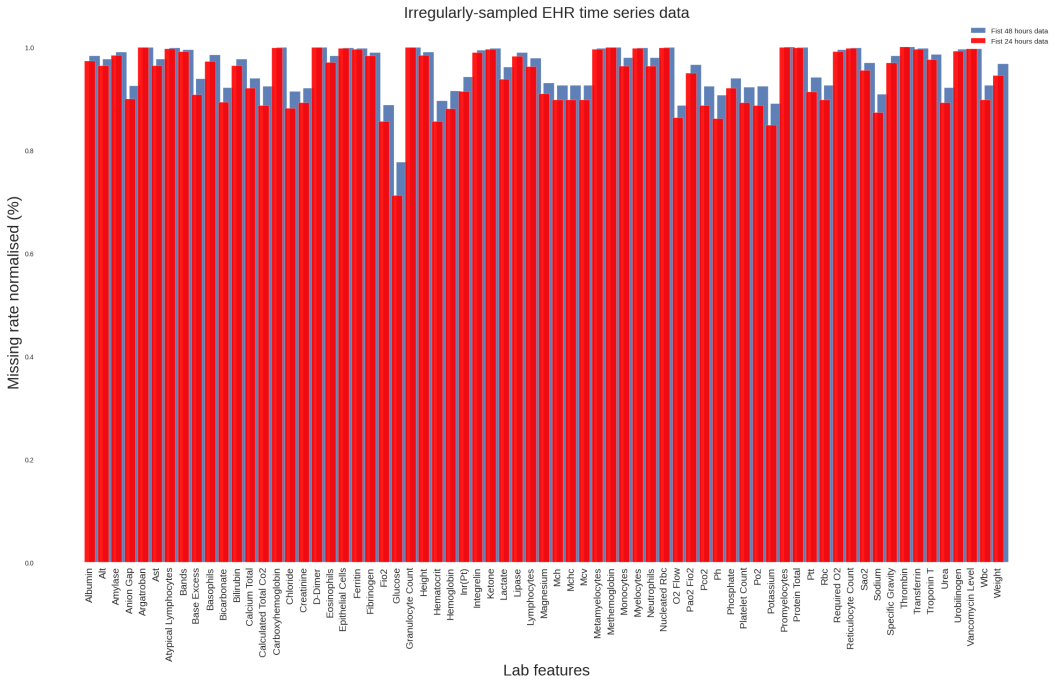


Figure 3.1: Normalised Missing ratio (%) for lab features within the first 24 and 48 hours of data collection (MIMIC III).

In our research, we concentrate on modeling a collection of irregularly-sampled time series observed over a shared time interval by proposing novel approaches capable of handling the aforementioned challenges related to EHR data analysis.

3.1.1 Notations

We note $n \in N$ the set of ICU stays considered in a medical study. To form our cohort of subjects, we represent each sample as

$$\boldsymbol{X} = \left\{ (\boldsymbol{X^{(n)}}, \boldsymbol{X^{(n)}_{last}}, \boldsymbol{F^{(n)}_{x}}, \boldsymbol{\Delta^{(n)}_{t}}, \boldsymbol{S^{(n)}}) \right\}_{n=1}^{N}, \tag{3.1}$$

where each sample is associated with a unique index n.

For each stay within our cohort of N stays, we examine a multivariate time series that incorporates D physiological variables, such as laboratory results and vital signs, observed over a duration of time T (window size):

$$\boldsymbol{X^{(n)}} = \{x_1, x_2, ..., x_T\} \in \mathbb{R}^{T \times D} \tag{3.2}$$

where each x_t^d represents the observation of the d^{th} variable (e.g. diastolic blood pressure) measured at timestamp $t \in \{1, 2, ..., T\}$ for stay n-th.

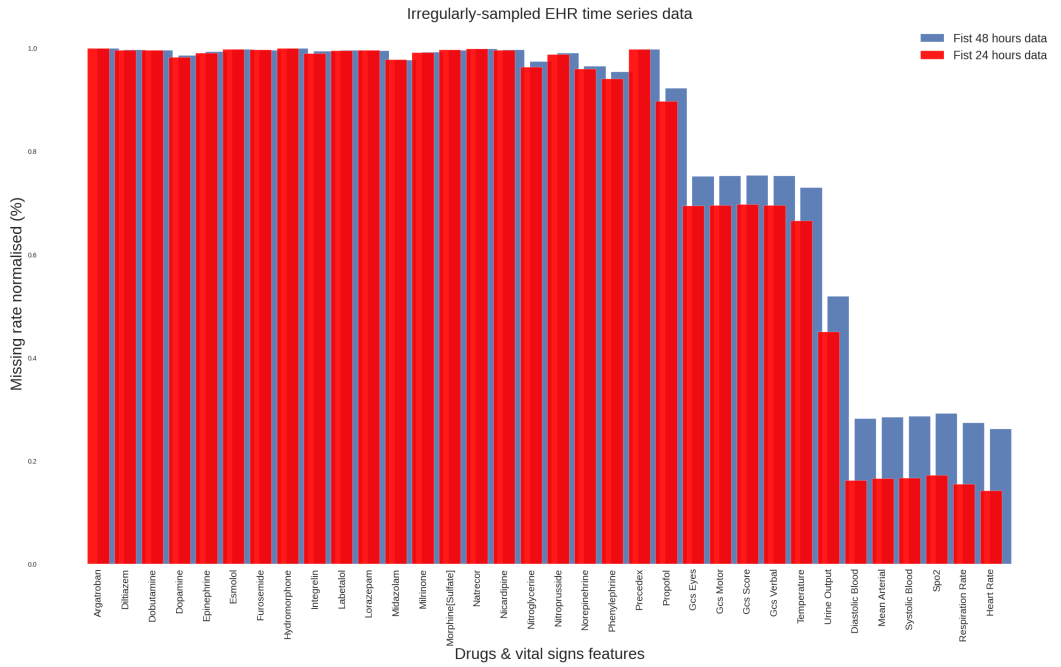


Figure 3.2: Normalised Missing ratio (%) for vital signs and drug features within the first 24 and 48 hours of data collection (MIMIC III).

To address missing values in $\boldsymbol{X}^{(n)}$, we introduce a masking vector in order to calculate $\boldsymbol{\Delta}_t^{d(n)}$, the elapsed times between data collections with M = $\{m_1, m_2, ..., m_T\} \in \mathbb{R}^{T\times D}$ of the same size as $\boldsymbol{X}^{(n)}$. This vector indicates which variables are observed or missing at each time step. We initialize M as follows:

$$\boldsymbol{m}_t^{d(n)} = \begin{cases} 1, & \text{if } x_t^d \text{ is observed in the input data} \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

We calculate the frequency measurements of $\boldsymbol{f}_{\boldsymbol{x}}^{(n)}$ based on the number of times n_x, x has been measured over a given windows size T.

$$\boldsymbol{f}_{\boldsymbol{x}}^{(n)} = n_{x^d}/T \tag{3.4}$$

In $\boldsymbol{X}_{\boldsymbol{last}}^{(n)}$, we store the last observation x_{last}^d of each physiological variable x^d within the past T time steps.

A set of static features, e.g. demographic features, denoted S is also extracted for each stay

$$\boldsymbol{S}_i^{(n)} = \left[s_{i,1}, s_{i,2}, \ldots, s_{i,n}\right] \in \mathbb{R}^S, \quad i = 1, 2, \ldots, N \tag{3.5}$$

To incorporate temporal information(intervals between data collections), we introduce a time interval variable $\Delta_t^d \in \mathbb{R}^D$ for each variable x_t^d. We assume that the initial observation

is made at timestamp $\Delta_1 = 0$. The definition of the time interval is expressed as follows:

$$\boldsymbol{\Delta_t^{d(n)}} = \begin{cases} 0, & t = 1 \\ \Delta_{t-1}^d + 1, & t > 1, m_{t-1}^d = 0 \\ 0, & t > 1, m_{t-1}^d = 1 \end{cases} \tag{3.6}$$

This formulation can be interpreted as follows:

If $t = 1$, then Δ_t^d is set to zero, representing the variable's value at the first time point, serving as the reference or baseline value.

If $t > 1$ and $m_{t-1}^d = 0$, then Δ_t^d is incremented by one unit compared to Δ_{t-1}^d. This implies a linear trend in the variable over time, with each time point one unit larger than the previous time point, as the subject has not experienced an event (indicated by $m_{t-1}^d = 0$) at time $t - 1$.

If $t > 1$ and $m_{t-1}^d = 1$, then Δ_t^d is reset to zero. This signifies a "restart" of the variable after an event has occurred (indicated by $m_{t-1}^d = 1$) at time $t - 1$.

- **Interventions:**

In this setting, we also considered that the elapsed time vector $\boldsymbol{\Delta_t^{(n)}}$ can be viewed as an intervention since recording a patient's clinical feature measurements over time is crucial for evaluating his/her health status. The frequency of these measurements is directly correlated to the patient's health condition, with shorter intervals between measurements potentially indicating a rapidly deteriorating condition, while longer intervals may suggest a stable state. Overall, monitoring changes in these measurements over time can enable necessary adjustments to the treatment plan, ensuring effective management of the patient's condition.

3.2 Multivariate event time series

We define multivariate event time series by a time-stamped sequence of events U = $\{u_j\}_j$, where each event $u_j = [e_j, t_j]$ is represented by a pair of an event type e_j and its time t_j. We assume there are |E| different event types defining the multivariate event time series. A univariate event time series would be defined by a single event type |E|=1.

The event time series with continuous time stamps can be directly modeled using point processes [65, 84]. Examples of such processes are a Poisson process [74] or a Hawkes process [135]. These models have been applied to various event sequence problems including clinical event prediction [98, 113]. However, these models are hard to optimize directly and the existing works only explore time series with a relatively small number of events. Because of these limitations, the event time series are often converted to discrete-time series with a non-overlapping segmentation window and then generate binary vector $y_i \in \{0, 1\}^{|E|}$ that represents all event occurrences during the timings of the window, as shown in Fig. 3.3 where the original event time series are segmented using a window spanning some fixed

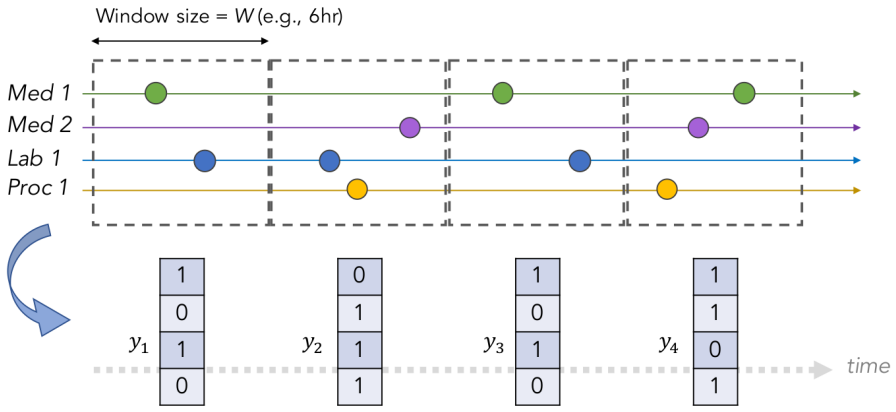


Figure 3.3: Overview of multivariate event time series processing. As seen in the upper part of the figure, the original EHR-based time series data consists of event occurrences on continuous time.

period of time, and events within the window are considered to co-occur in the discretized time.

3.3 Segmentation of event time series

We define the discrete-time event time series as follows:

- Discrete-time event time series Y = $\{y_i\}_i$ consist of a sequence of states y_i where $y_i \in \{0,1\}^{|E|}$ is a binary vector that represents occurrences of events of different types at a discrete time step i, and |E| denotes the total number of event types.
- Discrete-time event time series are generated from time stamped multivariate event time series U through segmentation of event occurrences with a time window W as described in Fig. 3.3.

In the following sections, we assume we have data that consists of N discrete-time event time series: D = $\{X_1, ..., X_N\}$. Next, we briefly review existing modeling approaches for discrete-time event time series.

3.4 Markov models

Markov models form a foundation of discrete-time series models. Given their simplicity and tractability, the majority of the event time series models are special cases of Markov models [106, 111]. Markov models represent an observed sequence of a random process over time as a sequence of states. The state is a categorical variable at a specific (discrete) time step. The Markov property assumes that the current state captures all necessary information

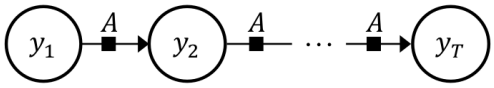


Figure 3.4: An illustration of a Markov model. The transition between observations y_i are defined by the transition matrix A in Eq. (3.9).

relating to the future and past. In other words, the next state depends only on the most recent state, and is independent of past states:

$$P(y_T|y_{T-1}, y_{T-2}, \ldots, y_1) = P(y_T|y_{T-1}) \tag{3.7}$$

In this case, the joint distribution of an observed sequence is modeled as a chain of the conditional probabilities:

$$P(y_1, y_1, \ldots, y_T) = p(y_1) \prod_{i=2}^{T} P(y_i|y_{i-1}) \tag{3.8}$$

The conditional probability defining a transition is parameterized by a transition matrix $A \in \mathbb{R}^{|E|\times|E|}$:

$$A_{m,n} = P(y_i = n|y_{i-1} = m) \tag{3.9}$$

where $\sum_{n=1}^{|E|} A_{m,n} = 1$ for all m . The transition matrix A can be learned by the maximum likelihood estimation [134]. The standard Markov models assume all states of the time series are directly observed. However, the states of many real-world processes are not directly observable. One way to resolve the problem is to define the state in terms of a limited number of past observations or features defined on past observations [54] and another one is to use the Markov models with hidden states [131, 150].

3.5 Attention mechanism

The attention mechanism, inspired by human cognitive processes [3, 115, 164] which selectively focus on crucial information, plays a pivotal role in deep neural networks. In neuroscience, attention systems are well-studied [126], highlighting their function in prioritizing relevant inputs for optimal output computation. For instance, in visual perception, attention directs focus towards objects rather than background elements, and in text processing, it identifies significant words while disregarding others [3].

The patients' health trajectories often span over several years, and only using the last hidden state to represent these long sequences leads to information loss and a weak patient representation. Consequently, aggregating the hidden states produced by the recurrent neural network from all-time steps to represent the patient's history offered a promising solution to capture all relevant medical states

$$\boldsymbol{R}_p = Agg(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N) \tag{3.10}$$

where a straightforward aggregation method like average pooling is commonly used. To enhance representation accuracy, assigning importance weights $\boldsymbol{\alpha}_i$ to each time observation $\boldsymbol{x}_t$ represented by $\boldsymbol{h}_i$ becomes crucial:

$$\boldsymbol{R}_p = \sum_{j=1}^{\boldsymbol{N}} \boldsymbol{\alpha}_i \boldsymbol{h}_i \tag{3.11}$$

The patient care pathway could be organized in a sequence of visits, in a time-window segment grouping a set of visits together or by episodes of care defined using medical expert rules. Therefore, each subsequence contains a variable amount of clinical information. Using a global attention weight to determine the importance of the overall sequence could over or underestimate specific medical events occurring within a particular segment. The hierarchical attention mechanism computes different levels of importance to address this limitation: inter-subsequence and intra-subsequence. The inter-subsequence models the global importance of a set of medical events occurring in the same period on the global patient health state. In comparison, intra-subsequence attention measures the local importance of each medical event on the near-term evolution of patient health. We note K_j the number of events recorded at a given timestamp j and $\boldsymbol{\gamma}_i^j$ the attention weight of a medical event x_{ij} (such as a numerical lab measurement or diagnosis code), the patient representation is then expressed as:

$$\boldsymbol{R}_p = \sum_{j=1}^{\boldsymbol{N}} \sum_{i=1}^{\boldsymbol{K_j}} \boldsymbol{\alpha}_i \boldsymbol{\gamma}_i^j \boldsymbol{h}_i \tag{3.12}$$

The attention mechanism empowers models to selectively focus on specific image regions or relevant input time steps, enhancing performance across various tasks. This section explores attention models, their variants, and their integration into RNNs for diverse learning objectives.

3.5.1 Attention model

The attention model depicted in Fig. 3.5 illustrates how to aggregate a summation vector $\boldsymbol{z}$ from a variable-size set $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$ given context $\boldsymbol{c}$. These vectors could represent RNN-generated hidden states, parts of an image, or object feature vectors. Traditional methods like sum or mean pooling treat all vectors equally. In contrast, the attention mechanism computes $\boldsymbol{z}$ as a weighted sum, with weights p_i for each $\boldsymbol{h}_i$, ensuring $\Sigma_{i=1}^{n} p_i = 1$:

$$\boldsymbol{a}_i = tanh(W_a \boldsymbol{h}_i + U_a \boldsymbol{c} + \boldsymbol{b}_a) \tag{3.13}$$

$$p_i = softmax(\boldsymbol{u}^\top \boldsymbol{a}_i) \tag{3.14}$$

$$\boldsymbol{z} = \sum_{i=1}^{\boldsymbol{n}} p_i \boldsymbol{h}_i \tag{3.15}$$

Here, $\boldsymbol{a}_i$ integrates $\boldsymbol{h}_i$ and $\boldsymbol{c}$, $\boldsymbol{u}$ measures contributions, and $\{\boldsymbol{W}_a, \boldsymbol{U}_a, \boldsymbol{b}_a\}$ and $\boldsymbol{u}$ are learned weights. Soft attention allows all vectors to contribute to $\boldsymbol{z}$, with complexity $O(n)$. Multiple attention mechanisms can enrich input information [103, 158].

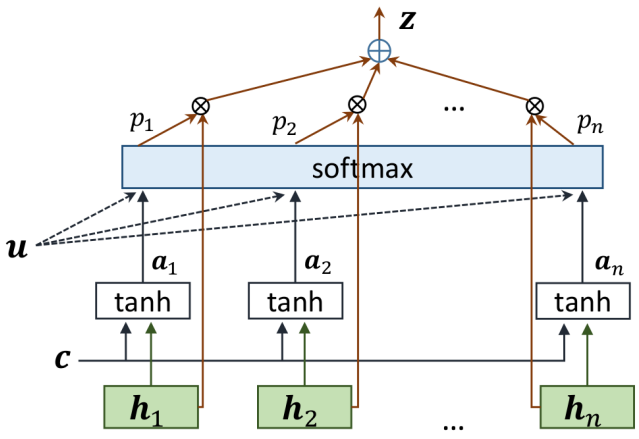


Figure 3.5: Attention mechanism on $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$[124] given a context $\boldsymbol{c}$.

Hard attention, an alternative, selects $\boldsymbol{z}$ directly from $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_n\}$ based on highest or random probabilities p_i, suitable for tasks like searching or sorting.

3.5.2 Transformer: self-attention mechanism

Instead of computing absolute importance $\boldsymbol{\alpha}_i$ of the position i related to the entire input sequence, the self-attention mechanism [157], also known as intra-attention, defines a local score. This score measures the relative importance of the position i taking into account the surrounding context positions $k < i$ and/or $k > i$ of a single input sequence. Their proposed score function computes how each position i (value vector $v_i \in \boldsymbol{R}^d$) is strongly correlated with the given output at i (query vector $q_i \in \boldsymbol{R}^d$) taking into account all surrounding positions (key matrix K $\in \boldsymbol{R}^{N \times d}$). In sequence modeling, both the keys and values are the neural network's hidden states. The score function of timestamp i returns relative weights overall positions and is formulated as:

$$\boldsymbol{\alpha}_i = \frac{1}{\sqrt{d}} q_i^T K \in \boldsymbol{R}^N \tag{3.16}$$

The Transformer, an encoder-decoder architecture, is entirely built upon self-attention mechanisms and was proposed by [157] as a replacement for recurrent network units. They showed that the multi-head self-attention mechanism coupled with a residual connection to the input word embedding representation is sufficient to capture the long-range dependencies between the given word and the other contextual positions.

3.6 Modeling temporal mechanisms of irregular medical time series

Modeling patient disease progression using EHRs data is critical to assist clinical decision-making. However, measurements in EHRs are commonly acquired with irregular intervals.

For example, when a patient is in a severe condition, events are likely to be recorded more frequently than when a patient is in a relatively "healthier" condition. Hence such varying time intervals can reveal patient's health status on certain impending conditions, and it is important to consider the time intervals between temporal events to capture latent progressive patterns of a disease. Therefore, when modeling a patient's longitudinal medical data, the model should account for the irregular periods between visits or clinical features as additional information to the event itself [151]. We categorize the work that considered learning to represent time into two main categories: (1) Considering time gaps as an additional feature to feed to the neural method and learn their embedding vector [24, 93]. (2) Extending the original neural architecture by injecting the time gaps as an additional parameter of the model [5, 93, 124, 176].

3.6.1 Time as an additional input variable

The initial multivariate event time series based on EHRs consist of continuously logged events. To efficiently process these clinical event time series, they are transformed into a discrete-time representation using window-based segmentation [90, 133] , mapping multiple events within a given time window into a fixed-sized vector.

Luo et al [102] designed HiTANet, a Hierarchical Time-aware Attention Network based on the Transformer model [157]. HiTANet uses as input a concatenation of visit sequence vectors and time vectors ($\boldsymbol{\Delta}_t$) between visits. It operates in two stages: local evaluation and global synthesis.

In the local evaluation stage, HiTANet addresses the challenge of incorporating historical patient information to assess current health risks. It models disease progression with a time-aware Transformer that learns time-sensitive attention weights for individual visits, combining both time and visit information to generate local attention weights and an overall patient representation. By embedding time information into vectors, it avoids the pitfalls of a monotonic time decay function.

To enhance interpretability of timesteps in disease progression, the global synthesis stage introduces a time-aware key-query attention mechanism. This mechanism uses the overall representation from the local evaluation stage as a query vector to generate global attention weights for each visit, utilizing temporal information once more. Finally, the two attention mechanisms are fused to create patient representations for disease risk prediction.

Shukla et al [144] have proposed mTAND, a transformer-like attention mechanism to re-represent an irregularly sampled time series at a fixed set of reference points by using time points as queries and the observed time points as keys. The encoder takes the irregularly sampled time series as input and produces a fixed-length latent representation over a set of reference points, while the decoder uses the latent representations to produce reconstructions conditioned on the set of observed time points. They evaluate on interpolation and classification tasks.

Ma et al [104] proposed ConCare, a transformer based model to handle the irregular EHR data and extract feature inter-relationship to perform individualized healthcare pre-

diction. They improved the multi-head self-attention via the cross-head decorrelation, so that the inter-dependencies among dynamic features and static baseline information can be effectively captured to form the personal health context.

Choi et al [24] proposed RETAIN, based on GRU [27], RETAIN employs a two-level attention mechanism to identify meaningful visits and specific features that contribute to the prediction. To provide precise interpretability, they use a reverse time attention mechanism in their RETAIN model. Moreover, the RETAIN model leverages irregular EHR data by concatenating the visits' timestamps in the inputs. This enables the model to capture temporal relationships in the data and generate more informative and accurate predictions.

3.6.2 Temporal based neural models: Time Aware models

Instead of treating elapsed time as an additional variable, studies have modeled the temporal structure of time series as an inherent channel within the neural architecture. Researchers have proposed temporal-based architectures, which can be categorized into (3) classes.

The first class uses a time decay formulation to simulate the diminishing effect of acute conditions over time. Examples include C-LSTM [124] and T-GRU [153], which incorporate exponential decay functions into the forget and memory cells of LSTM and GRU, respectively. For irregular time intervals $\mathbf{\Delta}_t$ between consecutive events, the formulation is $\mathbf{\Delta}_t = \frac{1}{log(e+\Delta_t)}$.

Baytas et al [5] introduced the T-LSTM approach in "Patient Subtyping via Time-Aware LSTM Networks," framing patient subtyping as an unsupervised learning task that benefits from including temporal information. This method modifies the LSTM [63] network to decompose the previous memory cell into short-term (D-STM, $\hat{C}^S_{t-1}$) and long-term memories (LTM, C_{t-1} - C^S_{t-1}). Only the short-term memory is discounted using the decay function, defined as $g(\mathbf{\Delta}_t) = \frac{1}{log(e+\Delta_t)}$ [124]

$$C^S_{t-1} = \tanh(W_d C_{t-1} + b_d) \qquad \text{(Short term memory)} \tag{3.17}$$

$$\hat{C}^S_{t-1} = g(\Delta_t) * C^S_{t-1} \quad \text{(Discounted short term memory)} \tag{3.18}$$

The adjusted previous memory combines the discounted short-term memory and long-term memory:

$$C^*_{t-1} = \overbrace{C_{t-1} - C^S_{t-1}}^{\text{LTM}} + \underbrace{\hat{C}^S_{t-1}}_{\text{D-STM}} \quad \text{(Adjusted previous memory)} \tag{3.19}$$

The current memory is then updated as:

$$C_t = f_t \odot C^*_{t-1} + i_t \odot \tilde{C}_t \qquad \text{(Current memory)} \tag{3.20}$$

This adaptation helps the model reduce the influence of previous memory as the time gap increases, improving its understanding of temporal context. However, T-LSTM's[5] dependence on the previous cell state limits its ability to assess the relative importance of events for current predictions.

Yu Zhu et al [176] proposed a modification to the standard LSTM unit to account for time irregularities in modeling user behavior. Their goal was to predict the next item a user might interact with by considering historical behavior, context, and temporal information. They introduced a "Time Gate" into the LSTM, which controls the influence of the last consumed item on current recommendations as follows:

$$T_m = \sigma(x_m W_{x_t} + \sigma_{\Delta_t}(\Delta t_m W_{t_t}) + b_t) \tag{3.21}$$

The Time Gate incorporates user actions and time information, enabling the Time-LSTM network to capture temporal patterns and make accurate predictions about the next item, while considering the relevance of the current time step. This incorporation of time information allows the Time-LSTM to adapt to dynamic user behavior.

To handle infrequent interactions over long periods, they also modified the memory cell and output gate of the standard LSTM unit:

$$\begin{aligned} c_m &= f_m \odot c_{m-1} + i_m \odot T_m \odot \tilde{c}_m \\ o_m &= \sigma(W_{x_o}[h_{m-1}, \boldsymbol{x}_m] + \Delta t_m W_{t_o} + w_{co} \odot c_m + b_o) \end{aligned} \tag{3.22}$$

The Time Gate can adjust its weight based on recent or older time intervals, indicating varying levels of user interest in an item over time. This makes the Time-LSTM [176] framework effective in predicting user interactions while accounting for time irregularities. In clinical settings, this framework can predict patient outcomes, detect early signs of deterioration, and optimize treatment plans by tracking a patient's health journey and historical data over time.

Zhang et al [171] proposed ATTAIN, an attention-based time-aware disease progression model for early prediction of septic shock, that incorporates the attention mechanism and models the time irregularity between events. Specifically, they adjusted the previous memory cell of LSTM $\boldsymbol{c}^{t-1}$ to accumulate previous information. Instead of adding memory from one previous event, they retrospected memories of all/several previous events and discounted them by weights generated from the attention mechanism and the time intervals between those events and the current event. The overall weights represent how important each previous event is for the current event to identify the progressing condition. Three attention mechanisms are explored: global ($\boldsymbol{g}$), local ($\boldsymbol{l}$), and flexible ($\boldsymbol{f}$), to generate the attention weights. On the other hand, the time intervals are transformed to decay weights through a decay function so that the outdated events are more likely to play a less important role than recent events when predicting the outcome of the current event as follows: $\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{\boldsymbol{t-1}} \alpha_{ti} \cdot \boldsymbol{c}^i \cdot \mathrm{g}(\Delta_{t_{ti}})$ where α_{ti} is the attention weight from $i-th$ event to the current event t, $\Delta_{t_{ti}}$ is the time interval between $i-th$ event to the current event, $\mathrm{g}(\cdot)$ is a decay function, and m stands for the number of events to look backwards.

On the other hand, GRU-D [18] defines the time decay function as a model trained jointly with the GRU network. This decay rate, given by $\boldsymbol{\gamma}_t = \exp^{-max(0, W_\gamma \Delta_T + \boldsymbol{b}_\gamma)}$, is used to weigh previous hidden states when computing the current one, where W_γ and $\boldsymbol{b}_\gamma$ are learnable parameters.

However, patient healthcare trajectories exhibit diverse temporal patterns. During follow-ups, some health conditions may improve while others worsen, and new chronic diseases

can develop. GRU-D's time decay does not account for all these patterns. To address this, C-LSTM [124] introduces a flexible parametric time component in the forget gate to learn temporal dependencies between different patterns. This trainable parameter, expressed as $\boldsymbol{Q}_f q_{\Delta_t}$, is added to the forget gate parameters as follows: $\boldsymbol{f}_t = \boldsymbol{\sigma}(W_f[\boldsymbol{h}_{t-1}, x_t, \boldsymbol{Q}_f q_{\Delta_{t-1}}] + \boldsymbol{b}_f)$, where $\boldsymbol{Q}_f$ is the weight matrix and q_{Δ_t} is a temporal vector derived from the time difference Δ_t.

The third research direction, temporal structure segmentation, tackles the same problem as parametric time. Instead of using one parametric network to learn the overall temporal structure of longitudinal EHR data, researchers aim to explicitly model each temporal dependency component. Jeong et al [89]. segmented the temporal trajectory of patients into three modules: the neural abstraction module (capturing long-term distant past), the recent context module (embedding recent event information through a linear layer's discriminative projection as a binary vector), and the periodicity mechanism (representing periodic events and recent intervals between the two most recent occurrences in the current event stream). However, this explicit modeling requires a pre-processing step to segment medical events at a certain temporal granularity and compute time gaps only between periodic events.

3.7 Clinical applications

In this section, we provide a comprehensive overview of the clinical applications of machine learning algorithms in the healthcare domain. Specifically, we focus on how these algorithms, particularly RNN-based, attention-based, and transformer-based models, are utilized to perform predictive clinical analyses across various tasks.

3.7.1 Downstream tasks

The widespread adoption of Electronic Health Records (EHRs) in healthcare has garnered significant interest from the machine learning community. Researchers aim to develop evidence-based clinical decision-making tools using various machine learning (ML) algorithms [162]. Numerous studies in the medical field have demonstrated the effectiveness of using sequential-based neural networks to represent patients' healthcare trajectories and conduct predictive clinical analyses. RNN-based, attention-based, and transformer-based models have been successfully applied to various clinical event prediction tasks, which we categorize into five major families.

The first category is binary classification [5, 18, 97, 102, 105, 124, 129, 147, 153] , which aims to predict the presence, absence, or future risk of a specific clinical outcome, such as the onset of heart failure [19, 23, 25, 105] , in-hospital mortality risk [18, 114, 129], and patient readmission [97, 124, 169]. The second category is multi-label classification, where the objective is to predict multiple outcomes, such as future diagnosis categories [5] and disease classifications [5, 18, 49, 114, 129]. The third category is regression, which predicts real-valued attributes and is used to estimate the temporal progression of a patient's health, fill in missing values of numerical indicators, and predict their future values [30, 129, 133,

144]. Examples include estimating the length of hospital stay in days , Qingxiong et al [153] filled the missing numerical rates based on surrounding context and Zhengping et al [18] predicted the future values of multivariate continuous outcomes. The fourth category is medical event prediction [19, 49, 87, 89, 124], which aids doctors in understanding the evolution of a patient's health trajectory and anticipating adverse events for better patient management. Finally, recommendation systems [22, 124, 172, 175] have been proposed to assist healthcare practitioners in selecting the most appropriate treatments or procedures based on a patient's history.

3.7.2 Deep learning architectures

RNN-based models, particularly LSTMs and GRU networks [5, 18, 124, 129, 153, 171], are widely used in modeling medical time series and have demonstrated promising results compared to traditional machine learning algorithms. Specifically, Choi et al [25] developed a two-stage process where they first defined embedding vectors for medical event codes using the skip-gram model [113]. These embeddings were then fed into a GRU layer to generate patient hidden representations for predicting the risk of heart failure.

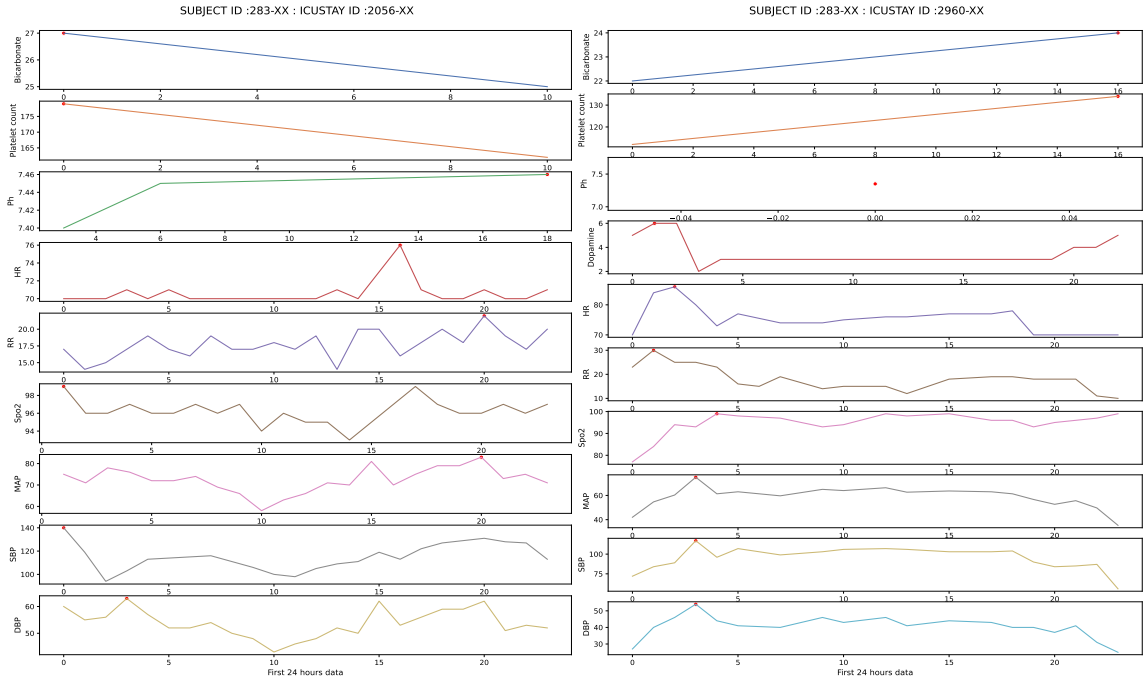
Esteban et al [35] employed an MLP projection layer to represent constant patient information while modeling the temporal dynamics of patient care history using a GRU network. They concatenated the static and dynamic hidden representations and applied a softmax layer to predict the outcomes of kidney transplantation procedures.

Ke et al [167] utilized bi-directional LSTMs (BiLSTM) to model the sequence of medical event embeddings, generating patient representations from the hidden states. They compared their approach with other aggregation methods, such as average and self-attention pooling, and found that BiLSTM yielded the best results.

More recently, researchers have explored the effectiveness of Transformer-based models for modeling medical time series, demonstrating superior performance over GRU/LSTM-based models [102, 104, 147]. Song et al [147] used a multi-dimensional embedding vector to represent each episode, encapsulating all medical event information. They processed these embeddings through a Transformer encoder with causal attention (where an observation t can only attend to past information $j < t$) and a dense interpolation layer to account for time gaps between observations. Tipirneni et al [155]. proposed a Transformer-based model featuring a novel Input Triplet Embedding component, which represents the time of the observation t, the features $(f_j)_{1\leq j\leq K}$ observed at t and their values $(v_j)_{1\leq j\leq K}$.

3.8 Limitations

As discussed in previous chapters, Electronic health records (EHR) are digital versions of patients' health information records that store a wide range of heterogeneous data related to their visits to hospitals, including diagnoses, lab results, prescribed medications, procedures, image data, demographic information by medical practitioners to report the patient's state and the medical events that occurred during his/her stay [122]. They are also char-



(a) First ICU admission (b) Second ICU admission

Figure 3.6: The illustration portrays a patient's health trajectories during his hospitalization, where the individual was admitted to the ICU twice, and each admission showed a distinct pattern. The X-axis indicates the time of measurement for each clinical feature (hourly), and the Y-axis represents the measured value.

acterized by high dimensionality and heterogeneity, ranging from continuous variables to event data such as interventions and drug administrations.

Due to the intricacies of EHRs data, current methods and standard LSTM architectures, designed for uniformly spaced sequence elements like in language modeling, face significant challenges in handling them. These methods struggle to capture temporal irregularities and retain relevant past medical history, risking the oversight of critical health episodes amid irrelevant data. In healthcare settings, the timing between patient data points varies considerably, rendering it crucial for understanding health status and disease evolution, as demonstrated in Fig. 3.6 with a patient's unique health trajectories during multiple ICU admissions. These patterns underscore the need for tailored interventions based on dynamic health statuses.

For instance, short intervals between lactate reading level may signal rapid changes and a deteriorating condition, necessitating immediate intervention. Conversely, longer intervals may indicate stability. Continuous monitoring of variables like lactate levels, norepinephrine dosage, and blood pressure at varying intervals is essential for evaluating a patient's health and treatment efficacy. Addressing the intricacies of EHR data calls for novel LSTM-based models capable of accommodating irregular timing and interactions among clinical parameters, thereby providing a more precise depiction of a patient's health trajectory.

Therefore, when developing models to learn from a patient's temporal representation

timeline, it is crucial to consider the heterogeneity and irregularity of the data, along with the complex relationships between clinical events. This can help to ensure that the model accurately reflects the patient's health status and can provide valuable insights to inform clinical decision-making.

We propose that addressing temporal irregularities and treating each feature differently based on their importance has the potential to significantly enhance the accuracy of predicting patient outcomes. This assertion is based on two key factors that underscore the advantages of this approach.

Firstly, analyzing temporal irregularities for each feature individually offers a more nuanced and precise understanding of the time span information. This enables us to gain a comprehensive insight into the evolution of each specific feature over time and its influence on patient outcomes.

Secondly, it is crucial to acknowledge that different features within the EHR data may exhibit distinct decaying patterns, with some features decaying more rapidly than others within the same time interval. Therefore, it becomes vital to handle these features in a customized and specific manner, considering their unique temporal characteristics. This approach allows us to accurately capture the diverse dynamics inherent in different features, facilitating a comprehensive analysis of their temporal behavior.

Furthermore, measuring the frequency of each clinical parameter offers a more accurate indication of a patient's management condition for several key reasons. Here are the three main reasons:

1. **Longitudinal Monitoring:** By tracking the frequency of clinical parameters over time, healthcare providers can identify trends and changes in a patient's condition. This is crucial for detecting gradual deterioration or improvement that may not be evident from a single measurement.

2. **Early Detection of Complications:** Regular monitoring can detect early signs of complications. For example, frequent glucose level checks in diabetic patients can help detect hyperglycemia or hypoglycemia early, allowing for interventions [156]. Similarly, in patients with chronic heart failure, frequent monitoring of blood pressure and heart rate can help detect worsening heart failure or the effects of medication adjustments.

3. **Personalized Medicine:** Frequent monitoring supports personalized treatment by providing detailed insights into how a patient responds to different therapies over time. This approach is beneficial in managing chronic conditions like heart failure, as it allows clinicians to make informed decisions about adjusting dosages or changing medications to achieve better outcomes.

Chapter 4

Recurrent Neural Networks & Training

We can only see a short distance ahead, but we can see plenty there that needs to be done.

– Alan Turing, Computing machinery and intelligence

In this chapter, we briefly review a family of Recurrent Neural Networks (RNN) [137], upon which this thesis is built, and an extension for the speed up training. In what follows, we start from the most basic form of RNNs and its properties (Training process). We then review Gated RNNs, a class of models with the gating mechanism to solve the long-term dependency in RNNs such as Long Short-Term Memory (LSTM) [63] and Gated Recurrent Unit (GRU) [27] and different summarise the real-world applications of Gated RNNs on

real-world events time series data.

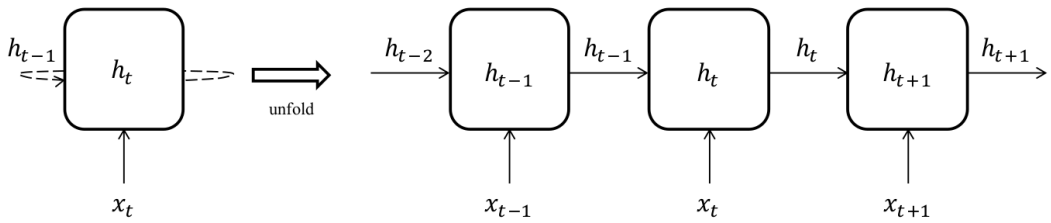


Figure 4.1: Folded graph of RNN (left) and the unfolded in time (right) during forward propagation. The new state $\boldsymbol{h}_t$ sequentially is updated by the current input $\boldsymbol{x}_t$ and the previous state $\boldsymbol{h}_{t-1}$.

4.1 Recurrent Neural Network & Properties

In general, RNNs are a learning model that updates new state $\boldsymbol{h}_t$ using previous state $\boldsymbol{h}_{t-1}$ and current input $\boldsymbol{x}_t$ recursively. It can also be described by a network, which is composed of multiple cells connecting in series along the time axis. Each cell in the network computes state at a certain time step. Fig. 4.1 illustrates forward propagation of RNN in two ways. Cyclic shape in the left is called 'folded graph' while acyclic shape in the right is called 'unfolded graph' which expands cyclic shape in time. Note that parameters in the unfolded representation share their values over the cells and the values are updated simultaneously within the optimization procedure.

4.1.1 Training RNN

In general, training means a process that a model learns the optimal parameters with a training set by minimizing defined error function which depends on trainable parameters. Training is composed of 3 steps forward propagation, backpropagation and parameter update respectively. In RNN, backpropagation step is called backpropagation through time (BPTT) [137] as the gradient of the error must be propagated through the unfolded version of the RNN graph.

4.1.1.1 Forward propagation

At each time step t, forward propagation of RNN updates values of the state $\boldsymbol{h}_t$, the forecast $\boldsymbol{p}_t$ and corresponding error $\boldsymbol{J}_t$ with respect to the input $\boldsymbol{x}_t$ and target $\boldsymbol{y}_t$. Note that the parameters, such as weights $\{\boldsymbol{W}_{hh}, \boldsymbol{W}_{xh}, \boldsymbol{W}_p\}$ and bias $\{\boldsymbol{b}_{hh}, \boldsymbol{b}_{xh}, \boldsymbol{b}_p\}$, remain unchanged during the forward propagation. To start forward propagation with sequential training data, initial state $\boldsymbol{h}_{t-1}$ is required to compute the first state $\boldsymbol{h}_t$. In general, initial state is set to zero [137]. Fig. 4.2 illustrates forward propagation process within a cell at a time step t.

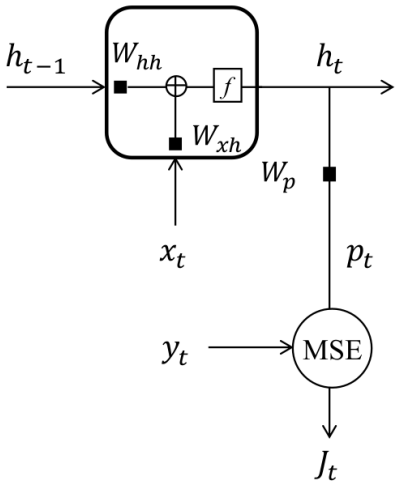


Figure 4.2: Forward propagation of a RNN at a time step t. The state $\boldsymbol{h}_t$, the forecast $\boldsymbol{p}_t$ and the error $\boldsymbol{J}_t$ are updated with parameters unchanged, such as weights $\{\boldsymbol{W}_{hh}, \boldsymbol{W}_{xh}, \boldsymbol{W}_p\}$ and bias $\{\boldsymbol{b}_{hh}, \boldsymbol{b}_{xh}, \boldsymbol{b}_p\}$, during forward propagation..

4.1.1.2 Trainable parameters

Trainable parameters can be grouped into two depending on the purpose, such as cell parameters and prediction parameters. The purpose of cell parameters are to update state $\boldsymbol{h}_t$ using previous state $\boldsymbol{h}_{t-1}$ and current input $\boldsymbol{x}_t$. On the other hand, prediction parameters exist to compute the output $\boldsymbol{p}_t$ using the updated state $\boldsymbol{h}_t$. Using dimension of $\boldsymbol{x}_t \in \mathbb{R}^{\boldsymbol{m}}$, $\boldsymbol{h}_t \in \mathbb{R}^{\boldsymbol{s}}$ and $\boldsymbol{p}_t \in \mathbb{R}^{\boldsymbol{m}}$, the dimension of the trainable parameters is specified.

$$\text{Cell parameters} : \boldsymbol{W}_{xh} \in \mathbb{R}^{\boldsymbol{s}\times m}, \boldsymbol{W}_{hh} \in \mathbb{R}^{\boldsymbol{s}\times s}, \boldsymbol{b}_h \in \mathbb{R}^{\boldsymbol{s}}, \boldsymbol{b}_{xh} \in \mathbb{R}^{\boldsymbol{s}}$$
$$\text{Prediction parameters} : \boldsymbol{W}_p \in \mathbb{R}^{\boldsymbol{m}\times s}, \boldsymbol{b}_p \in \mathbb{R}^{\boldsymbol{m}}$$

4.1.1.3 State $\boldsymbol{h}_t$

At each time step, new state $\boldsymbol{h}_t$ is updated by current input $\boldsymbol{x}_t$ and previous state $\boldsymbol{h}_{t-1}$. A set of parameters $\{\boldsymbol{W}_{hh}, \boldsymbol{W}_{xh}, \boldsymbol{b}_h\}$ participate in the update of state. Eq. (4.1) specifies the relationship, where $f(\cdot)$ is an activation function such that hyperbolic tangent or Relu. Note that all the elements in input $\boldsymbol{x}_t$ are usually normalized to have z-scores $\boldsymbol{x}_t \sim$ N(0, 1) or values within [0, 1],and expressed as follows:

$$\boldsymbol{h}_t = f(\boldsymbol{W}_{hh} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xh} \cdot \boldsymbol{x}_t + \boldsymbol{b}_h) \tag{4.1}$$

4.1.1.4 Output $\boldsymbol{p}_t$

The prediction for the next $\boldsymbol{p}_{t+1}$ is computed with current state $\boldsymbol{h}_t$ and prediction parameters $\boldsymbol{W}_p$ and $\boldsymbol{b}_p$. It is specified in Eq. (4.2), where function $g(\cdot)$ is usually linear but can be any other activation function and it needs to be selected to match the type of the target of the data. For instance, if the target variable is a multi-class variable Softmax function is

used. On the other hand, if the target is binary or is defined by a set of binary variables a sigmoid function (such as the logistic function) is used. The parameters of RNN are learned through stochastic gradient descent (SGD). Loss is determined by cross-entropy function (multi-class) or binary cross-entropy function (multi-label), summed over all time-steps of each sequence as well as across all sequences [137].

$$p_t = g(W_p \cdot \boldsymbol{h}_t + \boldsymbol{b}_p) \tag{4.2}$$

4.1.1.5 Error $\boldsymbol{J}$ and Error at a time $\boldsymbol{J}_t$

Error $\boldsymbol{J}$ is defined by mean square error (MSE) of the prediction $\boldsymbol{p}_t$ and target $\boldsymbol{y}_t$ for all time steps. Analogously, $\boldsymbol{J}_t$, an error at a time step t, is quadratic error of prediction at the time step. Both $\boldsymbol{J}$ and $\boldsymbol{J}_t$ are scalar values. Eq. (4.3) denotes the formal expression of $\boldsymbol{J}$ and $\boldsymbol{J}_t$.

$$\begin{aligned} MSE: \boldsymbol{J} &= \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{J}_t = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{m}\{(\boldsymbol{p}_t)_i - (\boldsymbol{y}_t)_i\}^2 \\ \boldsymbol{J}_t &= \sum_{i=1}^{m}\{(\boldsymbol{p}_t)_i - (\boldsymbol{y}_t)_i\}^2 \end{aligned} \tag{4.3}$$

This equation computes the MSE over all time steps T by summing the squared differences between the predicted $(\boldsymbol{p}_t)$ and target $(\boldsymbol{y}_t)$ values for each element i in the vectors at time step t, and then averaging them. $\boldsymbol{J}_t$ is the MSE at a specific time step t, and $\boldsymbol{J}$ represents the overall MSE for the entire sequence of predictions and targets.

4.1.2 Backpropagation

The purpose of backpropagation is to compute gradients that will be used for updating parameters. Gradients are derived from error $\boldsymbol{J}$ but have different calculation methods depending on the type of parameters, such as cell parameters which are used for updating the state $\boldsymbol{h}_t$ and prediction parameters used for computing the forecast $\boldsymbol{p}_t$. This is because parameters at each group participate in computing the error with different scheme. The chain rule is applied to compute gradients following the inverse direction of the forward scheme during backpropagation.

4.1.2.1 Gradients of J in terms of prediction parameters

As two parameters, $\boldsymbol{W}_p$ and $\boldsymbol{b}_p$ participate in computing the prediction using the updated state, they are not engaged in updating the state. Thus, the depth of backpropagation is bounded on the same time step when computing gradient of these parameters. That is, the gradients in terms of the prediction parameters don't backpropagate through time. Fig. 4.3 shows how the error at a time $\boldsymbol{J}_t$ is backpropagating to the prediction parameters. The gradient of J in terms of $\boldsymbol{W}_p$ can be represented by the mean value of the partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{W}_p$ over time. By the chain rule, the partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{W}_p$ becomes a product of two partial derivatives, the partial derivative of $\boldsymbol{J}_t$

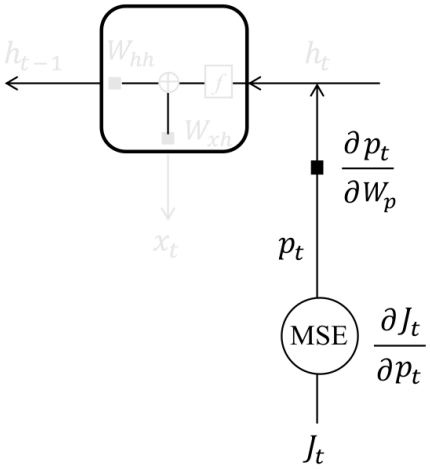


Figure 4.3: Backpropagation scheme for prediction parameters $\{\boldsymbol{W}_p, \boldsymbol{b}_p\}$. As the parameters are not engaged in updating states within the cell during forward propagation, the gradients of the error $\boldsymbol{J}_t$ in terms of prediction parameters are bounded within the time step. That is, the gradients do not backpropagate through time. Gradients in terms of the prediction parameters are computed by the chain rule.

with respect to $\boldsymbol{p}_t$ and the partial derivative of $\boldsymbol{p}_t$ with respect to $\boldsymbol{W}_p$. The former partial derivative is simplified by $2(\boldsymbol{p}_t\text{-}\boldsymbol{y}_t) \in \mathbb{R}^{\boldsymbol{m}}$ and the latter one is $\boldsymbol{h}_t \in \mathbb{R}^{\boldsymbol{s}}$, assuming that the function $g(\cdot)$ in Eq. (4.2) is linear. Considering the dimension of the gradient, cross product is operated between two vectors. Eq. (4.4) denotes the formal expression of the gradients, where $\otimes$ stands for cross product operation.

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_p} &= \frac{1}{T}\sum_{t=0}^{T-1}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{W}_p} \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{p}_t}\frac{\partial \boldsymbol{p}_t}{\partial \boldsymbol{W}_p} \\ &= \frac{2}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}_t - \boldsymbol{y}_t)\otimes \boldsymbol{h}_t\end{aligned} \tag{4.4}$$

Likewise, gradient of the error $\boldsymbol{J}$ with respect to $\boldsymbol{b}_p$ is also computed as:

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{b}_p} = \frac{2}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}_t - \boldsymbol{y}_t) \tag{4.5}$$

4.1.2.2 Gradients of J in terms of cell parameters

The purpose of a RNN cell is to update state $\boldsymbol{h}_t$ using previous state $\boldsymbol{h}_{t-1}$ and current input $\boldsymbol{x}_t$. But properties of the updated state $\boldsymbol{h}_t$ varies depending on the RNN cell architecture which means a computation process to update a new state $\boldsymbol{h}_t$ within a RNN cell.

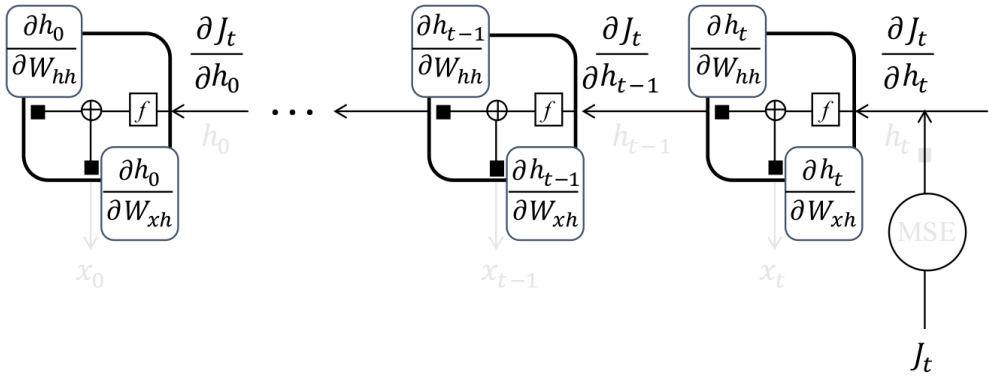


Figure 4.4: Schema of how the error $\boldsymbol{J}_t$ backpropagates to the first cell(unit) through recurrent connection with length 1, which carries gradients of the cell parameters.

The following discussion is based on the cell parameters or RNN, such as $\{\boldsymbol{W}_{xh}, \boldsymbol{W}_{hh}, \boldsymbol{b}_h\}$. Unlike prediction parameters, cell parameters are involved in updating the state $\boldsymbol{h}_t$ which propagates through time. Cell parameters compute current state $\boldsymbol{h}_t$ using previous state $\boldsymbol{h}_{t-1}$ which is computed by the same parameters and two-step previous state $\boldsymbol{h}_{t-2}$. This sequential chain lasts until it reaches to the first cell which computes $\boldsymbol{h}_0$.

Accordingly, the gradients in terms of cell parameters backpropagate through the path that the state $\boldsymbol{h}_t$ propagates through time. Theoretically, the error can backpropagate to the first cell, which is used to compute the first state $\boldsymbol{h}_0$. By the chain rule, the procedure that an error at a time step $\boldsymbol{J}_t$ backpropagates to cell parameters in every cell behind the time t can be specified. Fig. 4.4 illustrates how an error at time t, $\boldsymbol{J}_t$, backpropagates to the cell at time 0. At each time step, the gradient in terms of the state is expressed by the partial derivative of $\boldsymbol{J}_t$ with respect to the state $\boldsymbol{h}_{t-k}$, where $k = 0, 1, \ldots, t$ represents the number of time step that the error backpropagates. The gradients in terms of the cell parameters are derived from the partial derivative at each time step. It implies that the gradients of $\boldsymbol{J}_t$ in terms of cell parameters differ at each time step, unlike the same kind of parameters have identical values through time in the forward propagation scheme. Hence, for each parameter, the overall gradient is computed by the sum of gradients at every time step in order to maintain the gradient identical through time.

For the parameter $\boldsymbol{W}_{xh}$ within a cell at time step $t-k$, the gradient of an error $\boldsymbol{J}_t$ is expressed in Eq. (4.6) by the chain rule,

$$
\begin{aligned}
At \quad t-k^{th} \quad cell: \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{W}_{xh}} &= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}} \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} \\
&= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{h}_{t-2}} \cdots \frac{\partial \boldsymbol{h}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} \\
&= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t} \Big(\prod_{\tau=0}^{k-1} \frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}} \Big) \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}}
\end{aligned}
\tag{4.6}
$$

The overall gradient in Eq. (4.7) is the sum of the gradients at every time step induced by one error $\boldsymbol{J}_t$. The second last term of Eq. (4.7) within braces, products of the partial derivative,

is noteworthy because it causes vanishing or exploding gradient problem depending on its value, and that will be discussed later.

$$Overall: \quad \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{W}_{xh}} = \sum_{k=0}^{t} \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t} \left(\prod_{\tau=0}^{k-1} \frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}\right) \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} \tag{4.7}$$

The gradient of total error J is the average of gradient of an error $\boldsymbol{J}_t$ for T-length sequential data, denoted in Eq. (4.8).

$$\begin{aligned} \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_{xh}} &= \frac{1}{T} \sum_{t=0}^{T-1} \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{W}_{xh}} \\ &= \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=0}^{t} \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t} \left(\prod_{\tau=0}^{k-1} \frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}\right) \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} \end{aligned} \tag{4.8}$$

Three partial derivative terms in Eq. (4.8) can be expanded using the expressions in forward propagation.

$$\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t}, \quad \prod_{\tau=0}^{k-1} \frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}, \quad \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}}$$

The first term means the gradient of an error $\boldsymbol{J}_t$ in terms of the updated state ht at the same time step. Using expressions in Eqs. (4.2 and 4.3) , the term can be expanded in Eq. (4.9) ,where $g(\cdot)$ is linear.

$$\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t} = \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{p}_t} \frac{\partial \boldsymbol{p}_t}{\partial \boldsymbol{h}_t} = 2(\boldsymbol{p}_t - \boldsymbol{y}_t) \cdot \boldsymbol{W}_p \tag{4.9}$$

$$\begin{aligned} \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{p}_t} &= 2(\boldsymbol{p}_t - \boldsymbol{y}_t) \\ \frac{\partial \boldsymbol{p}_t}{\partial \boldsymbol{h}_t} &= \boldsymbol{W}_p \end{aligned} \iff \begin{aligned} \boldsymbol{J}_t &= \sum_{i=1}^{m} \{(\boldsymbol{p}_t)_i - (\boldsymbol{y}_t)_i\}^2 (4.3) \\ \boldsymbol{p}_t &= g(W_p \cdot \boldsymbol{h}_t + \boldsymbol{b}_p) (4.2) \end{aligned}$$

The second term, the product of the partial derivatives, implies the transmission of the gradient of $\boldsymbol{J}_t$ which backpropagates from time step t to $t-k$. As the gradient backpropagates in sequential manner, the process can be factorized by one time step backpropagation. Considering the formula of updating state $\boldsymbol{h}_t$ in Eq. (4.11) with activation function f as hyperbolic tangent, the product of factorized partial derivative is specified in Eq. (4.10) with element-wise product operator $\odot$.

$$\begin{aligned} \prod_{\tau=0}^{k-1} \frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}} &= \prod_{\tau=0}^{k-1} (1 - \boldsymbol{h}_{t-\tau}^2) \cdot \boldsymbol{W}_{hh} \\ &= \{(1 - \boldsymbol{h}_t^2) \cdot \boldsymbol{W}_{hh}\} \odot \{(1 - \boldsymbol{h}_{t-1}^2) \cdot \boldsymbol{W}_{hh}\} \ldots \\ &\quad \odot \{(1 - \boldsymbol{h}_{t-k+2}^2) \cdot \boldsymbol{W}_{hh}\} \odot \{(1 - \boldsymbol{h}_{t-k+1}^2) \cdot \boldsymbol{W}_{hh}\} \end{aligned} \tag{4.10}$$

$$\frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}} = (1 - \boldsymbol{h}_{t-\tau}^2) \cdot \boldsymbol{W}_{hh} \Leftrightarrow \boldsymbol{h}_{t-\tau} = tanh(\boldsymbol{W}_{hh} \cdot \boldsymbol{h}_{t-\tau-1} + \boldsymbol{W}_{xh} \cdot \boldsymbol{x}_{t-\tau} + \boldsymbol{b}_h) \tag{4.11}$$

The third term shows how the cell parameter $\boldsymbol{W}_{xh}$ is influenced by the backpropagated gradient at time step $t-k$. Using the expression of Eq. (4.1), the partial derivative is expanded in Eq. (4.12), where $\otimes$ represents cross product operation.

$$\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} = (1 - \boldsymbol{h}_{t-k}^2) \otimes \boldsymbol{x}_{t-k} \Leftrightarrow \boldsymbol{h}_{t-k} = tanh(\boldsymbol{W}_{hh} \cdot \boldsymbol{h}_{t-k-1} + \boldsymbol{W}_{xh} \cdot \boldsymbol{x}_{t-k} + \boldsymbol{b}_h) \tag{4.12}$$

Thus, Eq. (4.13) can be specified in the expression of Eq. (4.14) using Eqs. (4.9), 4.10, and 4.11)

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_{xh}} = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t}\left(\prod_{\tau=0}^{k-1}\frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}\right)\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{xh}} \tag{4.13}$$

$$= \frac{2}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\{(\boldsymbol{p}_t - \boldsymbol{y}_t)\cdot \boldsymbol{W}_p\}\{\prod_{\tau=0}^{k-1}(1-\boldsymbol{h}_{t-\tau}^2)\cdot \boldsymbol{W}_{hh}\}\{(1-\boldsymbol{h}_{t-k}^2)\otimes \boldsymbol{x}_{t-k}\} \tag{4.14}$$

Likewise, gradients in terms of $\boldsymbol{W}_{hh}$ and $\boldsymbol{b}_h$ are given in Eq. (4.15) and Eq. (4.16) respectively.

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_{hh}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t}\left(\prod_{\tau=0}^{k-1}\frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}\right)\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{W}_{hh}} \\ &= \frac{2}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\{(\boldsymbol{p}_t - \boldsymbol{y}_t)\cdot \boldsymbol{W}_p\}\{\prod_{\tau=0}^{k-1}(1-\boldsymbol{h}_{t-\tau}^2)\cdot \boldsymbol{W}_{hh}\}\{(1-\boldsymbol{h}_{t-k}^2)\otimes \boldsymbol{h}_{t-k-1}\}\end{aligned} \tag{4.15}$$

$$\begin{aligned}\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{b}_h} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t}\left(\prod_{\tau=0}^{k-1}\frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}}\right)\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{b}_h} \\ &= \frac{2}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\{(\boldsymbol{p}_t - \boldsymbol{y}_t)\cdot \boldsymbol{W}_p\}\{\prod_{\tau=0}^{k-1}(1-\boldsymbol{h}_{t-\tau}^2)\cdot \boldsymbol{W}_{hh}\}(1-\boldsymbol{h}_{t-k}^2)\end{aligned} \tag{4.16}$$

4.1.2.3 Parameters Update using Gradient Descent Algorithm

For parameters such that $\boldsymbol{W}_{hh}$, $\boldsymbol{W}_{xh}$, $\boldsymbol{W}_p$, $\boldsymbol{b}_h$ and $\boldsymbol{b}_p$, BPTT identifies each partial derivatives with respect to the error J. Identified gradients are used to update parameters using gradient decent algorithm with learning rate μ.

$$\boldsymbol{W}_{new} = \boldsymbol{W}_{old} - \mu\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_{old}} \tag{4.17}$$

4.1.2.4 Vanishing or Exploding Gradient

To ensure local stability, the network must operate in a ordered regime [9, 10]. However, the product of partial derivatives in Eq. (4.10) can cause the gradient to vanish or explode, that deteriorate training procedure. Two factors, length of the time step that error backpropagates and the value of the partial derivative are engaged in the problem [62]. While an error backpropagates from the cell corresponding to the current time step t to the first cell corresponding to $t = 0$, the partial derivative term is multiplied by itself at every time step.

$$f^{'}(t,\tau)\frac{\partial \boldsymbol{h}_{t-\tau}}{\partial \boldsymbol{h}_{t-\tau-1}} \tag{4.18}$$

For the partial derivative term $f^{'}(t,\tau)$ in Eq. (4.18), if the error backpropagates through sufficiently long time steps upon condition $|f^{'}(t,\tau)| < 1$, the gradient vanishes to a very

small value near zero before computing the appropriate gradient of the parameters in the first cell. On the other hand, an error that backpropagates upon condition $|f'(t,\tau)| > 1$ many time steps, it explodes to very large value, that causes the parameters never reach their optimal values, which hinder error function converge upon its minimum. In general, the network enters into a chaotic regime, where its computational capability is hindered [101].

Several solutions are proposed for the problem. One is to apply backpropagation on chunked sequences with a limited number of time steps (Truncated-BPTT) [121, 152, 161]. Another one is to add gates to produce paths where gradients can flow more constantly in longer-term without vanishing or exploding such as Long Short-Term Memory (LSTM) [63] and Gated Recurrent Units (GRU) [20].

4.1.3 Truncated backpropagation through time (Truncated BPTT)

Mini-batch training is a well-known method in neural networks for the advantages in terms of efficiency and robustness. However, to the best of my knowledge, mini-batch training has not been specifically described for Sequence tasks using RNNs. This section explains mini-batch training based on the learning algorithm called truncated backpropagation through time (truncated BPTT).

4.1.3.1 Truncated BPTT $(\boldsymbol{k}_2, \boldsymbol{k}_1)$

To avoid vanishing or exploding gradient problem, the time steps that RNNs backpropagate at a time in an unfolded graph is restricted. Roughly, a long sequence should be divided into several chunks with same length that RNNs can learn if the time series is longer than its limit. RNNs repeat forward and backpropagation for chunks that are sequentially fed. It is important to keep the sequential continuity between chunks because the last state at a chunk carries information of the sequence processed so far to the next chunk. Accordingly, shuffing the order of the chunks is not possible, unlike traditional Neural Networks. This learning procedure is called truncated backpropagation through time (truncated BPTT) [160].

Fig. 4.5 illustrates a scheme of truncated BPTT for a sequence with length T. The figure shows that a long sequence with length T which is divided into two chunks with length $\boldsymbol{k}_2$. The chunks are fed into the RNNs sequentially, the state $\boldsymbol{h}_{k_2-1}$ computed by the last input of the first chunk $\boldsymbol{x}_{k_2-1}$ is transferred to the initial state of the second chunk to compute $\boldsymbol{h}_{k_2}$. Two important hyperparameters are determined a-priori for truncated BPTT, which are the length of backward pass $\boldsymbol{k}_2$ and the length of forward pass $\boldsymbol{k}_1$.

4.1.3.1.1 Length of Backward Pass $\boldsymbol{k}_2$

The length of the backward pass $\boldsymbol{k}_2$ is the time steps at which the RNNs backpropagate. The length of $\boldsymbol{k}_2$ should be determined after considering whether the RNNs enable to backprop-

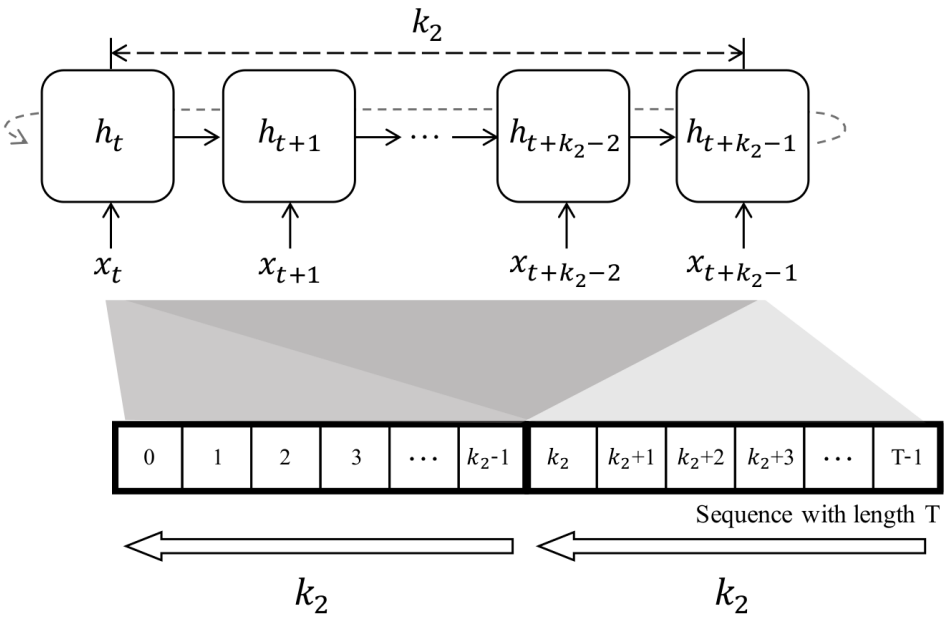


Figure 4.5: Folded graph of truncated $(\boldsymbol{k}_2, \boldsymbol{k}_1)$. The RNNs are fed by truncated sequences with length $\boldsymbol{k}_2$. The truncated sequences should be fed into network in sequential manner because the last state at a chunk carries information of the sequence processed so far to the next chunk.

agate to $\boldsymbol{k}_2$ time steps without vanishing or exploding gradient. In general, the RNNs using cells of gated architecture, such as LSTM and GRU, can backpropagate without vanishing gradient farther than the RNNs using Elman cell.

The length that RNNs can backpropagate also depends on the property of the input to learn, such as the expected maximum extent of time dependencies in the sequence or complexity. For example, in a periodic time series with period t, it may be unnecessary or even detrimental to set $\boldsymbol{k}_2 > t$ [7]. Meanwhile, a too small $\boldsymbol{k}_2$ can increase unnecessary computational cost because it makes the RNNs update parameters too often. The too often parameter update can make the RNNs concentrate more on the local minima within the chunk than global minima which can be obtained based on the long-term dependencies in some cases. Accordingly, it requires more iteration of training until the RNNs converge to the global minima. Therefore, the length of the backward pass $\boldsymbol{k}_2$ must be carefully tuned to achieve an effective training.

4.1.3.1.2 Length of Forward Pass $\boldsymbol{k}_1$

A chunk as shown in Fig. 4.5 can be interpreted as a sampled sequence by time window with length $\boldsymbol{k}_2$, where the window moves forward for every $\boldsymbol{k}_2$ time steps. The length of the window represents the length of backward pass. The length that the window moves forward is defined as the length of forward pass. Hence, the learning procedure is referred to truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_2$), where the first and second arguments are the length of backward and forward pass respectively.

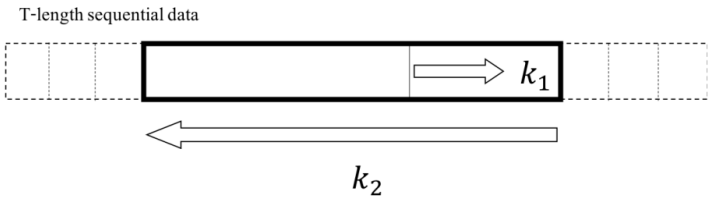


Figure 4.6: Procedure of a chunk generation for truncated BPTT $(\boldsymbol{k}_2, \boldsymbol{k}_1)$. The RNNs learns from the chunks with length $\boldsymbol{k}_2$ and a new chunk created for each $\boldsymbol{k}_1$ time steps.

Even though truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_2$) enables the RNNs to learn a long sequence, it still has a drawback. For an arbitrary sequence, truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_2$) does not guarantee that the gradients for all chunks can be backpropagated to the length of backward pass $\boldsymbol{k}_2$ without vanishing or exploding. That is, for some chunk in the sequence, the gradient may not be fully backpropagated, that harms the fidelity of learning.

The length of forward pass $\boldsymbol{k}_1$, where $\boldsymbol{k}_2 > \boldsymbol{k}_1 \geq 1$, is introduced to improve the drawback. Truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_1$) backpropagates to the length of $\boldsymbol{k}_2$ for every $\boldsymbol{k}_1$ time steps. Fig. 4.6 shows how a long sequence is transformed to chunks to be learned by truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_1$). Note that a chunk of truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_1$) have overlapped information of length $\boldsymbol{k}_2$-$\boldsymbol{k}_1$ with neighboring chunks, unlike the chunk of truncated BPTT($\boldsymbol{k}_2$, $\boldsymbol{k}_2$). This redundancy, obtained from the overlapped information, alleviates the impact that occurs in the drawback where the gradient is not fully backpropagated.

Truncated BPTT($\boldsymbol{k}_2$, 1), referred to true truncated BPTT($\boldsymbol{k}_2$), learns the sequence in the most detail but requires expensive computational cost that may be unnecessary. [160] reports that truncated BPTT($2h$, h) is a good trade-off between accuracy and computational cost, giving comparable accuracy with truncated BPTT($\boldsymbol{k}_2$, 1) and speed advantage.

4.2 Bidirectional RNN

The methods covered in this section are based on an assumption that current state is dependent on past states. That is why, we update and carry information in the hidden state $\boldsymbol{h}_t$ on forward direction from past to current (future) time $(t = 1, \ldots, T)$. Meanwhile, for some application areas such as sentence classification or speech recognition in NLP, the dependencies between inputs and outputs might be more complex, e.g., an output depends not only on the previous inputs but also on the future inputs or on the whole sequence. For example, in handwriting recognition, an unclear letter can be recognised more easily if the information of the letters is located before and after it is given information captured in the reverse direction from future to past $(t = T, \ldots, 1)$. The sequential structure of standard RNNs prevents the current representation reaching the future inputs.

Therefore, based on this motivation and to overcome this limitation, Bidirectional RNNs (BRNN) were first proposed in [141], which extends a regular RNN by introducing a backward RNN with the hidden states connecting in the opposite direction. The computational order is backward from the last input to the first. The information from both directions is

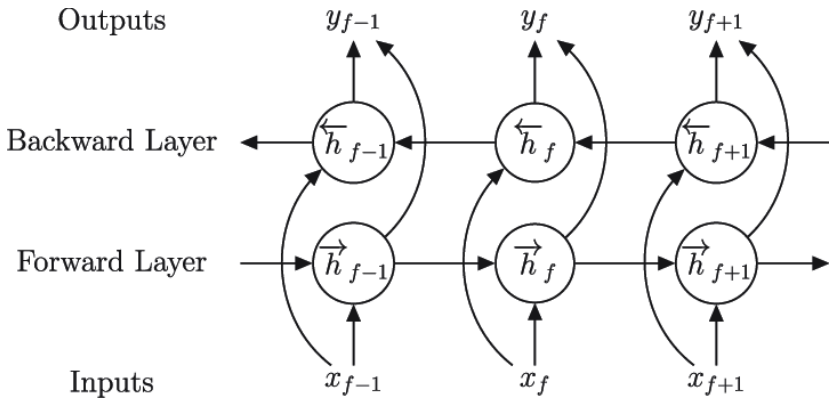


Figure 4.7: Bi-directional RNN. Note that, for brevity, the parameter matrices of the model are not represented in the figure.

then combined to produce a final output, as shown in Fig. 4.7. It independently updates two hidden states $h_{\overrightarrow{t}}$ and $h_{\overleftarrow{t}}$ in opposing forward and backward directions.:

$$\begin{aligned} \boldsymbol{h}_{\overrightarrow{t}} &= f(p_t, \boldsymbol{h}_{\overrightarrow{t-1}}), \qquad for \qquad t = 2, ..., T \\ \boldsymbol{h}_{\overleftarrow{t}} &= f(p_t, \boldsymbol{h}_{\overleftarrow{t-1}}), \qquad for \qquad t = T-1, ..., 1 \end{aligned} \tag{4.19}$$

where f(·) denotes an operation to update the hidden states such as Eq. (4.1) for RNN and Eq. (4.25) for LSTM or GRU. Two hidden states $\boldsymbol{h}_{\overrightarrow{t}}$ and $\boldsymbol{h}_{\overleftarrow{t}}$ are updated independently and once they are computed, the final hidden state is computed as a concatenation of the two:

$$\boldsymbol{h}_t = [\boldsymbol{h}_{\overrightarrow{t}}, \boldsymbol{h}_{\overleftarrow{t}}], \qquad for \qquad t = 1, ..., T \tag{4.20}$$

BRNN has been used effectively in many NLP applications such as phoneme classification [47], text-to-speech synthesis [38] and information extraction from free medical texts [68]. In the clinical domain, it is used to classify diagnosis codes based on a sequence of clinical events. Note that unlike how it is used in NLP, BRNN might not be directly applicable to the prediction of the event time series as the future information is not available at the time of the prediction. Similar to RNNs, Bidirectional RNNs (BRNNs) also experience issues with vanishing and exploding gradient problems, when learning from long sequences.

4.2.1 Challenges & Solutions

With more non-linear hidden layers, deep networks can theoretically model functions with higher complexity [26]. However, learning standard RNNs with many hidden layers is notoriously difficult [81], because they may suffer from vanishing gradients for long sequences [64], making gradient-based learning ineffective. A major reason is that many layers of non-linear transformations prevent the data signals and the gradients from flowing easily through the network. In the forward direction from data to outcomes, a change in data signals may not lead to any change in the outcomes, leading to the poor credit assignment problem. In the backward direction, a large error gradient at the outcomes may not be propagated back to the data signals. As a result, learning stops prematurely without returning an informative mapping from inputs to outputs. There have been several effective methods

to tackle the problem. The first line of work is to use non-saturated non-linear transforms such as rectified linear units (ReLUs) [42, 44, 56], whose gradients are non-zeros for a large portion of the input space. Another approach that also increases the level of linearity of the information propagation is through a *gating mechanism* [45, 63]. The gates are extra control neural units that let part of the information pass through a channel. They are learnable and have played an important role in state-of-the-art FNN architectures such as Highway Networks, which is an extension of FNNs, They can be used as recurrent models for single vector input in an iterative estimation scheme [149] and Residual Networks [57], and recurrent architectures such as Long Short-Term Memory (LSTM) [63] and Gated Recurrent Unit (GRU) [27]. In the next section, we review a family of models that utilise the *gating mechanism*.

4.2.2 Closing remarks

We have briefly reviewed the related background necessary for this thesis. We have provided an overview of the most basic form of neural networks and the training procedure with back-propagation and gradient descent, followed by a review of the popular regularisation methods used in deep neural networks, embedding methods and different complex data structures. In the next chapter, we narrow the subject to the main focus of this thesis, models that utilise the *gating mechanism* more specifically Long Short-Term Memory (LSTM) [45, 63] and their extensions such as Gated Recurrent Unit (GRU) [27] that leverage the models to solve a wide range of difficult tasks with better training.

4.3 Advanced RNN Architectures

Gating mechanisms like those in LSTM and GRU models [63, 27] effectively tackle long-term dependencies and non-linear transformations in deep neural networks. These mechanisms create learnable gates that regulate information flow within the network, enhancing linearity of information propagation. They are pivotal components in modern architectures such as Highway Networks [149], Residual Networks [57], LSTM [63], and GRU [27]. LSTM, introduced in 1997, has only recently become pivotal in sequential models [45, 46, 48].

LSTM and GRU, introduced by [63] and [27] respectively, excel in managing longer dependencies compared to traditional RNNs. They achieve this by redesigning the RNN cell architecture to mitigate issues like vanishing and exploding gradients. This enhancement is achieved through gated structures that update information without solely relying on non-linear activation functions, as illustrated in Fig. 4.8.

4.3.1 Long Short-Term Memory (LSTM)

LSTM[63] is a neural network architecture with recurrent connections that effectively handles the challenge of capturing long-term dependencies. It achieves this by introducing

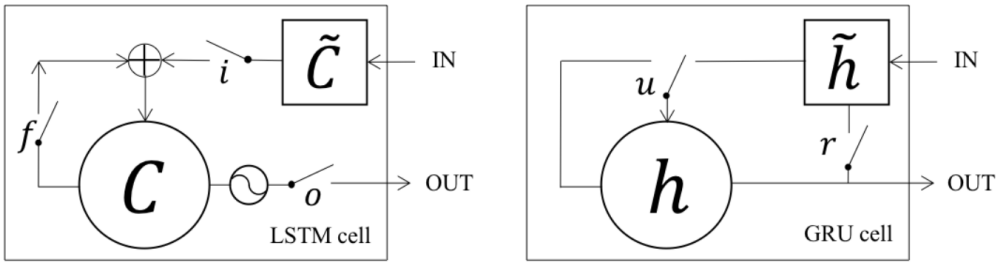


Figure 4.8: Schema of two RNN cells, LSTM (left) and GRU (right). GRU has a simpler architecture with the less number of gates than LSTM.

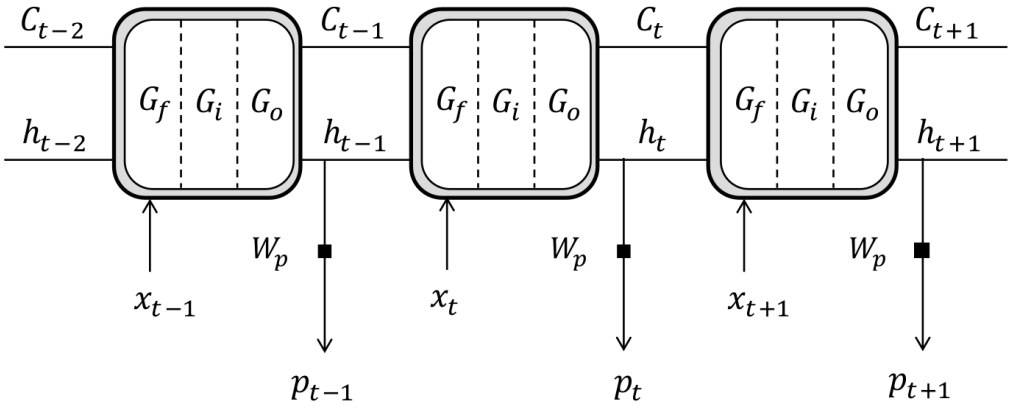


Figure 4.9: Unfolded graph of a RNN with LSTM cell that consist of three gates.

"*gates*" that regulate the information flow within the network. These *gating mechanisms* also mitigate the issue of vanishing or exploding gradients by re-parameterizing the recurrent model. Consequently, LSTM networks are particularly efficient at making predictions based on sequential data as demonstrated in studies such as [5, 18, 124]. Fig. 4.9 shows the unfolded graph of a RNN with LSTM cells.

4.3.1.1 Forward Propagation of LSTM

Fig. 4.10 depicts the architecture of LSTM cell when the RNNs propagate forward and Table 4.1 shows all trainable parameters and variables of LSTM cell with dimensional information. The three *gating mechanisms* are utilized to control the memory cell $\boldsymbol{c}_t$ and govern the flow of information for input, forgetfulness, and output operations at each time step. They act as selective filters, allowing or inhibiting the passage of information through the LSTM cell. Each LSTM cell comprises four components, as shown in Fig. 4.10. The input gate, denoted as $\boldsymbol{i}_t$, regulates the amount of new memory content added to the cell. The forget gate denoted as $\boldsymbol{f}_t$, determines the extent to which previous memory should be discarded. The output gate, denoted as $\boldsymbol{o}_t$, modulates the amount of memory content to be output. Additionally, the cell activation vector $\boldsymbol{C}_t$ consists of two components: the

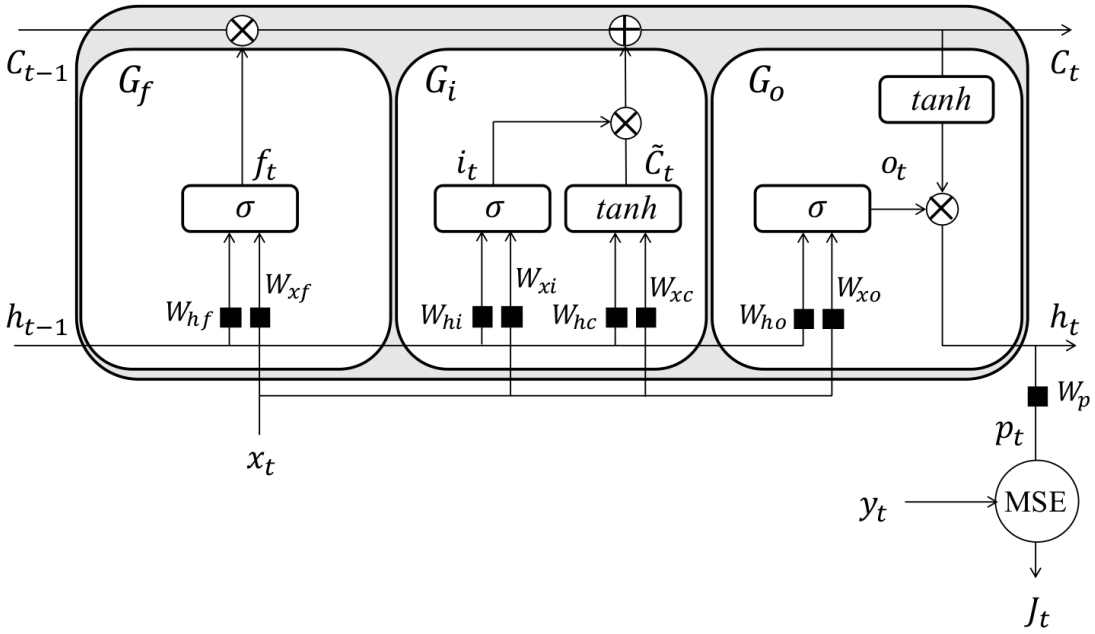


Figure 4.10: LSTM cell architecture. Note that the bypass without non-linear activation function for cell state $\boldsymbol{C}_t$ enables to avoid vanishing or exploding gradient problem and backpropagate the gradients to the further past [21].

Table 4.1: Variables and trainable parameters of LSTM cell.

Gate type	Variable	Parameters		
Forget gate $\boldsymbol{G}_f$	$\boldsymbol{f}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xf} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hf} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_f \in \mathbb{R}^s$
Input gate $\boldsymbol{G}_i$	$\boldsymbol{i}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xi} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hi} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_i \in \mathbb{R}^s$
	$\tilde{\boldsymbol{C}}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xc} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hc} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_c \in \mathbb{R}^s$
Output gate $\boldsymbol{G}_o$	$\boldsymbol{o}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xo} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{ho} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_o \in \mathbb{R}^s$
Prediction	$\boldsymbol{p}_t \in \mathbb{R}^m$	$\boldsymbol{W}_p \in \mathbb{R}^{m\times s}$		$\boldsymbol{b}_p \in \mathbb{R}^m$

partially forgotten previous memory $\boldsymbol{C}_{t-1}$ and the modulated new memory $\tilde{\boldsymbol{C}}_t$ facilitates effective memory management within the LSTM cell.We'll be discussing about each gate in the upcoming section.

4.3.1.1.1 Cell State $\boldsymbol{C}_t$

Cell state, which conveys the processed information so far to the next cell, plays a similar role like state $\boldsymbol{h}_t$ in RNN. However, the state $\boldsymbol{h}_t$ in RNN has non-linear property due to the activation function f(·) while cell state $\boldsymbol{C}_t$ in LSTM is expressed by a linear combination as shown in Eq. (4.21).

$$\boldsymbol{C}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{C}}_t \tag{4.21}$$

The additive nature of the memory cell sequence ensures linear gradient updates through the chain rule, preventing vanishing or exploding gradients during BPTT phase. As its

partial derivative doesn't generate product terms, therefore LSTM doesn't suffer from vanishing or exploding gradient problem like RNN does.

Instead, cell state is regulated by three variables $\boldsymbol{f}_t$, $\boldsymbol{i}_t$ and $\tilde{\boldsymbol{C}}_t$ which come from forget and input gate.

The memory cell's $\boldsymbol{C}_t$ ability to retain past experiences is attributed to the learnable forgetting gate $\boldsymbol{f}_t$. When $\boldsymbol{f}_t \to 1$, and $\boldsymbol{i}_t \to 0$, $\boldsymbol{C}_t$ only inherits from its previous cell state $\boldsymbol{C}_{t-1}$. In other words, cell state keeps same information between t and $t-1$, which can be interpreted as short-term memory all past memories are preserved, and new memories are updated with fresh inputs. Conversely, when $\boldsymbol{f}_t \to 0$, only the new experiences are incorporated, rendering the system memoryless. Furthermore, by optimizing parameters regarding the variables, the short-term memory can last for long period of time. Thus, for ideally trained LSTM network with a sequence with length T , cell state can reserve closely identical information through T-cells, that is why this RNN model is called long short-term memory.

4.3.1.1.2 Forget Gate $\boldsymbol{f}_t$

The role of the forget gate $\boldsymbol{f}_t$ is crucial in an LSTM unit as it creates a shortcut for the previous information flowing easily through the network. The range of the element values of $\boldsymbol{f}_t$ is restricted within the range (0, 1), so that a small portion of previous memory cell $\boldsymbol{C}_{t-1}$ can be add to the current memory cell $\boldsymbol{C}_t$ due to the sigmoid function $\sigma(\cdot)$ as shown in Eq. (4.22) by the element-wise multiplication. For example, at step t, the contribution of the first memory cell to the current memory state is $\boldsymbol{f}_2 * \ldots * \boldsymbol{f}_t * \boldsymbol{C}_1$. This enables the gradients to flow more directly to the inputs of the first steps, hence, effectively capturing the long-term dependency in long sequences.

$$\boldsymbol{f}_t = \boldsymbol{\sigma}(\boldsymbol{W}_{hf} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xf} \cdot \boldsymbol{x}_t + \boldsymbol{b}_f) \tag{4.22}$$

4.3.1.1.3 Input Gate $\boldsymbol{i}_t$ & New candidate memory $\tilde{\boldsymbol{C}}_t$

At each time step t, the input features $\boldsymbol{x}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$ are passed through a sigmoid and tanh function to compute the the new input features $\boldsymbol{i}_t$ Eq. (4.23) and modulated new candidate memory input features $\tilde{\boldsymbol{C}}_t$ Eq. (4.24). The element-wise product of the variable and the candidate participates in updating cell state $\boldsymbol{C}_t$ by adding to the element-wise product of $\boldsymbol{f}_t$ and $\boldsymbol{C}_{t-1}$ as shown in Eq. (4.21).

$$\boldsymbol{i}_t = \boldsymbol{\sigma}(\boldsymbol{W}_{hi} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xi} \cdot \boldsymbol{x}_t + \boldsymbol{b}_i) \tag{4.23}$$

$$\tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_{hc} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xc} \cdot \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{4.24}$$

The candidate $\tilde{\boldsymbol{C}}_t$ plays a role to collect and update new information from the training input data $\boldsymbol{x}_t$, like $\boldsymbol{h}_t$ in RNN. However, unlike $\boldsymbol{h}_t$ of RNN which conveys information to the next cell directly, the candidate $\tilde{\boldsymbol{C}}_t$ is adjusted by the variable $\boldsymbol{i}_t$ which element values spanned within 0 to 1. As definition of $\boldsymbol{i}_t$, shown in Eq. (4.23), is identical to the definition of $\boldsymbol{f}_t$ Eq. (4.22), which plays a similar role like $\boldsymbol{f}_t$ by determining the portion of the candidate $\tilde{\boldsymbol{C}}_t$ to be updated (added) to the cell state $\boldsymbol{C}_t$.

4.3.1.1.4 State h_t & Output Gate Variable o_t

At each time step t, a new output value is generated by $\boldsymbol{o}_t$ using current the input features $\boldsymbol{x}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$ unlike $\boldsymbol{i}_t$ and $\boldsymbol{f}_t$ does as shown in Eq. (4.26) and the current hidden state $\boldsymbol{h}_t$ use this new ouput generated by $\boldsymbol{f}_t$ and the cell state $\boldsymbol{C}_t$ to compute the prediction value for the current step as shown in Eq. (4.25).

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{C}_t) \tag{4.25}$$

$$\boldsymbol{o}_t = \boldsymbol{\sigma}(\boldsymbol{W}_{ho} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xo} \cdot \boldsymbol{x}_t + \boldsymbol{b}_o) \tag{4.26}$$

However, the information delivered by $\boldsymbol{h}_t$ is originated from the cell state $\boldsymbol{C}_t$ which keeps processed information so far and has only linear property, therefore state $\boldsymbol{h}_t$ requires activation function that enables the network to keep the non-linear property that allows the cells to be connected in series along the time. Therefore, hyperbolic tangent is applied to the cell state as an activation function, as shown in Eq. (4.25).

In summary, the variable $\boldsymbol{o}_t$ determines the portion of $\tanh(\boldsymbol{C}_t)$ to be delivered to the next cell by element-wise multiplication, as $\boldsymbol{f}_t$ and $\boldsymbol{i}_t$ do. These three variables can be interpreted as a dimming switch and their operating characteristic is defined by their parameters which are optimized during backpropagation phase.

4.3.1.1.5 Output p_t & error J

As the forecast $\boldsymbol{p}_t$ and error $\boldsymbol{J}_t$ are not included in LSTM cell, these expressions are identical to the ones for RNN. Please refer to Eqs. (4.2 and 4.3).

4.3.1.2 Back Propagation

For the LSTM network, parameters can be discriminated into two groups depending on the approach of computing gradients, prediction parameters, and cell parameters. Computing gradients of the prediction parameters such as $\boldsymbol{W}_p$, doesn't require BPTT. On the other hand, computing gradients of the cell parameters, such as $\boldsymbol{W}_{xf}$, $\boldsymbol{W}_{xi}$, $\boldsymbol{W}_{xc}$, $\boldsymbol{W}_{xo}$ etc, requires BPTT.

In terms of BPTT for the cell parameters, the gradients backpropagate through two paths, cell state $\boldsymbol{C}_{t-k}$ and state $\boldsymbol{h}_{t-k}$, where index $k = 0, 1, \ldots t$ denotes the number of time step that the gradients backpropagate. For the cell corresponding the $t-k^{th}$ time step which counts from the beginning, the gradients in terms of cell parameters are only dependent on the gradients in terms of cell state $\boldsymbol{C}_{t-k}$ and state $\boldsymbol{h}_{t-k}$.

In this section describes the gradients in terms of cell parameters following by the discussion of the gradients in terms of prediction parameters. Cell parameters are distinguished to two, depending on the paths where the gradients backpropagate, cell state $\boldsymbol{C}_{t-k}$ or state $\boldsymbol{h}_{t-k}$.

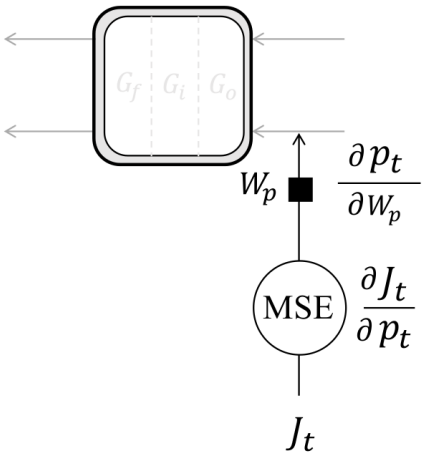


Figure 4.11: Gradient of $\boldsymbol{J}_t$ in terms of prediction parameters $\{\boldsymbol{W}_p, \boldsymbol{p}_p\}$. As prediction parameters locate out of the recurrent connections that enable the information over time, the error at a time step $\boldsymbol{J}_t$ only influences prediction parameters at the same time step, that is, the gradients of prediction parameters don't backpropagate through time.

4.3.1.2.1 Gradients of J in terms of Prediction Parameters

Fig. 4.11, shows how the error at one time step t is back-propagating to the prediction parameters. As illustrated in the Fig. 4.11, prediction parameters are located out of the LSTM cell. Thus, the parameters are not influenced by the gradients that backpropagate through time, unlike cell parameters and their backpropagation scheme is identical with the one of RNN, as shown in Eq. (4.4). For a sequence with length T , the gradients of total error $\boldsymbol{J}$ in terms of the prediction parameters are given in Eq. (4.27).

$$\begin{aligned} \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{W}_p} &= \frac{1}{T}\sum_{t=0}^{T-1}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{W}_p} = \frac{2}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}_t - \boldsymbol{y}_t)\otimes \boldsymbol{h}_t \\ \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{b}_p} &= \frac{2}{T}\sum_{t=0}^{T-1}(\boldsymbol{p}_t - \boldsymbol{y}_t) \end{aligned} \tag{4.27}$$

4.3.1.2.2 Gradients of J in terms of Cell Parameters

For the $t-k^{th}$ cell from the beginning, gradients in terms of cell parameters can be discriminated into two groups depending on the path where the gradients backpropagate : Path A, gradients through state $\boldsymbol{h}_{t-k}$, as shown in Fig. 4.12 and path B, gradients through cell state $\boldsymbol{C}_{t-k}$, as depicted in Fig. 4.13. The parameters in the output gate are involved in the path A, while the parameters in the input gate and forget gate are involved in the path B.

- **Path A : Gradients through $\boldsymbol{h}_{t-k}$**

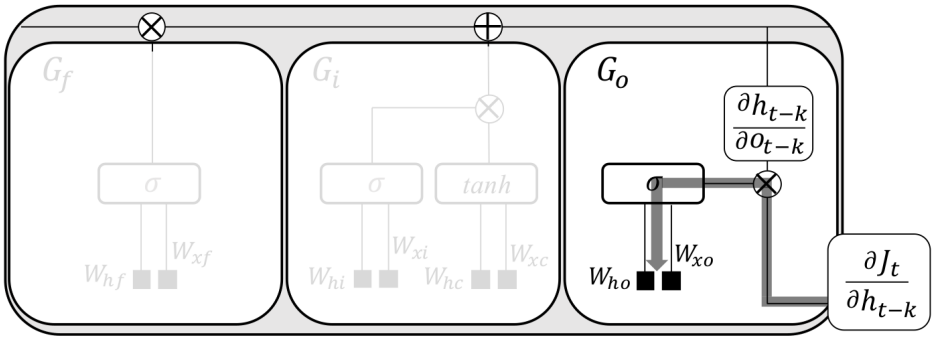


Figure 4.12: Back-propagating error through state $\boldsymbol{h}_{t-k}$. Output gate $\boldsymbol{G}_o$ only involves in this backpropagation. Gradients in terms of output gate parameters $\{\boldsymbol{W}_{xo}, \boldsymbol{W}_{ho}, \boldsymbol{b}_o\}$ can be expressed by the chain rule that is derived from the partial derivative of an error $\boldsymbol{j}_t$ with respect to $\boldsymbol{h}_{t-k}$ and $\boldsymbol{h}_{t-k}$ with respect to $\boldsymbol{o}_{t-k}$.

The gradients of output gate parameters, such as $\boldsymbol{W}_{xo}$, $\boldsymbol{W}_{ho}$ and $\boldsymbol{b}_o$, backpropagate through the path of $\boldsymbol{h}_{t-k}$ and $\boldsymbol{o}_{t-k}$. Overall gradient for a sequence with length T is expressed by the chain rule of three partial derivatives in Eq. (4.28), where the last two partial derivatives are expanded in Eq. (4.29). Discussion regarding the first partial derivative $\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}}$ will be followed later.

$$\begin{aligned}\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{xo}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}}\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{o}_{t-k}}\frac{\partial \boldsymbol{o}_{t-k}}{\partial \boldsymbol{W}_{xo}} \\ &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}} \odot tanh(\boldsymbol{C}_{t-k}) \odot \boldsymbol{o}_{t-k} \odot (1-\boldsymbol{o}_{t-k}) \otimes \boldsymbol{x}_{t-k}\end{aligned} \tag{4.28}$$

$$\begin{aligned}\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{o}_{t-k}} &= tanh(\boldsymbol{C}_{t-k}) \\ \frac{\partial \boldsymbol{o}_{t-k}}{\partial \boldsymbol{W}_{xo}} &= \boldsymbol{o}_{t-k} \odot (1-\boldsymbol{o}_{t-k}) \otimes \boldsymbol{x}_{t-k}\end{aligned} \quad \Longleftrightarrow \quad \begin{aligned}\boldsymbol{h}_t &= \boldsymbol{o}_t \odot tanh(\boldsymbol{C}_t) \\ \boldsymbol{o}_t &= \sigma(\boldsymbol{W}_{ho} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xo} \cdot \boldsymbol{x}_t + \boldsymbol{b}_o)\end{aligned} \tag{4.29}$$

For other parameters in the output gate $\{\boldsymbol{W}_{ho}, \boldsymbol{b}_o\}$, their gradients are in Eq. (4.30).

$$\begin{aligned}\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{ho}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}} \odot tanh(\boldsymbol{C}_{t-k}) \odot \boldsymbol{o}_{t-k} \odot (1-\boldsymbol{o}_{t-k}) \otimes \boldsymbol{h}_{t-k-1} \\ \frac{\partial \boldsymbol{J}}{\partial \boldsymbol{b}_o} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}} \odot tanh(\boldsymbol{C}_{t-k}) \odot \boldsymbol{o}_{t-k} \odot (1-\boldsymbol{o}_{t-k})\end{aligned} \tag{4.30}$$

- **Path B : Gradients through $\boldsymbol{C}_{t-k}$**

As the parameters in forget and input gate contribute to update cell state in the forward propagation, the gradients backpropagate through the same path. But the effect of the gradient through the cell state differs between input and forget gates. For the parameters

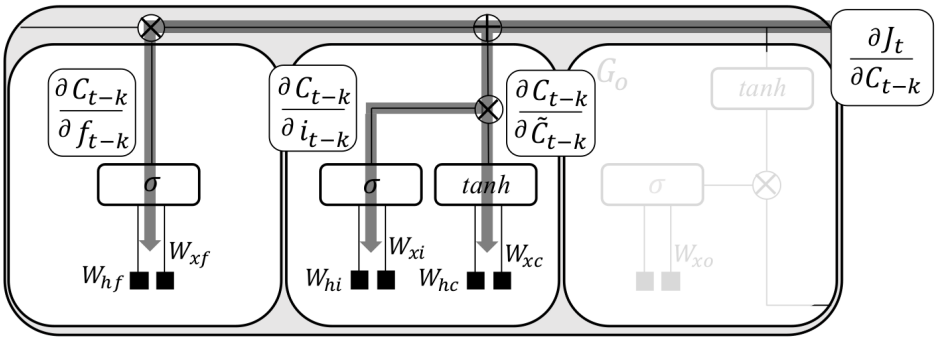


Figure 4.13: Back-propagating error through state $\boldsymbol{C}_{t-k}$. Forget and input gate, $\boldsymbol{G}_f$ and $\boldsymbol{G}_i$ involve in this backpropagation. Gradients in terms of parameters can be expressed by the chain rule that is derived from the partial derivative of an error $\boldsymbol{j}_t$ with respect to $\boldsymbol{C}_{t-k}$ and others depending on the parameters.

in input gate $\{\boldsymbol{W}_{hi}, \boldsymbol{W}_{xi}, \boldsymbol{W}_{hc}, \boldsymbol{W}_{xc}, \boldsymbol{b}_i, \boldsymbol{b}_c\}$, the backpropagating gradients branch off into two, one for the variable $\boldsymbol{i}_{t-k}$ and one for the candidate $\tilde{\boldsymbol{C}}_{t-k}$. As they experience different activation functions and paths, their expressions differ. Overall gradients of input gate parameters for a sequence with length T is expressed in Eqs. (4.31 and 4.32).

$$
\begin{aligned}
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{xi}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{i}_{t-k}}\frac{\partial \boldsymbol{i}_{t-k}}{\partial \boldsymbol{W}_{xi}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}} \odot \tilde{\boldsymbol{C}}_{t-k} \odot \boldsymbol{i}_{t-k} \odot (1-\boldsymbol{i}_{t-k}) \otimes \boldsymbol{x}_{t-k} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{hi}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{i}_{t-k}}\frac{\partial \boldsymbol{i}_{t-k}}{\partial \boldsymbol{W}_{hi}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}} \odot \tilde{\boldsymbol{C}}_{t-k} \odot \boldsymbol{i}_{t-k} \odot (1-\boldsymbol{i}_{t-k}) \otimes \boldsymbol{h}_{t-k-1} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{b}_i} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{i}_{t-k}}\frac{\partial \boldsymbol{i}_{t-k}}{\partial \boldsymbol{b}_i} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}} \odot \tilde{\boldsymbol{C}}_{t-k} \odot \boldsymbol{i}_{t-k} \odot (1-\boldsymbol{i}_{t-k})
\end{aligned}
\tag{4.31}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{xc}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \tilde{\boldsymbol{C}}_{t-k}}\frac{\partial \tilde{\boldsymbol{C}}_{t-k}}{\partial \boldsymbol{W}_{xc}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{i}_{t-k}\odot(1-\tilde{\boldsymbol{C}}_{t-k}^2)\otimes \boldsymbol{x}_{t-k} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{hc}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \tilde{\boldsymbol{C}}_{t-k}}\frac{\partial \tilde{\boldsymbol{C}}_{t-k}}{\partial \boldsymbol{W}_{hc}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{i}_{t-k}\odot(1-\tilde{\boldsymbol{C}}_{t-k}^2)\otimes \boldsymbol{h}_{t-k-1} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{b}_{c}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \tilde{\boldsymbol{C}}_{t-k}}\frac{\partial \tilde{\boldsymbol{C}}_{t-k}}{\partial \boldsymbol{b}_{c}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{i}_{t-k}\odot(1-\tilde{\boldsymbol{C}}_{t-k}^2)
\end{aligned}
\tag{4.32}
$$

For the parameters in forget gate $\{\boldsymbol{W}_{hf}, \boldsymbol{W}_{xf}, \boldsymbol{b}_f\}$, the gradients backpropagate through $\boldsymbol{C}_t$ and $\boldsymbol{f}_t$. Overall gradients of forget gate parameters for a sequence with length T is expressed in Eq. (4.33).

$$
\begin{aligned}
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{xf}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{f}_{t-k}}\frac{\partial \boldsymbol{f}_{t-k}}{\partial \boldsymbol{W}_{xf}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{C}_{t-k-1}\odot \boldsymbol{f}_{t-k}\odot(1-\boldsymbol{f}_{t-k})\otimes \boldsymbol{x}_{t-k} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{W}_{hf}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{f}_{t-k}}\frac{\partial \boldsymbol{f}_{t-k}}{\partial \boldsymbol{W}_{hf}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{C}_{t-k-1}\odot \boldsymbol{f}_{t-k}\odot(1-\boldsymbol{f}_{t-k})\otimes \boldsymbol{h}_{t-k-1} \\
\frac{\partial \boldsymbol{j}}{\partial \boldsymbol{b}_{f}} &= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\frac{\partial \boldsymbol{C}_{t-k}}{\partial \boldsymbol{f}_{t-k}}\frac{\partial \boldsymbol{f}_{t-k}}{\partial \boldsymbol{b}_{f}} \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{t}\frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k}}\odot \boldsymbol{C}_{t-k-1}\odot \boldsymbol{f}_{t-k}\odot(1-\boldsymbol{f}_{t-k})
\end{aligned}
\tag{4.33}
$$

4.3.1.2.3 Backpropagating Gradients between Neighboring Time Steps

In the following, the discussion concentrates on the inner paths of LSTM cell, where the gradients backpropagate through time. As the gradients in terms of cell parameters depend on the partial derivatives of $\boldsymbol{J}_t$ with respect to state $\boldsymbol{C}_{t-k}$ or state $\boldsymbol{h}_{t-k}$, it is important to understand how the partial derivatives are computed. In unfolded graph, the gradients backpropagate cell by cell through two paths, cell state $\boldsymbol{C}_{t-k}$ and state $\boldsymbol{h}_{t-k}$ where $k = 1, \ldots t$, shown in Fig. 4.14. The paths where the gradients backpropagate through the inner architecture of the LSTM cell between two neighboring time steps, $t-k$ and $t-k+1$ are discussed respectively.

- **Partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{h}_t$ and $\boldsymbol{C}_t$**

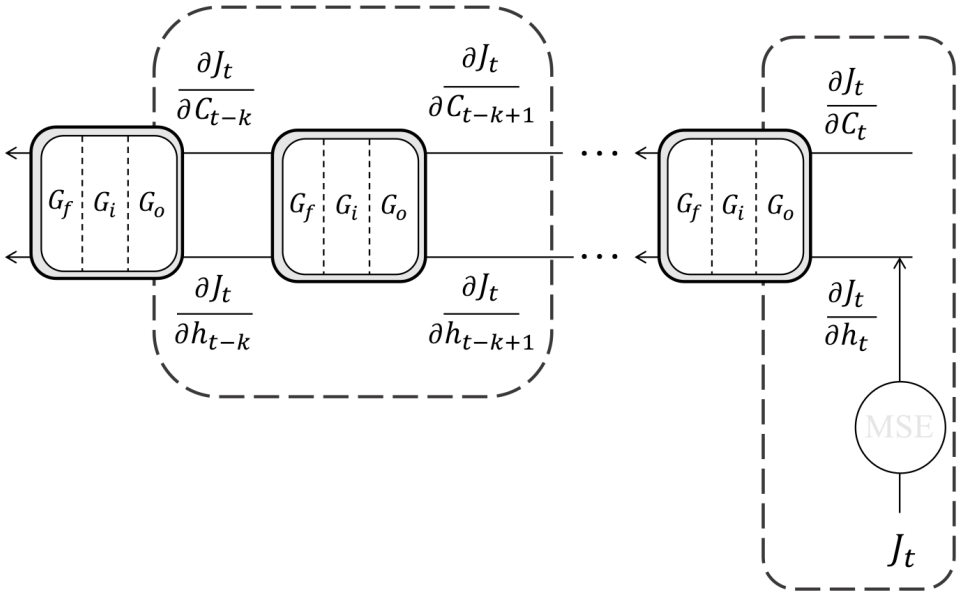


Figure 4.14: BPTT in terms of cell state and state over LSTM cells. An error ($\boldsymbol{J}_t$) at time t, backpropagate through the inner architecture of LSTM cell which has two different paths between the neighboring cells. The partial derivative of $\boldsymbol{J}_t$ with respect of $\boldsymbol{C}_{t-k}$ or $\boldsymbol{h}_{t-k}$ represents the effect of the error on the cell that k steps behind.

$$\begin{aligned}
\frac{\partial \boldsymbol{j}_t}{\partial \boldsymbol{h}_t} &= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{p}_t}\frac{\partial \boldsymbol{p}_t}{\partial \boldsymbol{h}_t} = (\boldsymbol{p}_t - \boldsymbol{y}_t)\cdot \boldsymbol{W}_p \\
\frac{\partial \boldsymbol{j}_t}{\partial \boldsymbol{C}_t} &= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_t}\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{C}_t} \\
&= 2\{(\boldsymbol{p}_t - \boldsymbol{y}_t \cdot \boldsymbol{W}_p\} \odot \boldsymbol{C}\boldsymbol{o}_t \odot \{(1 - tanh^2(\boldsymbol{C}_t))\}
\end{aligned} \tag{4.34}$$

As shown in the right box of Fig. 4.14, the partial derivatives of $\boldsymbol{J}_t$ with respect to $\boldsymbol{h}_t$ and $\boldsymbol{C}_t$ are simply computed as shown in Eq. (4.34) because the error $\boldsymbol{J}_t$ doesn't backpropagate through time yet. Note that the partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{C}_t$ is originated from the partial derivative with respect to $\boldsymbol{h}_t$.

- **Partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{h}_{t-k}$**

Fig. 4.15, depicts the four different paths within the LSTM cell through which the gradients backpropagate to $\boldsymbol{h}_{t-k}$. The paths (1) passing by the forget gate, and (2), (3) both passing by the input gate, start from $\boldsymbol{C}_{t-k+1}$ while the path (4) passing by output gate, starts from $\boldsymbol{h}_{t-k+1}$. Accordingly, the partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{h}_{t-k}$ can be expressed by the sum of the inflow from four paths, as shown in Eq. (4.35). Each path can be factorized

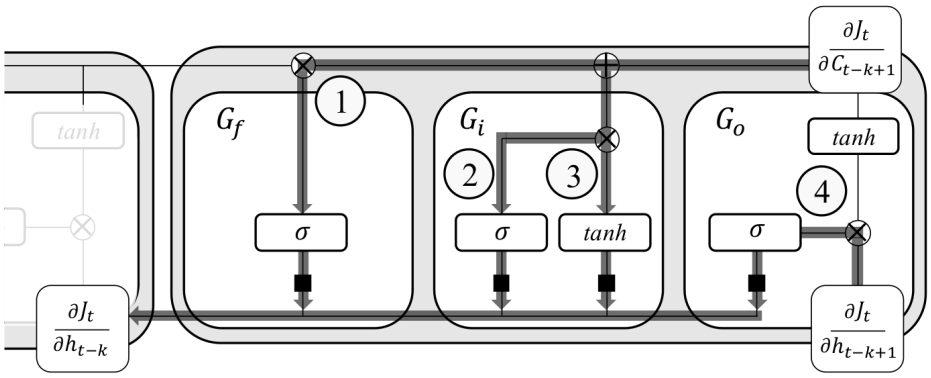


Figure 4.15: Four different paths that the gradients backpropagate to $\boldsymbol{h}_{t-k}$.

by the chain rule, as shown in Eq. (4.36).

$$\frac{\partial \boldsymbol{j}_t}{\partial \boldsymbol{h}_{t-k}} = \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{1} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{2} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{3} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{h}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{4} \tag{4.35}$$

$$\begin{aligned}
\frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{1} &= \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{f}_{t-k+1}} \frac{\partial \boldsymbol{f}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \\
&= \boldsymbol{C}_{t-k+1} \odot \{\boldsymbol{f}_{t-k+1} \odot (1 - \boldsymbol{f}_{t-k+1}) \cdot \boldsymbol{W}_{hf}\} \\
\frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{2} &= \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{i}_{t-k+1}} \frac{\partial \boldsymbol{i}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \\
&= \boldsymbol{C}_{t-k+1} \odot \{\boldsymbol{i}_{t-k+1} \odot (1 - \boldsymbol{i}_{t-k+1}) \cdot \boldsymbol{W}_{hi}\} \\
\frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{3} &= \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \tilde{\boldsymbol{C}}_{t-k+1}} \frac{\partial \tilde{\boldsymbol{C}}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \\
&= \boldsymbol{i}_{t-k+1} \odot \{(1 - \tilde{\boldsymbol{C}}^2_{t-k+1}) \cdot \boldsymbol{W}_{hc}\} \\
\frac{\partial \boldsymbol{h}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \textcircled{4} &= \frac{\partial \boldsymbol{h}_{t-k+1}}{\partial \boldsymbol{o}_{t-k+1}} \frac{\partial \boldsymbol{o}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \\
&= tanh(\boldsymbol{C}_{t-k+1}) \odot \{\boldsymbol{o}_{t-k+1} \odot (1 - \boldsymbol{o}_{t-k+1}) \cdot \boldsymbol{W}_{ho}\}
\end{aligned} \tag{4.36}$$

Fig. 4.15 and Eqs. (4.35 and 4.36) reveals that the partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{h}_{t-k}$ includes the non-linear activation function so that gradients backpropagating through the paths have risk of vanishing or exploding like RNN.

- **Partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{C}_{t-k}$**

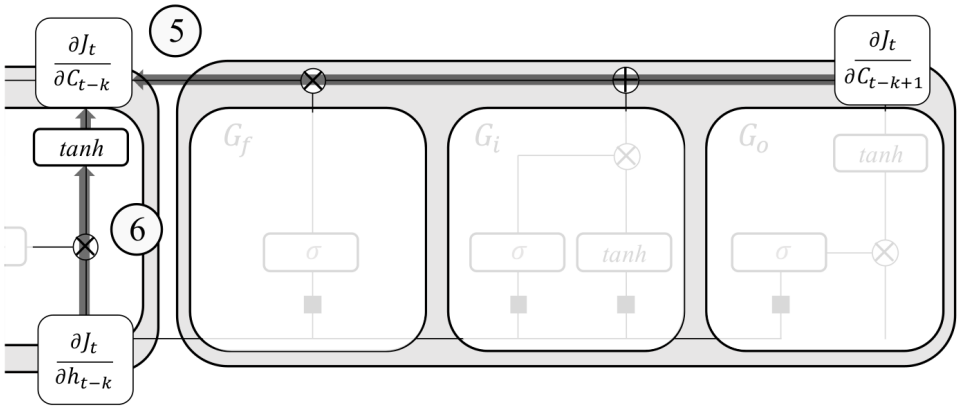


Figure 4.16: Two paths that the gradients backpropagate to $\boldsymbol{C}_{t-k}$. Gradients that backpropagate through the path (5) don't get vanishing or exploding thanks to the lack of activation function.

As shown in Fig. 4.16, the gradients backpropagate to the cell state $\boldsymbol{C}_{t-k}$ through two paths, (5) from $\boldsymbol{C}_{t-k+1}$ and (6) from $\boldsymbol{h}_{t-k}$. The partial derivative of $\boldsymbol{J}_t$ with respect to $\boldsymbol{C}_{t-k}$ is the sum of gradients from two paths (See Eq. (4.37)), where the path (6) stems from the partial derivative of $\boldsymbol{h}_{t-k}$ which is the sum of four path in the Fig. 4.15.

$$\begin{aligned}
\frac{\partial \boldsymbol{j}_t}{\partial \boldsymbol{C}_{t-k}} &= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{C}_{t-k}} \overset{(5)}{} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{h}_{t-k}} \frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{C}_{t-k}} \overset{(6)}{} \\
\frac{\partial \boldsymbol{j}_t}{\partial \boldsymbol{h}_{t-k}} &= \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \overset{(1)}{} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \overset{(2)}{} \\
&\quad + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \overset{(3)}{} + \frac{\partial \boldsymbol{J}_t}{\partial \boldsymbol{C}_{t-k+1}} \frac{\partial \boldsymbol{h}_{t-k+1}}{\partial \boldsymbol{h}_{t-k}} \overset{(4)}{}
\end{aligned} \tag{4.37}$$

The path (5) has a special property Eq. (4.38). As it doesn't have non-linear activation function on the path unlike the other paths, the gradients that backpropagate through the path (5) don't suffer from the vanishing or exploding issue. The path (5) provides a long-term bypass that enables the RNNs with LSTM cell to learn long-term dependencies in the sequence.

On the other hand, the path (6) Eq. (4.38) includes hyperbolic tangent which is a non-linear activation function. Partial derivative of $\boldsymbol{h}_{t-k}$ with respect to $\boldsymbol{C}_{t-k}$ returns $\boldsymbol{o}_{t-k} \odot \{1 - tanh^2(\boldsymbol{C}_{t-k})\}$, whose elements are equal to or less than one.

$$\begin{aligned}
\frac{\partial \boldsymbol{C}_{t-k+1}}{\partial \boldsymbol{C}_{t-k}} \overset{(5)}{} &= \boldsymbol{f}_{t-k} \\
\frac{\partial \boldsymbol{h}_{t-k}}{\partial \boldsymbol{C}_{t-k}} \overset{(6)}{} &= \boldsymbol{o}_{t-k} \odot \{1 - tanh^2(\boldsymbol{C}_{t-k})\}
\end{aligned} \tag{4.38}$$

4.3.2 Gated Recurrent Units (GRU)

A Gated Recurrent Unit (GRU) [27, 33] is a simpler variant of LSTMs, as shown in Fig. 4.18. Two notable differences compared with LSTM are found in Fig. 4.17. At first, GRU has two gates, referred to reset and update gate respectively. It implies that GRU has fewer parameters to train, as shown in Table 4.2, that allows to reduce computation load and time. The second, GRU doesn't have an independent memory cell, such as cell state $\boldsymbol{C}_t$ in LSTM. It suggests that the information is forward and backpropagates only through the state $\boldsymbol{h}_t$.

These two gates control the information flow of inputs, preventing the gradients from vanishing and capturing longer dependencies in the input sequence. Despite fewer gates and simpler architecture, it is known that GRU provides at least comparable prediction accuracy to LSTM and outperforms in terms of computation time [27, 174]. At each step $\boldsymbol{t}$, the model computes a candidate hidden state $\tilde{\boldsymbol{h}}_t$:

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_{hr} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{W}_{xr} \cdot \boldsymbol{x}_t + \boldsymbol{b}_r) \tag{4.39}$$

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W}_{xh} \cdot \boldsymbol{x}_t + \boldsymbol{W}_{hh} \cdot (\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h) \tag{4.40}$$

where $\boldsymbol{r}_t$ is a reset gate that controls the information flow of the previous state $\boldsymbol{h_{t-1}}$ to the candidate's hidden state $\tilde{\boldsymbol{h}}_t$. When $\boldsymbol{r}_t$ is close to $\boldsymbol{0}$, the previous hidden state $\boldsymbol{h_{t-1}}$ is ignored and the candidate hidden state $\tilde{\boldsymbol{h}}_t$ is reset with the current input, allowing the model to erase any irrelevant information from the previous step.

GRUs then define an update gate as a sigmoid function of the current input $\boldsymbol{x}_t$ and the previous hidden state $\boldsymbol{h}_{t-1}$:

$$\boldsymbol{u}_t = \boldsymbol{\sigma}(\boldsymbol{W}_{xu} \cdot \boldsymbol{x}_t + \boldsymbol{W}_{hu} \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}_u) \tag{4.41}$$

In LSTM, two variables $\boldsymbol{f}_t$ and $\boldsymbol{i}_t$ respectively determines the portion of the information, to inherit or to update. In GRU, $\boldsymbol{u}_t$ determines the portion in two ways. $(1 - \boldsymbol{u}_t)$ takes charge of the role of $\boldsymbol{f}_t$ while $\boldsymbol{u}_t$ takes $\boldsymbol{i}_t$'s role. Thus, update gate $\boldsymbol{u}_t$ is interpreted as a coupled gate of forget $\boldsymbol{f}_t$ and input gate $\boldsymbol{i}_t$ of LSTM cell. Therefore, it controls how much information from the previous step is brought to the current step t.

The hidden state $\boldsymbol{h}_t$ at current step t is computed as :

$$\boldsymbol{h_t} = (1 - \boldsymbol{u}_t) \odot \boldsymbol{h_{t-1}} + \boldsymbol{u}_t \odot \tilde{\boldsymbol{h}}_t \tag{4.42}$$

Here, GRUs do not have an explicit input gate but $\boldsymbol{1 - u_t}$ is used as the gate to control the new information updated in the current step. Hence $\boldsymbol{h}_t$ is a linear interpolation of the candidate hidden state $\tilde{\boldsymbol{h}}_t$ and the previous hidden state $\boldsymbol{h_{t-1}}$. This prevents the amount of information in the hidden states from exploding.

4.4 Closing remarks

We've explored Recurrent Neural Networks (RNNs), a versatile model family applicable across various machine learning tasks. Utilizing features such as gating, attention mechanisms, and memory, RNNs have excelled in numerous applications, particularly with large-scale and complex data.

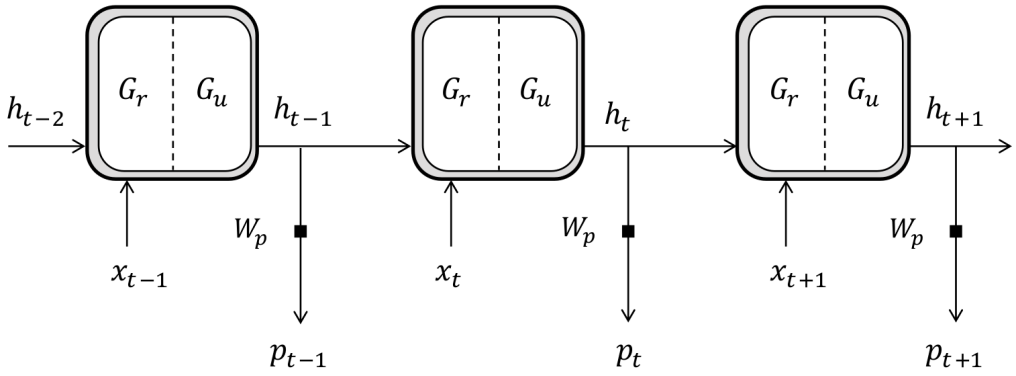


Figure 4.17: Unfolded graph of a RNN with GRU cell that consist of two gates.

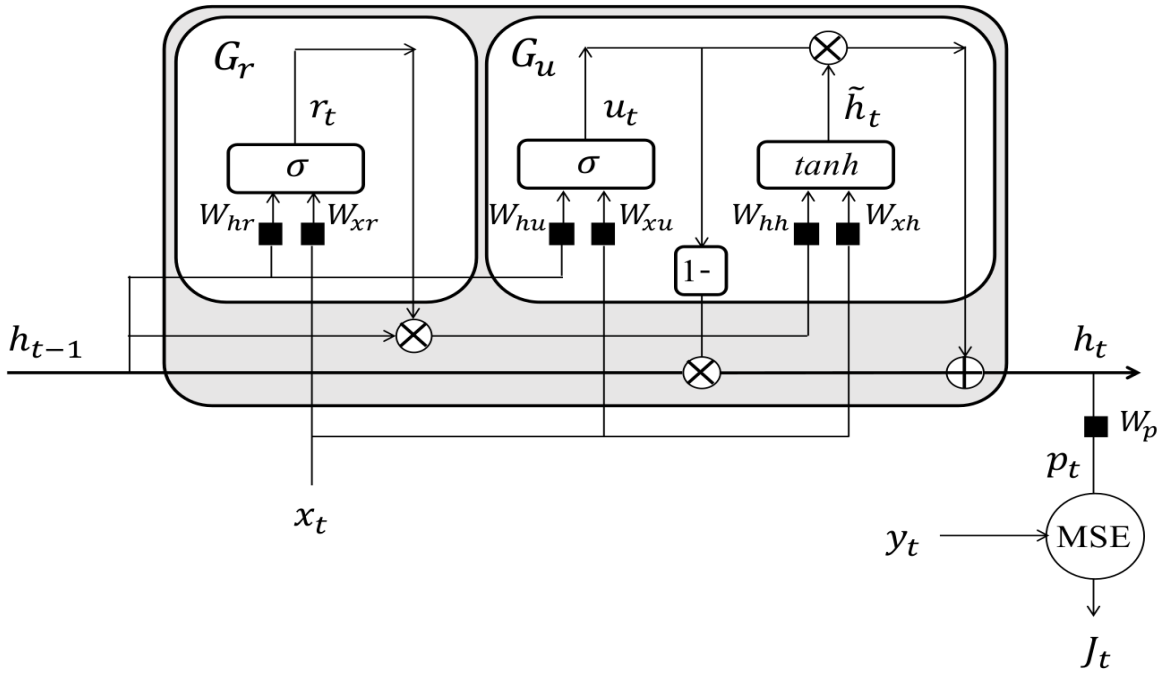


Figure 4.18: Gated recurrent unit (GRU) neural network unit structure. $\boldsymbol{x}_t$ is the current input, $\boldsymbol{u}_t$ is the update gate Eq. (4.41), $\boldsymbol{r}_t$ is the reset gate Eq. (4.39), $\tilde{\boldsymbol{h}}_t$ is the candidate hidden state of the currently hidden input Eq. (4.42) , $\boldsymbol{h}_t$ is the current hidden state, $\boldsymbol{x}_t$ is the input of the current neural network, and $\boldsymbol{h}_{t-1}$ is the hidden state at the previous moment. $\boldsymbol{\sigma}$ is the activation function sigmoid. $\tilde{\boldsymbol{h}}_t$ records all important information through the reset gate and input information, source from [92].

Table 4.2: Variables and trainable parameters of GRU cell.

Gate type	Variable	Parameters		
Reset gate $\boldsymbol{G}_r$	$\boldsymbol{r}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xr} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hr} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_r \in \mathbb{R}^s$
Update gate $\boldsymbol{G}_u$	$\boldsymbol{u}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xu} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hu} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_u \in \mathbb{R}^s$
	$\tilde{\boldsymbol{h}}_t \in \mathbb{R}^s$	$\boldsymbol{W}_{xh} \in \mathbb{R}^{s\times m}$	$\boldsymbol{W}_{hh} \in \mathbb{R}^{s\times s}$	$\boldsymbol{b}_h \in \mathbb{R}^s$
Prediction	$\boldsymbol{p}_t \in \mathbb{R}^m$	$\boldsymbol{W}_p \in \mathbb{R}^{m\times s}$		$\boldsymbol{b}_p \in \mathbb{R}^m$

RNNs are not limited to sequential data, they are models with self-connections that iteratively refine data representation across multiple computation steps. This characteristic allows RNNs to serve as effective alternatives to Feed forward Neural Networks (FNNs) for vector-to-vector mapping. While deep FNNs can handle complex functions, they are prone to overfitting with limited data. An alternative approach involves parameter sharing across hidden layers, converting an FNN into an RNN. This method enables learning of intricate functions by adding hidden layers while maintaining a constant number of parameters, thus mitigating overfitting.

Research confirms RNNs' effectiveness in handling vector data [123] and image data [107]. Beyond these applications, our hypothesis suggests that RNNs can tackle challenges posed by complex structured data types such as episodic data (EHRs), relational data, multiple input/output data, and graph data, topics explored further in Chapters 5.

Part III

Contribution

Chapter 5

Multi-Way adaptive Time Aware LSTM (MWTA-LSTM)

"If the doors of perception were cleansed everything would appear to man as it is, infinite."

– William Blake:The Doors of Perception

In personalized predictive medicine, accurately modeling a patient's illness and care processes is essential, given their inherent long-term temporal dependencies. However, electronic medical records contain episodic and irregularly timed data due to patients visiting hospitals based on treatment needs, resulting in unique patterns for each hospital stay. Consequently, when constructing a personalized predictive model, it is crucial to consider these factors in order to accurately capture the patient's health journey, as shown in Fig. 3.6.

To address this challenge, we present a novel deep dynamic memory neural network

called Multi-Way adaptive Time Aware LSTM (MWTA-LSTM). The primary objective of MWTA-LSTM is to leverage medical records, memorize illness trajectories and care processes, estimate current illness states, and predict future risks, thereby providing a high level of precision and predictive power. To enhance its capabilities, MWTA-LSTM extends the conventional Long Short-Term Memory (LSTM) model in three key ways. Firstly, it incorporates frequency measurement and the most recent observation to enhance personalized predictive modeling of patient illnesses. This inclusion allows for a more accurate understanding of the patient's condition. Secondly, MWTA-LSTM parameterizes time to effectively handle irregular timing, enabling the modeling of interventions and their impact on the course of illness and disease progression through the use of two decay mechanisms. Lastly, we introduce a novel adaptive pooling strategy that specifically targets and resolves outlier issues that can potentially occur during the analysis of EHR data. By incorporating these features, MWTA-LSTM significantly improves its ability to capture the temporal dynamics of healthcare data, accommodating variations and irregularities in event and observation timing.

Our contributions can be summarized as follows:

- We introduce a novel framework called MWTA-LSTM, which effectively handles timing irregularities, and is capable to capture the complex interactions between different clinical features at different stages. MWTA-LSTM extends the standard LSTM [63] $gates$ ($forget$, $input$, and $output$) by an additional $time\ gate$ to adjust the memory cell in a way that diminishes the influence of previous memory as the elapsed time increases. We achieve this adjustment through the use of two time gate mechanisms to reflect the intervention effects on the current output, one of which was proposed by Baytas et al [5], while the other is our proposed approach.
- Our model takes a unique approach by treating each feature differently, learning their decaying parameters based on frequency measurements and time intervals between events. This methodology empowers the model to effectively address timing irregularities and overcome the challenge of sparse records commonly found in EHR data.
- Additionally, we propose a novel adaptive pooling strategy to address outlier issues that may arise during EHR data analysis, enhancing the robustness and reliability of our framework.

Hence, our model not only captures long-term temporal dependencies in time series observations, but it also effectively addresses timing irregularities by leveraging frequency measurements, resulting in significantly enhanced prediction performance. Through empirical experiments conducted on two real-world clinical datasets and three time series datasets, we have demonstrated the superiority of our proposed model over the current state-of-the-art models, as well as other robust baselines, which will be discussed in greater detail in Section 7.

To the best of our knowledge, our research represents the first attempt to jointly incorporate the time interval between events as intervention data and integrate it into the LSTM unit for adjusting the cell state. Additionally, we leverage both the time interval and frequency measurement as decay mechanisms to handle timing irregularities and address

missing data issues in EHR data analysis effectively. Overall, the MWTA-LSTM framework, combined with the customized treatment of features and the adaptive pooling strategy, presents a comprehensive and advanced solution for handling timing irregularities, capturing complex feature interactions, and addressing outliers in EHR data analysis.

To address the above challenges encountered in EHR data analysis, as discussed in section 3.8, we introduce MWTA-LSTM, which stands for Multi-Way adaptive Time Aware Long Short-Term Memory an end-to-end deep neural network. MWTA-LSTM leverages the power of Long Short-Term Memory (LSTM) and takes into account three essential pieces of information: the elapsed time between events in a sequence, the frequency measurement, and the last observed value of the feature. By incorporating these three factors into the MWTA-LSTM unit, we effectively navigate the processing of irregularly spaced multivariate time series and adeptly manage timing irregularities while minimizing additional noise in sparse records encountered during EHR data analysis.

MWTA-LSTM is an advanced end-to-end deep-learning model designed to capture the complex temporal dynamics of sequential data with time irregularities. It has four main layers. The bottom layer is built upon an extended modified version of T-LSTM and Time-LSTM which were first proposed in [5] and [176], whose memory cells are modified to handle the temporal irregularity of a patient's medical records. The subsequent layers include a multi-head attention pooling layer, a gating mechanism layer, a temporal convolutional block, and an output layer, as depicted in Fig. 5.1. In the subsequent discussion, we delve into the specifics of each layer.

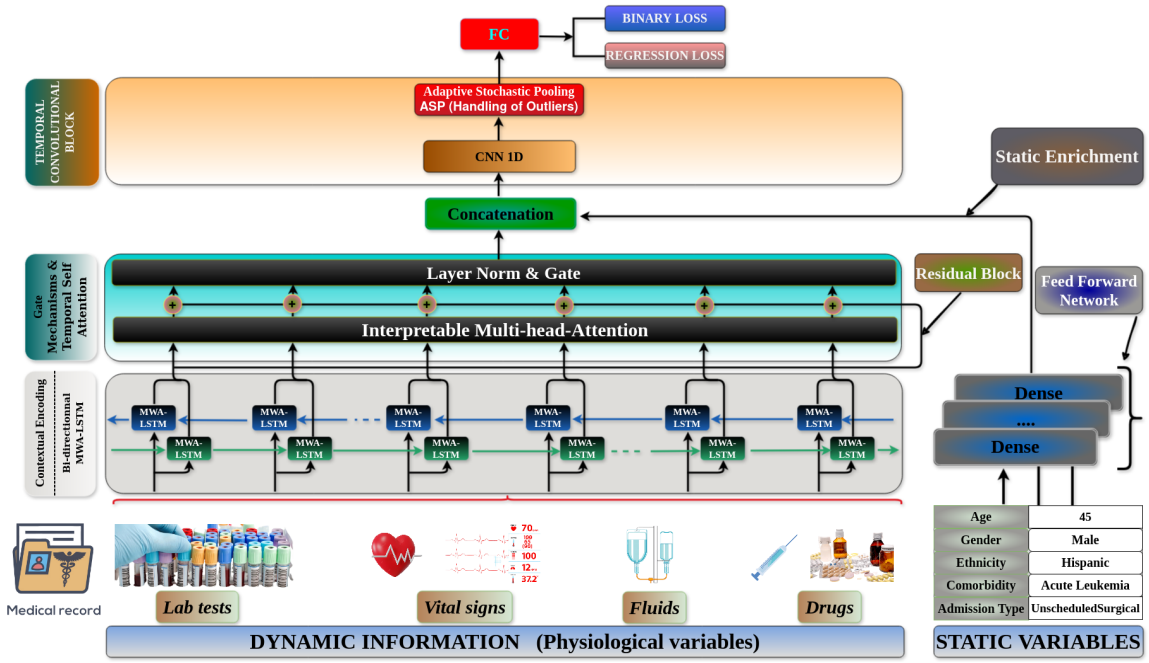


Figure 5.1: Multi-Way adaptive Time-Aware LSTM (MWTA-LSTM).

5.1 MWTA's Unit 1

The time intervals between successive events during a patient's hospitalization can exhibit considerable variation, ranging from mere minutes to several weeks. As the significance of past events diminishes with time, it becomes crucial to mitigate their influence on current assessments. Therefore, considering elapsed time is pivotal for ensuring accurate pre-

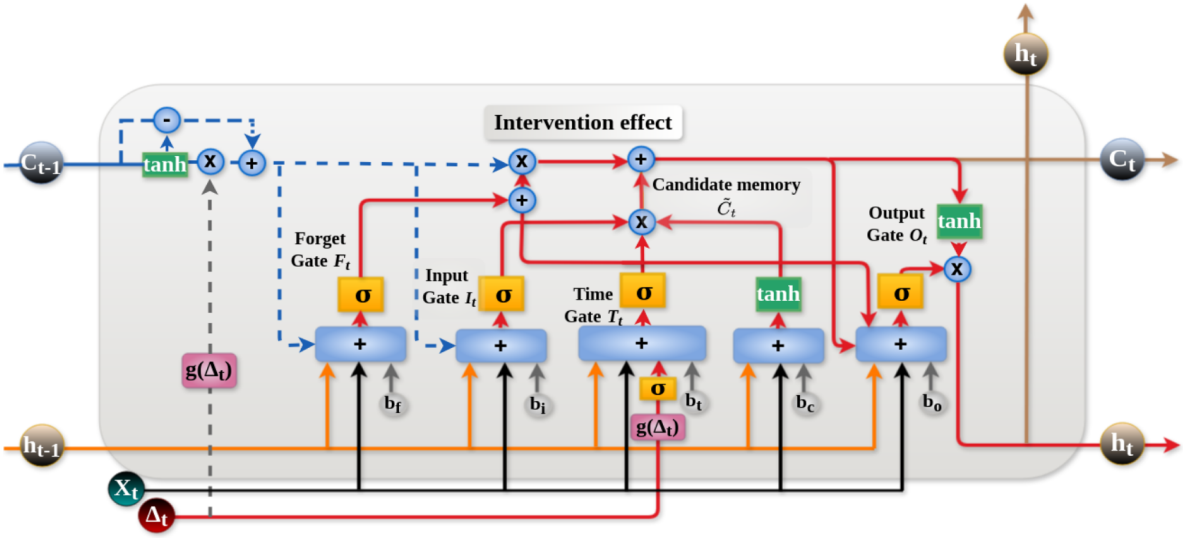


Figure 5.2: MWTA-LSTM 1 (unit) on analyzing healthcare records. Red arrows correspond to the modifications compared to standard LSTM. Shade blue boxes indicate networks and shade blue circles denote point-wise operators.

dictions of present outcomes. To enhance prediction performance, it is essential to adopt an architecture capable of handling this irregularity, particularly in scenarios where the frequency and volume of patient records can vary greatly. Rather than viewing varying elapsed times as obstacles, they should be regarded as integral components of a patient's medical history and leveraged during the processing of EHRs data. In order, to address the temporal irregularity in a patient's medical records, the contextual encoding layer of MWTA-LSTM employs a bidirectional layer to capture the patient's historical medical state.

In this version we use two data related to the patient, more specifically, the input sequence is represented by the temporal patient data denoted by $\boldsymbol{X}$ and the elapsed time $\boldsymbol{\Delta}_t$ between successive records for each clinical feature related to the patient stay which is shown in Fig. 5.2.

As first proposed in [5], we decomposed the previous memory into two distinct components, namely the long-term and short-term memories using equations from Eq. (3.17 to 3.19). Subsequently, the short-term memory undergoes a reduction in value through the application of a time decay factor calculated based on the time interval between consecutive records ($\boldsymbol{\Delta}_t$). Ultimately, the discounted short-term memory (D-STM) and long-term memory (LTM) are merged to yield an updated memory. This adjustment allows for customization of the memory content in the new cell state unit, reducing the impact of past memory on the current output, especially when there is a significant time gap between successive recordings of a clinical feature. This is crucial since $\boldsymbol{x}_t$ contains information on the patient's present diagnosis, reflecting his/her short and long-term medical condition, while $\boldsymbol{c}_{t-1}$ tracks his/her long-term medical history based on past medical records.

To better account for the varying time intervals between successive medical records,

an additional time gate has been included to regulate the state update process. The main premise behind this modification is that a patient's medical condition is constantly evolving and that each prior medical encounter affects his/her current state $\boldsymbol{c}_t$ in a different way. Therefore, it's essential to capture the impact of varying time intervals between records. The time gate is leveraged to regulate the influence of these two factors, taking into account the time gap $\boldsymbol{\Delta}_t$.

$$\boldsymbol{T}_t = \sigma(\boldsymbol{W}_{x_t}\boldsymbol{x}_t + \boldsymbol{\sigma}_{\boldsymbol{\Delta}_t}(g(\boldsymbol{\Delta}_t)) + \boldsymbol{b}_t) \tag{5.1}$$

where σ is the sigmoid function, $\boldsymbol{\Delta}_t$ is the time interval between records for each clinical feature at time span $t-1$, and t, this time gap can vary from hours to days even years.

5.1.1 Modeling effect of interventions

As discussed in section 3.1.1, assessing a patient's health status involves recording clinical feature measurements over time, and the elapsed time vector plays a pivotal role in this process. Considered an intervention, this vector monitors the frequency of measurements for each clinical feature, directly reflecting the patient's health condition. Shorter intervals between measurements for vital clinical features may signal a rapidly deteriorating condition, while longer intervals may indicate stability. Hence, the elapsed time vector serves as a crucial tool in evaluating a patient's health status over time, aligning with the goal of interventions to either cure diseases or alleviate the patient's illness.

This intervention effect is modeled by regulating the cell state $\boldsymbol{c_t}$ using Eqs. (6.6 and 5.1), as illustrated in Fig. 5.2.

$$\boldsymbol{c_t} = (f_t + T_t) \odot C^*_{t-1} + i_t \odot T_t \odot \tilde{c}_t \tag{5.2}$$

In this way, the input gate i_t is not only responsible for filtering input information but also the time gate T_t, which plays a vital role in regulating the influence of x_t on the patient's current state and modeling their long-term medical condition.

In healthcare, the significance of certain events or interventions may vary based on their temporal context. Consequently, the output gate, which governs illness states, is also influenced by the current intervention, as outlined below:

$$\boldsymbol{o_t} = \sigma(W_o[c_t, h_{t-1}, x_t] + T_t + b_o) \tag{5.3}$$

Through the inclusion of $\boldsymbol{c_t}$ and the time gate $\boldsymbol{T}_t$ in the output gate, the MWTA-LSTM is capable of assigning distinct weights to features, considering their correlation with specific events. This capability allows the model to distinguish between events occurring in close proximity or further apart, capturing the nuanced impact of these events on the patient's health state. Furthermore, it facilitates the learning of patient-specific patterns, such as variations in response to treatments or the progression of diseases, resulting in more precise and personalized predictions.

5.2 MWTA's Unit 2

Two more pieces of information are added in this version, which are the last observation $\boldsymbol{X}_{last}$ and the frequency measurement of each physiological variable $\boldsymbol{f}_X$, as shown in Fig. 5.3.

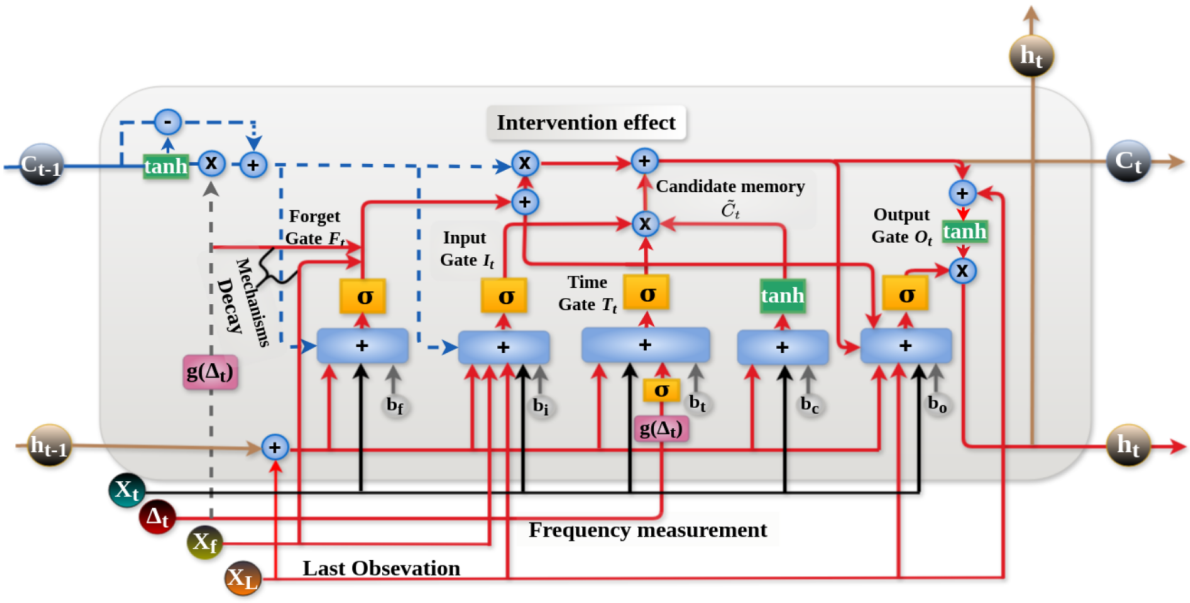


Figure 5.3: MWTA-LSTM 2 (unit) on analyzing healthcare records. Goldenrod arrows correspond to the modifications compared to MWTA-LSTM 1 (unit). Shade blue boxes indicate networks and shade blue circles denote point-wise operators.

Using of $\boldsymbol{X}_{last}$ and $\boldsymbol{f}_X$ has several benefits when building personalized predictive model:

Firstly, by using the last observation of each clinical variable, we can better represent the current health status of the patient. This is because the most recent value is likely to be more reflective of the patient's current health status compared to earlier measurements. It's mathematically expressed as follows:

$$\boldsymbol{L}_{\boldsymbol{x}_t} = \mathbf{ELU}(\boldsymbol{X}_{last} W_{X_{last}} + \boldsymbol{b}_{\boldsymbol{X}_{last}}) \tag{5.4}$$

Where $\boldsymbol{ELU}$ is the activation function [28] and expressed as:

$$\boldsymbol{ELU} = \begin{cases} x & \text{if } x \geq 0, \\ \alpha(e^x - 1) & \text{otherwise.} \end{cases}$$

with $\boldsymbol{\alpha}$ being the hyperparameter that controls the saturation of the function for negative inputs.

The choice to use the ELU activation function in analyzing EHRs data is driven by its advantages: ELU effectively captures complex non-linear relationships in EHR data, improving pattern recognition. It is more robust to outliers than ReLU, handling the irregularities

common in EHRs data and avoiding large gradients for large negative inputs. Additionally, ELU acts as a regularizer, preventing overfitting by limiting output magnitude for large negative inputs, leading to a more generalized model.

Secondly, examining the frequency measurement of each clinical variable offers a more nuanced understanding of their temporal changes. This provides a comprehensive view of the patient's health status, facilitating the identification of significant trends that may serve as crucial predictors, as stated in section 3.8.

In conclusion, frequency measurement provides a comprehensive view of the patient's health status, encompassing various physiological and biochemical parameters. This comprehensive data is essential for creating a detailed health profile that can guide holistic patient management. It also allows for a better understanding of the interplay between different aspects of the patient's health, such as the relationship between clinical parameters. Furthermore, incorporating both the last observation and the frequency measurement of each clinical variable is beneficial for addressing issues related to missing data, especially when certain clinical variables are not measured at every encounter. This approach ensures a more complete and accurate picture of the patient's health.

5.2.1 Mitigating missing data issues

Within an LSTM unit, the hidden state $\boldsymbol{h_t}$ functions as the model's memory, retaining information from prior time steps. Our proposal involves modifying the previous hidden state $\boldsymbol{h_{t-1}}$ and the cell state $\boldsymbol{c_t}$ by incorporating the last observation $\boldsymbol{L_{x_t}}$ Eq. (5.4) of each feature. This implies that, at each time step, the hidden state is adjusted to encompass the most recent recorded data for each feature.

$$\begin{aligned} \boldsymbol{h}^{*}_{t-1} &= \overbrace{\boldsymbol{h}_{t-1} + \boldsymbol{L}_{\boldsymbol{x}_t}}^{\text{Adjusted previous memory}} \\ \boldsymbol{h}_t &= \overbrace{\boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t + \boldsymbol{L}_{x_t})}^{\text{Adjusted current memory}} \end{aligned} \tag{5.5}$$

The purpose of this adjustment is to enable the model to incorporate the latest information into its memory. Adjusting $\boldsymbol{h}_{t-1}$ and $\boldsymbol{c}_t$, the MWTA-LSTM unit can enhance the model's ability to capture the dynamic nature of a patient's health over time. This can result in more precise predictions about future outcomes as the most pertinent and up-to-date information is integrated into the hidden state representation.

We have also made adjustments to the output gate Eq. (5.3) to incorporate $\boldsymbol{L_{x_t}}$ Eq. (5.4), expressed as follows:

$$\boldsymbol{o_t} = \sigma\big(W_o[c_t, h^{*}_{t-1}, x_t] + T_t + L_{x_t} + b_o\big) \tag{5.6}$$

By doing this the output gate acts as a filter, which gives the ability to the model to selectively filter out irrelevant information and focus on the most important features. This is because the last observation of a feature can provide important context about the patient's current state and can help the MWTA-LSTM unit make more informed decisions about which information to retain and which to discard because some features may be more predictive at certain times than others:

5.2.2 Capturing time irregularity

When utilizing LSTM to model a patient's medical history, the memory cell holds valuable information regarding his/her illness trajectory. However, it's important to note that this memory should not remain constant, as the patient's health status is subject to change over time. To address this issue, we introduce two forget mechanisms (Time-based decay and Frequency-based decay) within the MWTA-LSTM, which adjust the forget gate Eq. (4.22) and the input gate Eq. (4.23) to help the memory cell adapt and update accordingly. These two terms are defined as follows:

$$\boldsymbol{t}_d = \overbrace{g(\boldsymbol{\Delta}_t)}^{\text{Time-based}} \tag{5.7}$$

$$\boldsymbol{f}_d = \overbrace{e^{-\beta * \boldsymbol{f}\boldsymbol{x}_d} W_{f_d} + \boldsymbol{b}_{f_d}}^{\text{Frequency-based}} \tag{5.8}$$

where $\beta \in [0, 1]$ is the hyper-parameter (constant) that will determine the rate at which the transformed values decrease as the feature's frequency measurement increases. W_{f_d}, W_{t_d}, $\boldsymbol{b}_{f_d}$ and $\boldsymbol{b}_{t_d}$ are the weights and biases for the two terms.

The modified input gate and updated forget gate value are expressed as follows:

$$\boldsymbol{i}_t = \sigma(W_i[c_{t-1}, h^*_{t-1}, x_t] + L_{x_t} + f_d + b_i) \tag{5.9}$$

The rationale for this modification stems from the nature of EHR data, which encompasses a patient's complete medical history with irregularly spaced measurements. The frequency measurement serves as a valuable metric, revealing insights into the consistency and continuity of the data by indicating how often measurements were taken. By incorporating the frequency measurement and the last observation into the input gate Eq. (5.9) of the MWTA-LSTM unit, it acts as a filter, enabling the model to effectively sift through unnecessary information and concentrate on the essential features, thereby improving the model's ability to grasp the temporal context of each variable in the patient's medical history.

This integration allows for the effective utilization of longitudinal information, leading to a deeper understanding of the patient's health progression over time. Additionally, it facilitates the dynamic adjustment of variable weights within the MWTA-LSTM unit so that variables that occur more frequently can be assigned higher weights, reflecting their relative importance or relevance in the sequence, while still paying attention to variables with low frequencies to ensure their significance is not disregarded during the learning process.

$$\boldsymbol{f_t} = \overbrace{f_t * t_d}^{\text{Time-based decay}} + \overbrace{(1 - f_t) * f_d}^{\text{Frequency-based decay}} \tag{5.10}$$

where $\boldsymbol{f}_t$ is the updated forget gate, $\boldsymbol{f}_t$ the forget gate Eq. (4.22).

Incorporating these two decay terms when adjusting the forget gate results in several benefits, which are outlined below.

- **Time-based decay:** The first term in the equation (Eq. 5.10) introduces time-based decay, reducing the forget gate's activation value as more time elapses since the last

update. This is particularly relevant for Electronic Health Record (EHR) data, where the significance of past medical history diminishes over time. The MWTA's Unit, through time-based decay, can dynamically regulate its forgetfulness based on the elapsed time since the previous time step. This ensures the model does not retain outdated information, which is crucial in applications like EHR data analysis, language modeling, or speech recognition where recent information holds greater importance.

- **Frequency-based decay:** EHR data often involves varying frequencies of medical conditions or events. The second term in the equation (Eq. 5.10) introduces a frequency based decay, adjusting the forget gate's activation value based on the frequency of input data. This mechanism allows the forget gate to prioritize frequently occurring information, preventing it from being forgotten too quickly. The MWTA's Unit can adapt its memory retention to capture essential patterns in the data, even if they are less frequent, while efficiently discarding irrelevant or redundant information for frequently occurring events.

- **Adaptive behavior:** Combining time and frequency based decays empowers the MWTA's Unit to adapt the forget gate behavior dynamically. It can selectively forget or retain information based on its relevance and importance, leading to efficient memory utilization and enhanced performance across various sequence modeling tasks.

Algorithm A.1 describes the whole training of **MWTA LSTM**, as shown in Fig. 5.3

5.3 Multi-head attention

Time series data often have dependencies that span long intervals. Multi-head attention can effectively capture these long-range dependencies by allowing each head to focus on different parts of the sequence. In order, to enhance the learning capability, we suggest incorporating the multi-head attention mechanism proposed in [95] into MWTA-LSTM. This approach aims to capture meaningful features specific to a given dataset and facilitate the understanding of long-term dependencies across different time steps.

In the context of attention mechanisms, the values $\boldsymbol{V} \in \mathbb{R}^{N \times d_v}$ are typically scaled based on the relationship between the keys $\boldsymbol{K} \in \mathbb{R}^{N \times d_{attn}}$ and queries $\boldsymbol{Q} \in \mathbb{R}^{N \times d_{attn}}$. The attention function is defined as:

$$Attention(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = A(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V} \tag{5.11}$$

where A() is a normalization function, and N is the number of time steps feeding into the attention layer, represented as k + t_{max}. Scaled dot-product attention is a commonly used normalization function, as proposed by [157]:

$$A(\boldsymbol{Q}, \boldsymbol{K}) = Softmax(\boldsymbol{Q}, \boldsymbol{K}^t / \sqrt{d_{attn}}) \tag{5.12}$$

Algorithm 1 MWTA-LSTM forward propagation. The process is repeated by n_{epochs} epochs

Inputs:
$[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M]$: the input vectors
$[\boldsymbol{\Delta}_{t_1}, \ldots, \boldsymbol{\Delta}_{t_M}]$: Elapsed time
$[\boldsymbol{f}_{x_1}, \ldots, \boldsymbol{f}_{x_M}]$: Frequency measurement
$[\boldsymbol{x}_{last_1}, \ldots, \boldsymbol{x}_{last_M}]$: Last observation
$[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_M]$: the corresponding outputs
for e = 1 to n_{epochs} **do**
  for *each step* $t = 1, \ldots, T$ **do**
    $*$ Compute the adjusted previous memory cell $\boldsymbol{c}^*_{t-1}$ using (Eq. 3.19)
    $*$ Compute the adjusted $\boldsymbol{h}^*_{t-1}$ (Eq. 5.5)
    $*$ Compute 4 gates: $\boldsymbol{T}_t$ (Eq. 5.1), $\boldsymbol{i}_t$ (Eq. 5.9), $\boldsymbol{f}_t$ (Eq. 5.10), $\boldsymbol{o}_t$ (Eq. 5.6)
    $*$ Compute $\boldsymbol{h}_t$ (Eq. 5.5) and $\boldsymbol{c}_t$ (Eq. 5.2)
  end for
  if the task is a classification **then**
    $*$ Compute the predictive probability using (Eq. 7.1)
    $*$ Compute the log loss.
  else
    $*$ Compute the mse probability using (Eq. 7.2)
    $*$ Compute the difference between the predicted value and the ground truth.
  end if
end for

To improve the learning capacity of the standard attention mechanism, multi-head attention was introduced by [157]. This involves employing different heads for different representation subspaces:

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = [\boldsymbol{H_1}, ..., \boldsymbol{H}_{mH}]\boldsymbol{W_H} \tag{5.13}$$

Here, $\boldsymbol{H_h}$ denotes the output of the attention function for the h^{th} head, and is computed as follows:

$$\boldsymbol{H_h} = Attention(\boldsymbol{Q}\boldsymbol{W}_Q^{(h)}, \boldsymbol{K}\boldsymbol{W}_K^{(h)}, \boldsymbol{V}\boldsymbol{W}_V^{(h)}) \tag{5.14}$$

The weights $\boldsymbol{W}_K^{(h)}$, $\boldsymbol{W}_Q^{(h)}$, and $\boldsymbol{W}_V^{(h)}$ are head-specific and belong to the set of $\mathbb{R}^{d_{model} \times d_{attn}}$ and $\mathbb{R}^{d_{model} \times d_v}$, respectively. The outputs from all heads are concatenated and linearly combined using $\boldsymbol{W}_H$, which belongs to $\mathbb{R}^{(mh.d_v) \times d_{model}}$.

However, the use of varying values in each head makes it challenging to ascertain the importance of a specific feature based solely on attention weights in standard multi-head attention. To tackle this challenge, a modification to multi-head attention has been proposed by [95]. This modification entails sharing values across all heads and employing additive aggregation of all heads. This is expressed in equations (5.15) through (5.17).

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \tilde{\boldsymbol{H}}\boldsymbol{W_H}, \tag{5.15}$$

$$\tilde{\boldsymbol{H}} = \tilde{\boldsymbol{A}}(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V}\boldsymbol{W_V} \tag{5.16}$$

$$= \left\{ \frac{1}{\boldsymbol{m_H}} \sum_{h=1}^{\boldsymbol{m_H}} \boldsymbol{A}\left(\boldsymbol{Q}\boldsymbol{W}_{\boldsymbol{Q}}^{\boldsymbol{h}}, \boldsymbol{K}\boldsymbol{W}_{\boldsymbol{K}}^{\boldsymbol{h}}\right) \right\} \boldsymbol{V}\boldsymbol{W_V}, \tag{5.17}$$

$$= \frac{1}{\boldsymbol{m_H}} \sum_{h=1}^{\boldsymbol{m_H}} Attention(\boldsymbol{Q}\boldsymbol{W}_Q^{(h)}, \boldsymbol{K}\boldsymbol{W}_K^{(h)}, \boldsymbol{V}\boldsymbol{W}_V^{(h)}), \tag{5.18}$$

where $\boldsymbol{W_V} \in \mathbb{R}^{d_{model} \times d_v}$ are value weights shared across all heads, and $\boldsymbol{W_H} \in \mathbb{R}^{d_{attm} \times d_{model}}$ is used for final linear mapping.

5.4 Gating mechanisms

The relationship between inputs and targets can be ambiguous, complicating the identification of relevant variables, non-linear processing, and handling small or noisy datasets. To tackle these challenges, we propose using component gating layers based on Gated Linear Units (GLUs) [31], which offer model flexibility by suppressing irrelevant features and enhancing efficiency in predicting patient outcomes. Applying GLUs to EHRs data provides several benefits. GLUs refine clinical feature representation by selectively gating inputs, identifying crucial features, and ignoring noise. They mitigate the impact of missing data by activating available information, ensuring robustness. GLUs excel at modeling complex

relationships between features and patient outcomes, capturing intricate non-linear interactions. Additionally, GLUs improve gradient flow, leading to faster training and better performance.

Mathematically, the GLU activation function is expressed as follows:

$$GLU_{\boldsymbol{w}}(\boldsymbol{h}) = \sigma(\boldsymbol{W_{1,w}h} + \boldsymbol{b_{1,w}}) \odot (\boldsymbol{W_{2,w}h} + \boldsymbol{b_{2,w}}) \tag{5.19}$$

Here, $\boldsymbol{h}$ corresponds to the hidden state, σ is the sigmoid activation function, $\odot$ represents the element-wise Hadamard product, and $W_{1,w}$, $W_{2,w}$, $b_{1,w}$, and $b_{2,w}$ are weights and biases $\in \mathbb{R}^D$.

5.5 Adaptive stochastic pooling for handling outliers

To enhance the model's capability, we integrated a CNN layer [79] into the architecture. This addition enables the handling of multiple input feature maps and generates stacked feature maps with increasing layers, determined by the encoded input. CNN has demonstrated impressive performance in various tasks such as image, voice, video, and object detection [79, 145]. Their effectiveness lies in capturing time-invariant features within short temporal regions, a task traditional fully connected neural networks may find challenging. When applying 1D convolution and pooling operations to the temporal dimension of sensor signals, we obtain a feature map denoted as $z_{i,j}^{l,k}$ in the lth convolutional layer(the index i traverses the spatial domain, while j spans the temporal domain). The computation of this feature map can be expressed as follows:

$$z_{i,j}^{l,k} = f\left(\sum_{k'=1}^{K'} \sum_{y=1}^{Y} x_{i,j+y-1}^{l-1,k'} \circledast w_{j,k'}^{l-1,k} + b^{l-1,k}\right) \tag{5.20}$$

where $x_{i,j}$ represents a set of input feature maps, $\circledast$ indicates the convolutional operation, K' is the number of feature maps in the $(l-1)$-th layer, Y is the size of the convolution kernel running over the temporal domain of a signal, $w^{l-1,k} \in \mathbb{R}^{Y \times K'}$ is a local filter weight matrix, and $b^{l-1,k} \in \mathbb{R}$ is the bias.

However, the feature maps generated by each convolutional block can often be unwieldy in terms of their width, length, and number of channels, making them challenging to manage. This can lead to overfitting during training and high computational costs. An efficient solution to these challenges is applying a non-linear downsampling (NLDS) technique, known as a pooling layer, to reduce redundant features. Two well-known pooling techniques are average pooling and max pooling [12]. The pooling layer reduces the spatial size of the representation, thereby reducing the number of parameters and mitigating overfitting. The output map $p_{i,j}^{l,k}$ is obtained by taking the maximum or mean value over non-overlapping regions of the feature maps $z_{i,j}^{l,k}$ with a filter size of $\boldsymbol{Q}$, it's computed as follows:

$$p_{i,j}^{l,k} = \underset{(j-1)Q+1 \le j' \le jQ}{down} z_{i,j'}^{l-1,k} \tag{5.21}$$

where the downsampling function (down(·)) can be either average pooling or max pooling. While average and max pooling are commonly used for their simplicity and computational

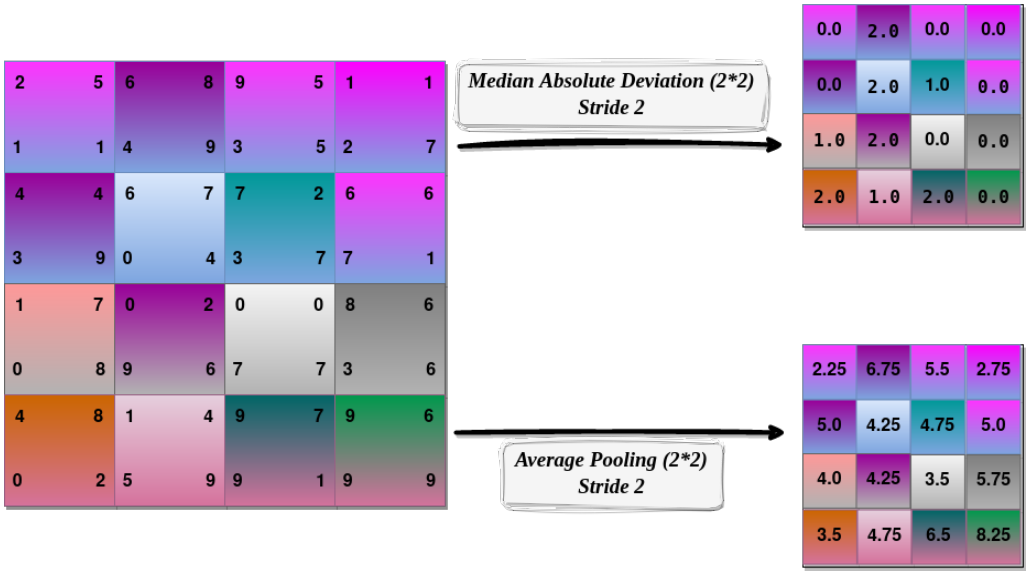


Figure 5.4: Adaptive Stochastic Pooling (ASP) process

efficiency, they may not be suitable for EHRs time series data containing outliers. These pooling operations assume homogeneous data, where spatial or temporal relationships can be summarized by aggregated values. However, in the presence of outliers, average and max pooling can lead to distorted representations and impact overall performance. They are sensitive to the magnitude of outliers, meaning a single extreme outlier can significantly influence the pooled summary, even if it is an anomaly or measurement error. This sensitivity can result in unstable and unreliable representations, particularly with noisy or heterogeneous EHR time series data.

Max pooling, which selects the maximum value within each pooling window, is especially problematic when outliers are present. Outliers can dominate the pooled representation, leading to biases and distortions in the output. On the other hand, average pooling performs poorly when there are many zero elements in the input features, causing a significant reduction in the feature map's characteristics [166]. Outliers can also strongly influence the average value, pulling it towards extreme values and resulting in an inaccurate representation of the data, especially if outliers are frequent or have a significant impact on overall statistics.

To address the issue of outliers in EHR time series data, it is beneficial to consider outlier-resistant pooling techniques or alternative methods that can handle outliers effectively. We propose a novel approach called Adaptive Stochastic Pooling (ASP), as illustrated in Fig. 5.4 which incorporates the Median Absolute Deviation (MAD) [91] into the pooling process. MAD, a measure of data spread, is less influenced by extreme values and allows the pooling operation to handle noisy or anomalous data points more robustly. By leveraging ASP with MAD, we can improve the pooling process and mitigate the negative impact of outliers in EHR time series analysis. This approach is mathematically expressed as follows:

$$ASP_{i,j}^{l,k} = p_{i,j}^{l,k} + \underset{(j-1)Q+1 \le j' \le jQ}{mad} z_{i,j'}^{l-1,k} \cdot \mathcal{N}(0,1)(p_{i,j}^{l,k}) \tag{5.22}$$

where: $ASP_{i,j}^{l,k}$ represents the pooled value at position (i,j) in the l-th layer and k-th

channel. $p_{i,j}^{l,k}$ is the pooling operation used for each pool. $\mathcal{N}(0,1)$ denotes a sample drawn from a standard normal distribution and mad is the median absolute deviation computed for each pool, expressed as:

$$\underset{(j-1)Q+1\leq j'\leq jQ}{mad} z_{i,j'}^{l-1,k} = M(|z_{i,j'}^{l-1,k} - M(z_{i,j'}^{l-1,k})|) \tag{5.23}$$

with M being the Median operation.

ASP also introduces randomness through sampling from a normal distribution, making the pooling operation non-deterministic. This helps prevent overfitting and provides regularization benefits [120, 168]. The random sampling allows the pooling operation to capture the inherent variability within each pool, which is particularly advantageous for capturing diverse patterns and fluctuations in EHR time-series data. It also enables the model to explore different possible pooling outcomes during training, leading to improved generalization. Furthermore, ASP retains more information compared to average and max pooling by combining the mean or max and MAD-adjusted samples within each pool. This approach captures both central tendencies and variability in the pooled representations, resulting in richer and more informative feature representations. This is particularly valuable when dealing with complex and non-linear relationships within EHR time-series data, where different patients or events may exhibit varying dynamics.

5.6 Limitations & Perspectives

Despite their considerable success, existing models [5, 124, 171] and MWTA-LSTM encounter substantial challenges in accurately modeling patient health trajectories, particularly when faced with varying measurement frequencies. This is because they commonly overlook a crucial factor (the frequency of clinical feature measurements) when adjusting the previous cell state. Instead, they primarily rely on the elapsed time to accommodate timing irregularities in patient EHR records. While this approach is valid for some variables with delayed impact on patient outcomes, it doesn't universally apply. For instance, a high level for variables like lactate or troponin T, while not frequently measured, can exert a prolonged and significant influence on a patient's illness course. Their infrequent measurement doesn't diminish their immense clinical significance, especially in conditions like tissue hypoxia or other serious medical situations.

Additionally, certain clinical interventions, such as renoprotective strategies [72], may not yield immediate observable effects on biomarkers like kidney function. However, they can wield a substantial influence on patient outcomes over time. Recognizing and tracking the long-term impact of such variables is pivotal for accurate patient management. Numerous studies [72, 146, 154] have emphasized the critical role of this consideration. Singer et al [146] have particularly emphasized the significance of elevated lactate levels in critically ill patients, linking them to increased mortality rates and extended hospital stays. Their study underscores the importance of factoring in earlier lactate measurements over an extended duration. Therefore, relying solely on elapsed time for adjusting the cell state could lead to an underestimation of the impact of variables that aren't measured frequently, potentially missing out on the full complexity of the clinical context.

In summary, the importance of features in different time-scales varies among different patients. For example, for the patient suffering from chronic disease, the feature extracted in the long term may be more representative for depicting the health status or in the case of end-stage renal disease (ESRD) patients, a sustained decline in blood albumin levels over the long term serves as a robust sign of malnutrition and declining health [8]. Conversely, for the patient diagnosed with acute disease, the short-term feature describes the health risk more precisely.

Hence, when adapting the cell state to account for timing irregularities, it's imperative to incorporate contextual information from the patient's history, encompassing both measurement frequency and elapsed times. This approach aids in distinguishing between short-term and long-term memory contributions within the cell state.

In addition to their triumphs, these models also face significant hurdles when applied to EHR data. Firstly, many of them lack the ability to provide interpretable results, a crucial aspect in healthcare domains. In real clinical settings, it's often more vital to discern interpretable patterns that capture informative disease progression than to achieve absolute predictive accuracy. This has led to the development of various attention-based neural networks aimed at generating interpretations for EHR data [24, 171] using attention mechanisms to identify meaningful visits and specific features that contribute to predictions.

Chapter 6

Adaptive Multi-Way Interpretable Time-Aware LSTM (AMITA)

"It's always good to take an orthogonal view of something. It develops ideas."

– Ken Thompson, C, Unix and beyond an interview with:

As discussed previously in section 5.6, existing models face significant hurdles in effectively modeling patient health trajectories. To address these limitations, we introduce a novel deep dynamic memory neural network named Adaptive Multi-Way Interpretable Time-Aware LSTM for irregularly collected sequential data (AMITA). The primary aim of AMITA is to harness medical records to remember illness trajectories and care processes, estimate current illness states, and predict future risks with a high degree of accuracy and predictive power.

To enhance its capabilities, AMITA extends the standard LSTM model in two key ways. Firstly, it incorporates frequency measurement and the most recent observation to enhance personalized predictive modeling of patient illnesses, enabling a more accurate understanding of the patient's condition. Secondly, it parameterizes the cell state to handle irregular timing effectively, utilizing both elapsed times and a frequency-based decay factor, which considers both measurement frequency and contextual information. Furthermore, the model capitalizes on both to comprehend the impact of interventions on the course of

illness on the cell state, facilitating the memorization of illness courses and improving its ability to capture the temporal dynamics of healthcare data, accommodating variations and irregularities in event and observation timing.

In alignment with the foundational concept introduced in [5], we have systematically dissected the preceding memory $\boldsymbol{c}_{t-1}$ into two distinct constituents. One is the long-term component, designed to encapsulate the enduring impact of earlier events, and the other involves adjusted short-term decayed memories, tailored to capture recent developments and immediate influences.

Following this decomposition, the adjusted short-term memory undergoes a discounting mechanism, integrating a temporal decay factor. This factor is intricately derived from a fusion of elapsed time ($\boldsymbol{\Delta}_t$) and a frequency-based decay factor, which considers both measurement frequency and contextual information. Thus, prior data's significance is thoughtfully weighted, ensuring variables with infrequent measurements, potentially bearing crucial implications for a patient's overall health, maintain relevance even with extended time gaps between measurements.

Ultimately, the long-term and adjusted short-term decayed memories converge, resulting in an updated memory that seamlessly integrates the prolonged impact of historical events with the immediate influences and recent developments for a comprehensive understanding of the patient's evolving health status.

This adjustment enables the customization of memory content in the new cell state unit, fostering a more comprehensive understanding of the patient's health trajectory from their EHR data. The merging of these memory components allows the model to leverage the strengths of both short-term and long-term information, offering a balanced and accurate reflection of the patient's evolving condition over time. Through the combination of these memory components, the model gains the ability to make more reliable predictions about future events based on the patient's historical data, resulting in predictions that are both accurate and interpretable.

In healthcare, patient's data often arrives at irregular intervals due to various reasons, like differences in clinical protocols or the nature of the condition being monitored. To handle this, we've extended the forget gate mechanism of AMITA. This extension takes into consideration both the time gap between consecutive events, how frequently measurements are taken and the contextual information from the patient's history. This enables the model to adapt to the unique time patterns in each patient's data. The mechanism also promotes sparsity in time, focusing on critical data points without being overly influenced by less informative or infrequent measurements. Objectively, this research makes the following contributions:

- Innovative Framework for Handling Timing Irregularities: AMITA advances LSTM's standard mechanisms by incorporating the time interval between events, measurement frequency, and contextual data from patient's medical history. This enables nuanced memory cell adjustments diminishing the influence of previous memories with increasing time lags while preserving the impacts of significant past events. Moreover, we integrate dual gate mechanisms to better capture the effects of clinical

interventions on the current states of illness.

- Enhanced Forget Gate Mechanism: Recognizing the irregularity of data acquisition in healthcare, we have refined AMITA's forget gate to consider not just the timing but also the frequency of events and the broader historical context of each patient. This improvement helps the model adapt to unique patient timelines, ensuring a focus on critical data while minimizing the impact of less relevant or infrequent data points.
- Furthermore, we introduce an innovative approach to enhance interpretability by conducting a thorough analysis that incorporates both attention values and frequency weights for each feature. This method allows us to identify and emphasize the most critical features relevant to each prediction task, including ICU mortality and Length of Stay (LOS).

In section 7, we showcase the superior performance of our proposed model through empirical experiments conducted on two real-world clinical datasets, surpassing both the current state-of-the-art models and other robust baselines.. To the best of our knowledge, our research represents the first attempt to jointly incorporate the time interval between events, the frequency of measurements and the contextual information from the patient's history into the LSTM unit for adjusting the cell state.

AMITA is an advanced end-to-end deep-learning architecture tailored for capturing intricate temporal patterns within sequential data that may exhibit irregularities in timing. This model comprises three primary layers. The first layer (1) is constructed upon an extended and adapted version of LSTM [63]. Here, the memory cells are specifically tailored to accommodate the temporal irregularities present in a patient's medical records. The second layer (2) employs a mixture attention mechanism strategy pooling, while the third layer (3) serves as the output layer, as shown in Fig. 6.1. The concept behind AMITA is to utilize a hidden state matrix and devise a corresponding update scheme. This approach ensures that each element, such as a row, within the hidden matrix exclusively captures information related to a specific variable from the input data. Next, we discuss each layer in detail.

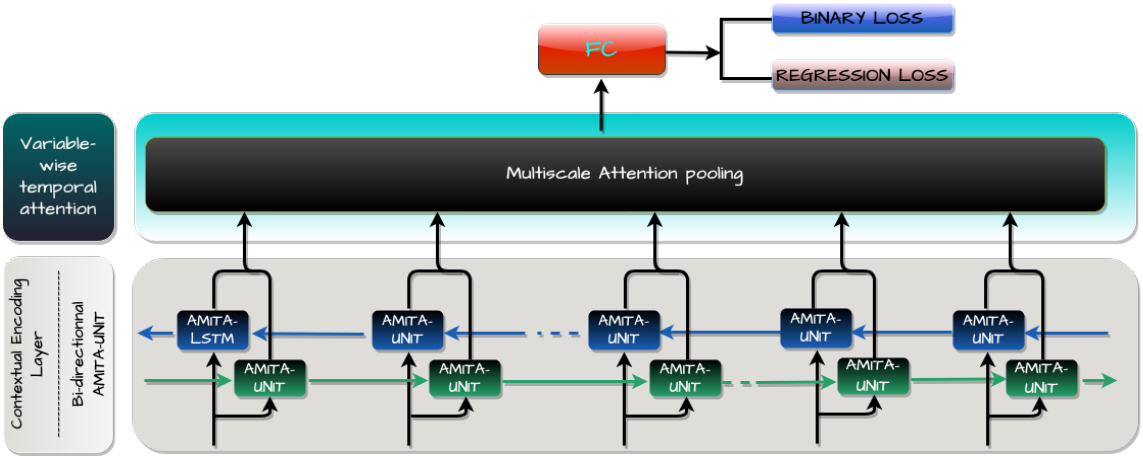


Figure 6.1: Adaptive Multi-Way Interpretable Time-Aware LSTM (AMITA).

6.1 AMITA's Unit

The duration between successive events within a patient's hospitalization can vary widely, from minutes to weeks. Given the diminishing relevance of past events over time, mitigating their influence on current assessments becomes imperative. Thus, considering elapsed time is crucial for accurate predictions of present outcomes. Moreover, these diverse intervals can serve as indicators of shifts in a patient's health status; frequent events may signify acute conditions, whereas infrequent occurrences could suggest relative stability. Our adoption of the AMITA-unit is motivated by its adeptness in managing temporal irregularities and comprehending the complexities of a patient's medical trajectory. Through the incorporation of measurement frequency and elapsed time, we attain a nuanced comprehension that distinguishes between immediate effects, where frequent measurements are pivotal, and long-term impacts, where earlier data points retain significance. AMITA's Unit is illustrated in Fig. 6.2.

As $\boldsymbol{c}_{t-1}$ systematically monitors the long-term history, and the preceding hidden state $\boldsymbol{h}_{t-1}$ encapsulates recent trends and patterns, we leverage both components to enable the model to capture temporal patterns within $\boldsymbol{x}_t$, which contains information on the patient's present diagnosis, providing insights into both short-term and long-term medical condi-

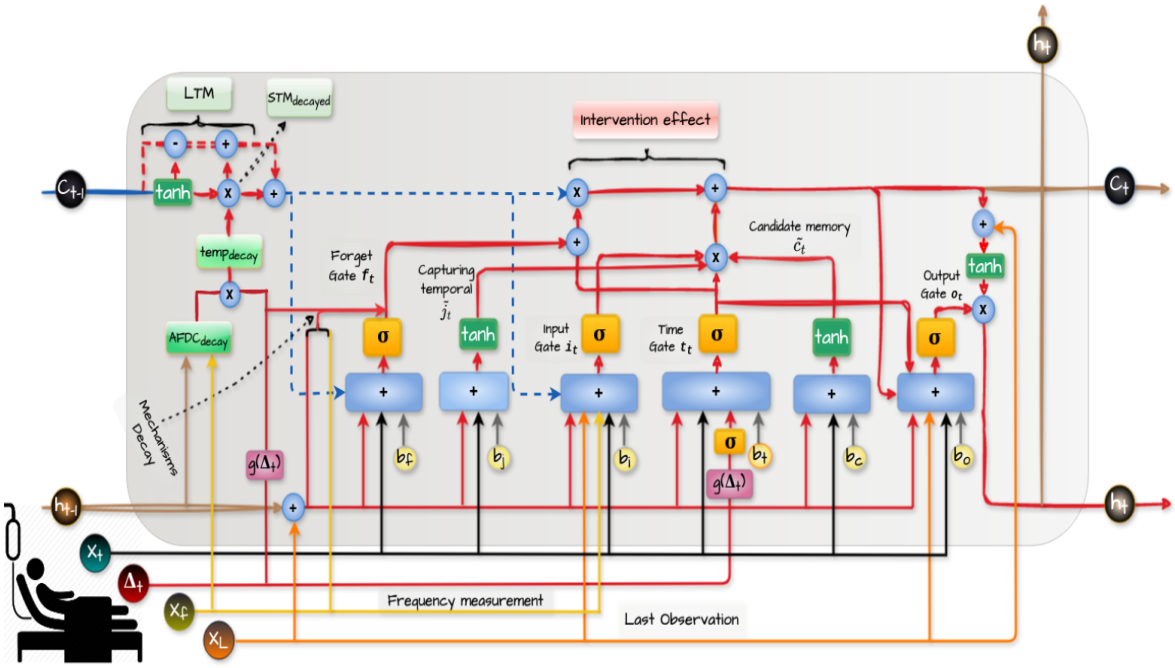


Figure 6.2: AMITA Unit on analyzing patient's EHR records. Sketch red arrows correspond to the modifications compared to LSTM Unit. Shade blue boxes indicate networks and shade blue circles denote point-wise operators.

tions.

In patient's records EHR data, short-term changes in clinical features can be crucial for immediate care decisions. For example, sudden changes in vital signs may signal a medical emergency. $\boldsymbol{c}^S_{t-1}$ expressed in Eq. (6.1), is a representation of how the current state of the cell memory $\boldsymbol{c}_{t-1}$ emphasizes recent information.

$$\boldsymbol{c}^S_{t-1} = \overbrace{\tanh(\boldsymbol{W}_s \circledast \boldsymbol{c}_{t-1} + \boldsymbol{b}_s)}^{\text{Short term memory[5]}} \tag{6.1}$$

where $\circledast$ is defined as the product of two tensors. $\boldsymbol{c}^S_{t-1}$ ensures that the model gives more weight to recent measurements, reflecting their importance in short-term patient assessment.

In real-world healthcare settings, measurements and interventions are often recorded at varying frequencies, which can lead to irregular time intervals between consecutive events. Therefore, when adjusting the cell state to account for this irregularity, it's crucial to apply the frequency-based decay so that the model may accurately capture the evolving nature of a patient's condition. By incorporating the frequency-based decay inclunding both the frequency of measurements and the contextual information from the patient's history ($\boldsymbol{h}_{t-1}$), the model can appropriately weights the importance of previous data. This ensures that infrequently measured variables, which may have a significant impact on a patient's overall health, are appropriately weighted, preventing the loss of critical information even when the time gap between measurements widens. Therefore, leading to more accurate predictions and a better understanding of patient outcomes over time, where different clinical

variables may be measured with varying frequencies and play diverse roles in patient care. It's modelled as follows:

$$freq_d = \overbrace{\sigma(e^{-\beta * \boldsymbol{f}_{\boldsymbol{x}_d}} \circledast \boldsymbol{W}_{\boldsymbol{f}_{\boldsymbol{x}_d}} + \boldsymbol{h}_{t-1} \circledast \boldsymbol{W}_{h_{f_{\boldsymbol{x}_d}}} + \boldsymbol{b}_{\boldsymbol{f}_{\boldsymbol{x}_d}})}^{\text{Adaptive frequency decay with contextual information(AFDC)}} \tag{6.2}$$

where e stands for exponential, $\beta \in [0, 1]$ is the hyper-parameter (constant) that will determine the rate at which the transformed values decrease as the feature's frequency measurement increases. $\boldsymbol{W}_{\boldsymbol{f}_{\boldsymbol{x}_d}}$, $\boldsymbol{W}_{h_{f_{\boldsymbol{x}_d}}}$, $\boldsymbol{b}_{\boldsymbol{f}_{\boldsymbol{x}_d}}$ represents the weights and biases.

The frequency-based decay as expressed in Eq. (6.2) and illustrated in Fig. 6.2, dynamically adjusts the influence of clinical variables in the model based on their measurement frequency and historical context. The purpose of this factor is to modulate the impact of short-term memory on the current state by ensuring that both frequently and infrequently measured features are appropriately weighted in the prediction process, reflecting their clinical significance accurately.

In patient's EHR records data, different clinical variables may have varying temporal profiles as the impact of some variables may become evident immediately, while some may not be immediately evident but becomes more pronounced over time. Therefore, it's important to adjust the contribution of short-term memory accordingly based on historical profile of each clinical variable, it's expressed in Eq. (6.3) as follows:

$$temp_{decay} = \overbrace{g(\boldsymbol{\Delta}_t) \odot freq_d}^{\text{Temporal decay factor}} \tag{6.3}$$

where $g(\boldsymbol{\Delta}_t) = (\boldsymbol{W}_{\boldsymbol{\Delta}_t} \circledast \frac{1}{log(e+\boldsymbol{\Delta}_t)})$ [124], particularly suitable for datasets that involve large elapsed times.

The temporal decay factor actively modulates the influence of short-term memory according to the time elapsed between clinical events. This adjustment ensures that vital historical information is not prematurely discarded, thereby improving the model's ability to adapt to the unique clinical context.

In clinical settings, the evolution of certain clinical features can be rapid, while others may change more gradually. Monitoring these temporal changes is crucial for gaining insights into a patient's disease progression. This temporal modeling is represented by Eq. (6.4), where the adjusted short-term memory decayed (STM) is calculated as:

$$\hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}} = \overbrace{\boldsymbol{c}_{t-1}^{S} \odot temp_{decay}}^{\text{Adjusted short-term memory decayed (STM)}} \tag{6.4}$$

$\hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}}$ as illustrated in Fig. 6.2, plays a vital role in the model's ability to capture the dynamic adaptation of short-term memory to changing clinical conditions, providing a mechanism for more accurate representation of disease progression over time. By appropriately discounting short-term memory, the model considers the long-term effects of various clinical variables, contributing to more precise and reliable predictions crucial for clinical decision-making, especially considering that certain clinical features, interventions, or

treatments may not immediately manifest their full impact but can significantly influence patient outcomes over time. Moreover, it ensures that less frequently measured variables are not overshadowed or discarded, preserving their importance in the overall patient assessment, leading to more accurate model predictions and improved interpretability, essential aspects in the context of clinical applications.

The long-term memory component, denoted as $\boldsymbol{c}_{t-1}^{\text{LTM}}$, responsible for preserving and incorporating the historical context of a patient's health journey over an extended period is computed as follows:

$$\boldsymbol{c}_{t-1}^{\text{LTM}} = \overbrace{\boldsymbol{c}_{t-1} - \boldsymbol{c}_{t-1}^{S} + \hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}}}^{\text{Long-term memory}} \tag{6.5}$$

$\boldsymbol{c}_{t-1}^{\text{LTM}}$ as drawn in Fig. 6.2, ensures that the influence of previous events and measurements that occurred over a more extended timeframe is appropriately captured. Unlike short-term memory, which is more sensitive to recent events, $\boldsymbol{c}_{t-1}^{\text{LTM}}$ encapsulates enduring patterns and trends in a patient's health data. $\boldsymbol{c}_{t-1}^{\text{LTM}}$ plays a crucial role in adjusting the cell state to account for timing irregularities in EHR timeseries data. Its utility is akin to that of $\hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}}$ Eq. (6.4) in the sense that it contributes to the overall memory of the cell state, albeit with a focus on long-term information retention.

In summary, $\boldsymbol{c}_{t-1}^{\text{LTM}}$ serves as a reservoir of enduring clinical information, allowing the model to grasp the sustained trends and patterns in a patient's health data. Including $\boldsymbol{c}_{t-1}^{\text{LTM}}$ into the current cell also adds stability to the model's outputs, as it helps prevent erratic predictions based solely on recent events, providing a more reliable basis for forecasting the patient's future health outcomes.

In EHR data, irregular time intervals between events are common. Combining memories with different time dependencies allows the model to handle such irregularities more effectively, capturing the influence of both recent and distant events without relying solely on uniform time gaps. Therefore the updated cell memory is computed as follows:

$$\boldsymbol{c}_{t-1}^{*} = \overbrace{\hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}} + \boldsymbol{c}_{t-1}^{\text{LTM}}}^{\text{Adjusted previous memory}} \tag{6.6}$$

In summary, the combination of short-term and long-term memory on the current cell state, achieved by adding $\hat{\boldsymbol{c}}_{t-1}^{S_{\text{decayed}}}$ Eq. (6.4) and $\boldsymbol{c}_{t-1}^{\text{LTM}}$ Eq. (6.5), results in a comprehensive cell state $\boldsymbol{c}_{t-1}^{*}$ that appropriately accounts for both immediate and prolonged influences on the patient's health trajectory, aiding the model in discerning between transient fluctuations and sustained trends.

In an LSTM, the hidden state $\boldsymbol{h_t}$ serves as a memory for the model, capturing information from previous time steps. We propose to adjust the hidden state $\boldsymbol{h}_t$, the previous hidden state $\boldsymbol{h}_{t-1}$, expressed in Eq. (6.8) and the output gate $\boldsymbol{o}_t$ as expressed in Eq. (6.14) with the last observation $\boldsymbol{L}_{x_t}$ Eq. (6.7) of each feature. This means that at each time step, the hidden state is modified to include the latest record data of each feature.

It's mathematically expressed as follows:

$$\boldsymbol{L}_{\boldsymbol{x}_t} = \mathbf{ELU}(\boldsymbol{W}_{\boldsymbol{x}_{\text{last}}} \circledast \boldsymbol{X}_{\text{last}} + \boldsymbol{b}_{\boldsymbol{x}_{\text{last}}}) \tag{6.7}$$

where **ELU** is the activation function.

$$\begin{aligned} \boldsymbol{h}^*_{t-1} &= \overbrace{\boldsymbol{h}_{t-1} + \boldsymbol{L}_{\boldsymbol{x}_t}}^{\text{Adjusted previous memory}} \\ \boldsymbol{h}_t &= \overbrace{\boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t + \boldsymbol{L}_{x_t})}^{\text{Adjusted current memory}} \end{aligned} \tag{6.8}$$

The objective of this modification is to empower the model with the ability to assimilate the most recent information into its memory. Given that the latest record is likely a more accurate reflection of the patient's current health status compared to earlier measurements, this adjustment not only enhances the model's responsiveness to real-time changes but also serves as a strategy to alleviate the challenges associated with missing data.

6.1.1 Capturing temporal dependencies

EHR data often contains longitudinal information about a patient's medical history, to better capture temporal dependencies and relationships between past (knowledge from previous patient histories) and current patient records. We update the previous hidden state $\boldsymbol{h}^*_{t-1}$ by the current input $\boldsymbol{x}_t$, the hidden state update is defined as:

$$\tilde{\boldsymbol{j}}_t = \tanh(\boldsymbol{W}_j \circledast \boldsymbol{h}^*_{t-1} + U_j \circledast \boldsymbol{x}_t + \boldsymbol{b}_j) \tag{6.9}$$

where $\tilde{\boldsymbol{j}}_t = [\boldsymbol{j}^1_t, \ldots, \boldsymbol{j}^N_t]^T$ has the same shape as hidden state matrix $\boldsymbol{h}^*_{t-1} \in \mathbb{R}^{N \times d}$. Each element $\boldsymbol{j}^n_t \in \mathbb{R}^d$ corresponds to the update of the hidden state w.r.t. input variable d. Term $W_j \circledast \boldsymbol{h}^*_{t-1}$ and $U_j \circledast \boldsymbol{x}_t$ respectively capture the update from the hidden states at the previous step and the new input. The tensor-dot operation $\circledast$ is defined as the product of two tensors along the N axis, e.g., $W_j \circledast \boldsymbol{h}^*_{t-1} = [\boldsymbol{W}^1_j \boldsymbol{h}^1_{t-1}, \ldots, \boldsymbol{W}^N_j \boldsymbol{h}^N_{t-1}]^T$ where $W^n_j \boldsymbol{h}^n_{t-1} \in \mathbb{R}^d$.

The new updated hidden state $\tilde{\boldsymbol{j}}_t$ allows the model to explicitly capture and consider temporal dependencies by incorporating information from both the previous hidden state $\boldsymbol{h}^*_{t-1}$ and the current input $\boldsymbol{x}_t$, therefore contributing to a more informed and context-aware representation of patient's medical history.

6.1.2 Capturing time irregularity

To address time irregularity, we make use of the frequency measurement of each clinical variable, to gain a better understanding of how these variables are changing over time, which can give a more comprehensive picture of the patient's health status to identify trends that may be important predictors. This can also help addressing irregularity timing so mitigate issues related to missing data as a clinical variable may not be measured at every encounter.

The modified input gate $\boldsymbol{i}_t$ and updated forget gate value $\boldsymbol{f}_t$ are expressed as follows by Eq. (6.10) and Eq. (6.11):

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{i_h} \circledast \boldsymbol{h}^*_{t-1} + \boldsymbol{U}_i \circledast \boldsymbol{x}_t + \boldsymbol{W}_{i_c} \odot \boldsymbol{c}^*_{t-1} + \mathit{freq}_d + \boldsymbol{L}_{x_t} + \boldsymbol{b}_i) \tag{6.10}$$

By incorporating the frequency decay, as expressed in Eq. (6.2) and the last observation $\boldsymbol{L}_{x_t}$ by Eq. (6.7) into the input gate $\boldsymbol{i}_t$ Eq. (6.10), it can acts as a discerning filter. This refinement empowers the model to efficiently sift through irrelevant information, honing in on essential features and thereby enhancing its ability to comprehend the temporal context of each variable in the patient's medical history. This integration also facilitates the effective use of longitudinal information, providing a more profound understanding of the patient's health progression over time. Furthermore, it enables the dynamic adjustment of variable weights within the model, ensuring that variables occurring more frequently receive higher weights, indicative of their relative importance in the sequence. Simultaneously, attention is paid to variables with lower frequencies, preventing their neglect and underscoring their significance in the learning process.

$$\boldsymbol{f_t}^{\text{new}} = \overbrace{\boldsymbol{f}_t \odot g(\boldsymbol{\Delta}_t)}^{\text{Time-Based Decay}} + \overbrace{(1 - \boldsymbol{f}_t) \odot \mathit{freq}_d}^{\text{Frequency-Weighted Decay}} \tag{6.11}$$

where $\boldsymbol{f_t}^{\text{new}}$ Eq. (6.11) represents the updated forget gate value, $\boldsymbol{f}_t$ is the original forget gate value, as expressed in Eq. (4.22) and freq_d represents the frequency decay term expressed in Eq. (6.2).

Incorporating these two decay terms when adjusting the forget gate results in several benefits, which are outlined below.

- **Time-Based Decay**: The initial component in Eq. (6.11) introduces a temporal decay mechanism, influencing the activation value of the forget gate based on the time elapsed since the last update. This imparts a dynamic capability to the model, allowing it to adapt its forgetfulness proportionately as time progresses. This adaptation is particularly pertinent in medical context, where the significance of past medical history naturally diminishes over time. Such adaptability becomes crucial to prevent the model from retaining outdated information, a pivotal feature with diverse applications including EHR data analysis, language modeling, and speech recognition.

- **Frequency-Weighted Decay**: An inherent challenge in handling EHR data lies in the varying frequencies at which different medical conditions or events manifest. When the forget gate treats all inputs uniformly, there's a risk of overlooking less frequent yet critical information. The second component in Eq. (6.11) introduces an adaptive decay mechanism to the forget gate's activation value, leveraging both the frequency of input data and contextual insights from the patient's history (illness state).

  This adaptive decay mechanism plays a crucial role in preserving sequential context and long-term dependencies within patient records. The incorporation of the previous hidden state $\boldsymbol{h}^*_{t-1}$ in Eq. (6.2) ensures that the forget gate retains a memory of past states, a pivotal aspect in comprehending the patient's evolving condition. The mechanism strikes a balance between forgetting less relevant features and retaining essential sequential patterns. By allowing the model to learn patient-specific patterns and dynamically adjust its memory retention strategy, this approach enhances interpretability. It provides a transparent view of how the model's decision-making process is influenced not only by feature frequency but also by historical context.

In essence, it promotes a nuanced understanding of the interplay between the frequency of medical events and their contextual relevance, resulting in more informed and interpretable outcomes.

6.1.3 Modeling the Impact of Interventions

In understanding the dynamic health status of a patient, continuous monitoring of clinical feature measurements over time is essential. The elapsed time vector functions as a pivotal intervention, allowing the tracking of measurement frequency for each clinical feature, providing a direct reflection of the patient's evolving health condition. Brief intervals between measurements for critical features may signify rapid deterioration, whereas longer intervals may suggest stability. Consequently, the elapsed time vector is a fundamental tool for the ongoing evaluation of a patient's health status, aligning with the overarching objectives of interventions, whether to cure diseases or alleviate the patient's illness. The computation of the elapsed time vector is defined by:

$$\boldsymbol{t}_t = \sigma(\boldsymbol{W}_{x_t} \circledast \boldsymbol{x}_t + \boldsymbol{\sigma}_{\boldsymbol{\Delta}_t}(g(\boldsymbol{\Delta}_t)) + \boldsymbol{b}_t) \tag{6.12}$$

Here, σ denotes the sigmoid function, $\boldsymbol{\Delta}_t$ represents the time interval between records for each clinical feature at time spans $\boldsymbol{t-1}$ and $\boldsymbol{t}$, and this time gap can vary from hours to days or even years.

The intervention effect on the current cell state, denoted as $\boldsymbol{c_t}$ is captured through Eq. (6.6) and Eq. (6.12):

$$\boldsymbol{c}_t = (\boldsymbol{f_t}^{\text{new}} + \boldsymbol{t}_t) \odot \boldsymbol{c}^*_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{j}}_t \odot \boldsymbol{t}_t \odot \tilde{\boldsymbol{c}}_t \tag{6.13}$$

By consolidating $\boldsymbol{f_t}^{\text{new}}$ and $\boldsymbol{t}_t$ before interacting with the previous cell state $\boldsymbol{c}^*_{t-1}$, we establish a transparent and interpretable link between the intervention effect ($\boldsymbol{t}_t$) and the inherent memory mechanism of the model ($\boldsymbol{f_t}^{\text{new}}$). This approach diminishes reliance on previous memory as the time gap between measurements widens, aligning with the clinical insight that earlier measurements may have a diminishing impact on current health status.

This adaptability is crucial, allowing the model to recognize the significance of prior information (captured by $\boldsymbol{f_t}^{\text{new}}$) while considering the impact of recent interventions (influenced by $\boldsymbol{t}_t$). This is especially pertinent in scenarios where recent changes in clinical status better indicate the patient's current condition. The time gate $\boldsymbol{t}_t$ not only filters input information through the input gate $\boldsymbol{i}_t$ but also controls the impact of $\boldsymbol{x}_t$ on the patient's current state, modeling their long-term medical condition.

By selectively updating the cell state and controlling the influence of past illness and current information, the model becomes less sensitive to noise, enhancing interpretability. This clarity in understanding how past information influences the current state makes the model applicable to various healthcare tasks, including predicting patient outcomes, disease progression modeling, anomaly detection, and treatment response prediction.

In the dynamic landscape of healthcare, the implications of events or interventions can vary depending on their temporal context. To account for this, the output gate Eq. (4.27), governing illness states, is dynamically influenced by the ongoing intervention Eq. (6.12) and the latest information Eq. (6.7), as expressed in Eq. (6.14):

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{o_h} \circledast \boldsymbol{h}^*_{t-1} + \boldsymbol{U}_o \circledast \boldsymbol{x}_t + \boldsymbol{W}_{o_c} \odot \boldsymbol{c}_t + \boldsymbol{t}_t + \boldsymbol{L}_{x_t} + \boldsymbol{b}_o) \tag{6.14}$$

Integrating the most recent feature observations and the temporal context designed by $\boldsymbol{t_t}$, the output gate $\boldsymbol{o}_t$ acts as a discerning filter, empowering the model to selectively emphasize crucial information while adeptly filtering out extraneous details. Moreover, this adaptability enables the model to unravel patient-specific patterns, such as responses to treatments or disease progression, culminating in more precise and personalized predictions tailored to the intricacies of individual healthcare journeys.

6.2 Mixture Attention

After processing a sequence of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ through AMITA Unit, we generate a sequence of hidden state matrices corresponding to each event $\{\boldsymbol{h}^e_1, \ldots, \boldsymbol{h}^e_T\}$. The specific sequence of hidden states related to variable $\boldsymbol{x}^d$ is then extracted as $\left\{\boldsymbol{h}_1^{\boldsymbol{x}^d}, \ldots, \boldsymbol{h}_T^{\boldsymbol{x}^d}\right\}$.

A mixture attention pooling mechanism, initially proposed in [50], is then applied. This mechanism involves initially applying temporal attention to the sequence of hidden states corresponding to each variable, resulting in a summarized history for each variable. Subsequently, by utilizing the history-enriched hidden state of each variable, variable attention is computed to merge variable-specific states. These two steps are integrated into a probabilistic mixture model [11, 177], facilitating subsequent learning, prediction, and interpretation. Specifically, the context vector $\boldsymbol{g}_t^{\boldsymbol{x}^d}$ of the variable-wise temporal attention mechanism regarding $\boldsymbol{x}^d$ is computed as weighted sum of hidden states:

$$\boldsymbol{g}_t^{\boldsymbol{x}^d} = \sum_{k=1}^{t_0} \alpha_{t,k}^{\boldsymbol{x}^d} \boldsymbol{h}_{t-t_0+k}^{\boldsymbol{x}^d} \tag{6.15}$$

where the temporal attention weight $\alpha_{t,k}^{\boldsymbol{x}^d}$ is determined (equivalent to applying the softmax function) by:

$$\alpha_{t,k}^{\boldsymbol{x}^d} = \frac{\exp(f^n_{\boldsymbol{x}^d}(\boldsymbol{h}_{t-t_0+j}^{\boldsymbol{x}^d}))}{\sum_{j=1}^{t_0} \exp(f^n_{\boldsymbol{x}^d}(\boldsymbol{h}_{t-t_0+j}^{\boldsymbol{x}^d}))} \tag{6.16}$$

Then, we obtain the variable state $\boldsymbol{s}_t^{\boldsymbol{x}^d}$ by concatenating $\boldsymbol{h}_t^{\boldsymbol{x}^d}$ and $\boldsymbol{g}_t^{\boldsymbol{x}^d}$:

$$\boldsymbol{s}_t^{\boldsymbol{x}^d} = [\boldsymbol{h}_t^{\boldsymbol{x}^d} \oplus \boldsymbol{g}_t^{\boldsymbol{x}^d}] \tag{6.17}$$

In each prediction time step, the variable states of all input variables are collectively employed to build variable attention, thereby enabling the representation of variable importance. The variable attention weight is calculated by:

$$\beta_t^{\boldsymbol{x}^d} = \frac{\exp(f^s_{\boldsymbol{x}^d}(\boldsymbol{s}_t^{\boldsymbol{x}^d}))}{\sum_{k=1}^{D} \exp(f^s_{\boldsymbol{x}^d}(\boldsymbol{s}_t^k))} \tag{6.18}$$

For $p(y_{T+1}|\boldsymbol{z}_{T+1} = \boldsymbol{x}^d, \boldsymbol{h}_T^{\boldsymbol{x}^d} \oplus \boldsymbol{g}^{\boldsymbol{x}^d})$, without loss of generality, a Gaussian output distribution parameterized by $[\mu_t^{\boldsymbol{x}^d}, \boldsymbol{s}_t^{\boldsymbol{x}^d}]$ is used as follows:

$$\begin{aligned}\mu_t^{\boldsymbol{x}^d} &= \varphi_t^{\boldsymbol{x}^d}(\boldsymbol{s}_t^{\boldsymbol{x}^d}) \\ &= \boldsymbol{W}_{\mu_t^{\boldsymbol{x}^d}} \boldsymbol{s}_t^{\boldsymbol{x}^d} + \boldsymbol{b}_{\mu_t^{\boldsymbol{x}^d}}\end{aligned} \tag{6.19}$$

$$\hat{y}_t = \sum_{\boldsymbol{x}^d=1}^{D} \beta_t^{\boldsymbol{x}^d} \mu_t^{\boldsymbol{x}^d} \tag{6.20}$$

Here, $f_{\boldsymbol{x}^d}^n(\cdot)$, $f_{\boldsymbol{x}^d}^s(\cdot)$ and $\varphi_{\boldsymbol{x}^d}$ can take the form of any distribution, such as a feed-forward neural network.

Building upon this attention pooling mechanism, we investigate four (4) distinct pooling strategies: **sum**, **max**, **mean**, and **global mean** (min, max, mean, std). The rationale behind their utilization is outlined below:

(1) **Normalized sum:** In situations where a patient's condition is influenced by multiple variables, the sum attention method ensures a comprehensive interpretation by taking into account all contributing factors in proportion to their impact on the overall decision-making process.

(2) **Normalized mean:** In the clinical context, understanding the average contribution of each variable is essential for clinical decision support systems. This strategy helps in balancing the overall influence of variables, making it particularly valuable when there are multiple factors contributing to a patient's health status and also preventing undue emphasis on outliers.

(3) **Normalized max:** In the ICU setting, where extreme values or critical events hold significant importance in evaluating a patient's condition. This approach offers insights into the dominant factors carrying substantial information that influences a patient's health trajectory. It proves invaluable in emergency scenarios or situations requiring swift medical interventions, enabling targeted and personalized care.

(4) **Normalized global mean:** Gathering information from various statistical metrics for variables provides a comprehensive understanding of their combined impact. This aggregation method aims to condense and interpret attention weights by calculating the mean across maximum, minimum, standard deviation, and mean values for variables. The inclusion of these metrics improves interpretive robustness by considering the influence of outliers in the attention weights on the overall interpretation.

Algorithm A.2 describes the whole training of **AMITA**, as shown in Fig. 6.2.

Algorithm 2 AMITA forward propagation. The process is repeated by $\boldsymbol{n}_{epochs}$ epochs

Inputs:
$[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]$: Input features
$[\boldsymbol{\Delta}_{t_1}, \ldots, \boldsymbol{\Delta}_{t_N}]$: Elapsed times
$[\boldsymbol{f}_{x_1}, \ldots, \boldsymbol{f}_{x_N}]$: Frequency measurement
$[\boldsymbol{x}_{last_1}, \ldots, \boldsymbol{x}_{last_N}]$: Last observation
$[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N]$: Corresponding outputs
for $\boldsymbol{e} \leftarrow 1$ to $\boldsymbol{n}_{epochs}$ **do**
    $\boldsymbol{T}_i$ being the number of time steps of the series for a given patient $\boldsymbol{i}$
    for each step $\boldsymbol{t} \leftarrow 1$ to $\boldsymbol{T}_i$ **do**
        $*$ Compute the adjusted previous memory cell $\boldsymbol{c}^*_{t-1}$ using (Eq. 6.4) and Eq. (6.5)
        $*$ Compute the adjusted $\boldsymbol{h}^*_{t-1}$ using (Eq. 6.7)
        $*$ Compute 5 gates: $\boldsymbol{t}_t$ (Eq. 6.12), $\boldsymbol{i}_t$ (Eq. 6.10), $\boldsymbol{f_t}^{\text{new}}$ (Eq. 6.11), $\boldsymbol{o}_t$ (Eq. 6.14) and $\tilde{\boldsymbol{j}}_t$ (Eq. 6.9)
        $*$ Compute $\boldsymbol{h}_t$ (Eq. 6.8) and $\boldsymbol{c}_t$ (Eq. 6.13)
    end for
    if the task is a classification **then**
        $*$ Compute the predictive probability using (Eq. 7.1)
        $*$ Compute the log loss.
    else
        $*$ Compute the mse probability using (Eq. 7.2)
        $*$ Compute the difference between the predicted value and the ground truth.
    end if
end for

Chapter 7

Experiments settings & Results

"The best and safest method of philosophizing seems to be first to inquire diligently into the properties of things, and establishing those properties by experiments, and then to proceed more slowly to hypotheses for the explanation of them."

– Isaac Newton, The Life of

Machine learning algorithms applied to EHR datasets for predicting medical events and mining clinical insight have been demonstrated to potentially provide powerful tools to aid clinicians in their work [18, 52, 129, 133]. It's a challenging and time-consuming process for clinicians to get a full overview of patient EHR data. Time becomes critical when a patient is admitted to the ICU, where quick clinical insight into a patient's health data can be life-saving and aid health organizations to structure and plan better.

In order to develop algorithms to perform these tasks, large amounts of data are required. Unfortunately, EHR datasets are difficult to access without a vigorous screening process from the health organizations, or even altogether impossible, because of the sensitivity of the data. Luckily, there are some datasets available that contain real, non-fabricated data that have been anonymized to protect patients' and clinicians' identities.

In this chapter, we experiment with predicting valuable medical information from the MIMIC-III critical care database [70] and eICU [125], using traditional machine learning algorithms. The experiment revolves around the prediction of two cases: mortality and length of stay. Both cases need to be assessed early when a patient is admitted to a hospital. By trying to predict the mortality of a patient, the trained model can give us an indicator that a patient is going to pass away during the hospital visit. Length of stay (LOS) is a factor that helps assess the efficiency of hospital management and quality of care's quality

received by the patient.

This experiment will be used as an indicator to see if it's possible for a researcher to extract valuable information out of EHR data using regular machine learning algorithms and standardized preprocessing techniques. By doing so we hope to expose and shed light on some of the difficulties and problematic areas of performing EHRs analytics with machine learning algorithms on real EHRs data from a hospital.

7.1 Cohort selection

To define our relevant cohort, we initiated the process by excluding patients under the age of 15 and those lacking essential temporal data records. It's noteworthy that a patient's hospital admission may encompass multiple episodes of intensive care, involving transfers between various intensive care units during their stay. For the purpose of this study, we considered all instances of intensive care associated with a patient throughout their hospitalization. After applying these exclusion criteria, we identified a total of 53,103 distinct intensive care unit stays eligible for our prediction tasks in the case of MIMIC III. Within this cohort, the median age was 65.7 years, with an interquartile range (Q1-Q3: 52.0 - 76.9), and in-hospital and ICU mortality rates were 12.4% and 8.8%, respectively. The average length of stay in intensive care units was 4.15 days. Turning to eICU, we obtained 190,935 unique intensive care unit stays with a median age of 64.0 years and an interquartile range (Q1-Q3: 52.0 - 75.0). To ensure consistency across tasks, we focused on the first 24 and 48 hours of ICU data for each patient. Further details are provided in Table 7.1.

Table 7.1: Baseline characteristics and in-hospital mortality outcome measures in our cohort. LOS (Length of Stay). Continuous variables are presented as Median [InterQuartile Range Q1–Q3]; binary or categorical variables as Count (%).

		MIMIC III			eICU			P-Value
		Overall	Alive (HOSPITAL)	Dead (HOSPITAL)	Overall	Alive (HOSPITAL)	Dead (HOSPITAL)	Kruskal-Wallis/ Chi-squared
TOTAL ICU ADMISSIONS		53103	46509 (87.6)	6594 (12.4)	190935	174214 (91.2)	16721 (8.8)	
AGE, median [Q1,Q3]		65.7 [52.8,77.8]	64.7 [52.0,76.9]	73.4 [60.1,82.8]	64 [52 , 75]	64 [52 , 74]	70 [60 , 79]	<0.001
ICU LOS (days), median [Q1,Q3]		2.1 [1.2,4.2]	2.1 [1.2,4.0]	3.1 [1.3,7.2]	2.30 [1.2, 4.9]	2.21 [1.2,4.6]	3.9 [1.5,8.9]	<0.001
Variables	Norepinehrine, median [Q1,Q3]	0.1 [0.0,0.2]	0.1 [0.0,0.1]	0.2 [0.1,0.3]	0.1 [0.0,0.2]	0.1 [0.0,0.1]	0.1 [0.1,0.3]	<0.001
	Dobutamine, median [Q1,Q3]	3.2 [2.5,5.0]	3.0 [2.5,4.8]	4.5 [2.5,7.5]	3.1 [2.5,5.0]	2.9 [2.4,4.5]	4.3 [2.5,7.2]	<0.001
	Dopamine, median [Q1,Q3]	3.0 [0.2,5.9]	3.0 [0.1,5.0]	4.7 [0.2,9.9]	3.0 [0.1,5.2]	2.7 [0.1,4.8]	4.0 [0.2,9.1]	<0.001
	Phenylephrine, median [Q1,Q3]	0.4 [0.2,0.8]	0.4 [0.2,0.6]	1.0 [0.5,2.1]	0.4 [0.2,0.7]	0.3 [0.2,0.6]	0.9 [0.4,2.0]	<0.001
	Bilirubin, median [Q1,Q3]	0.6 [0.4,1.1]	0.6 [0.4,1.0]	0.8 [0.5,2.2]	0.6 [0.4,1.1]	0.6 [0.4,1.0]	0.8 [0.4,2.1]	<0.001
	Creatinine, median [Q1,Q3]	1.0 [0.7,1.4]	0.9 [0.7,1.3]	1.3 [0.9,2.3]	0.9 [0.7,1.4]	0.9 [0.7,1.3]	1.3 [0.8,2.3]	<0.001
	Glucose, median [Q1,Q3]	130.4 [113.4,154.3]	129.0 [112.7,151.5]	144.0 [121.0,175.9]	128.5 [111.8,151.3]	127.3 [111.2,148.8]	140.4 [117.7,170.7]	<0.001
	Lactate, median [Q1,Q3]	1.8 [1.3,2.5]	1.7 [1.3,2.4]	2.3 [1.6,4.1]	1.7 [1.3,2.4]	1.6 [1.2,2.2]	2.2 [1.5,3.9]	<0.001
	Pao2/Fio2, median [Q1,Q3]	248.4 [170.1,350.0]	256.9 [180.0,355.2]	209.2 [128.1,312.0]	257.4 [178.8,358.0]	265.6 [188.1,364.1]	217.2 [137.5,321.9]	<0.001
	Gcs Score, median [Q1,Q3]	14.2 [11.4,15.0]	14.5 [12.4,15.0]	9.2 [5.9,13.9]	14.2 [11.5,15.0]	14.5 [12.4,15.0]	9.5 [5.9,13.9]	<0.001
	Diastolic Blood, median [Q1,Q3]	60.6 [54.1,68.2]	61.0 [54.6,68.6]	57.2 [50.1,65.2]	59.8 [53.2,67.5]	60.3 [53.9,67.9]	55.8 [48.3,63.6]	<0.001
	Systolic Blood, median [Q1,Q3]	119.3 [109.3,131.9]	120.1 [110.2,132.3]	112.5 [101.0,127.3]	118.0 [108.0,130.6]	118.9 [109.1,131.2]	110.2 [98.7,125.2]	<0.001
	Heart Rate, median [Q1,Q3]	84.4 [74.6,95.0]	83.9 [74.4,94.0]	89.5 [77.4,102.6]	83.8 [74.0,94.2]	83.4 [74.0,93.5]	87.2 [74.6,100.2]	<0.001
	Respiration Rate, median [Q1,Q3]	19.1 [16.8,21.8]	18.9 [16.7,21.5]	20.8 [17.7,24.4]	18.8 [16.5,21.5]	18.6 [16.4,21.2]	20.2 [17.1,23.8]	<0.001
GENDER, n (%)	F	23185 (43.7)	20179 (43.4)	3006 (45.6)	86761 (45.0)	79091 (45.4)	7670 (46.0)	<0.001
	M	29918 (56.3)	26330 (56.6)	3588 (54.4)	104174 (55.0)	95123 (54.6)	9051 (54.0)	<0.001
ADMISSION TYPE , n (%)	Medical	32447 (61.1)	27591 (59.3)	4856 (73.6)	16592 (8.7)	14483 (8.3)	2109 (12.6)	<0.001
	ScheduledSurgical	6266 (11.8)	6074 (13.0)	192 (2.9)	68349 (35.8)	61575 (35.3)	6774 (40.5)	<0.001
	UnscheduledSurgical	14390 (27.1)	12844 (27.7)	1546 (23.5)	105994 (55.5)	98156 (56.4)	7838 (46.9)	<0.001
Other Outcomes								
ICU, n (%)	Alive	48384 (91.1)			180768 (94.7)			<0.001
	Dead	4719 (8.9)			10167 (5.3)			<0.001
2-DAYS ICU, n (%)	Alive	51207 (96.4)						
	Dead	1896 (3.6)						
3-DAYS ICU, n (%)	Alive	50639 (95.4)						
	Dead	2464 (4.6)						
30-DAYS ICU, n (%)	Alive	44652 (84.1)						
	Dead	8451 (15.9)						
1-YEAR ICU, n (%)	Alive	37640 (70.9)						
	Dead	15463 (29.1)						

The extracted benchmark datasets are processed to obtain the features that will be used for the prediction tasks. We collect 5 static demographic features (Age, Gender, Admission type, Comorbidity, Ethnicity), which are fully observed and 103 physiological variables including (Labs, Vitals & Drug) measurements that vary over time for each patient stay. Categorical variables (Gender, Admission type, Comorbidity, Ethnicity) are addressed using one-hot encoding.

We chose two sets of features, as described below, to allow for an exhaustive bench-

marking comparison study.

- Feature Set A: This feature set consists of the 15 features used in the SAPS-II score calculation [85]. We have merged the data for each feature based on medical knowledge. For example, to calculate the Glasgow Coma Scale score, we add the GCSVerbal, GCSMotor, and GCSEyes values; to calculate urine output, we sum the features representing urine output, for body temperature, we convert Fahrenheit to Celsius scale and also calculate PaO2/FiO2 ratio instead of considering them as individual features.

- Feature Set B: This feature set consists of 103 features and includes the feature set A, as outlined in Table 7.20a and Table 7.21a.

7.1.1 Handling irregular time intervals

Physiological variables are often measured at irregular intervals, posing a challenge for machine learning techniques that are not specifically designed for time-series data. While some methods can be adapted for streaming data, they usually assume regular time sampling. To overcome this limitation and improve the applicability to irregularly sampled data, we transformed each feature into hourly data for the first 24 and 48 hours following ICU admission. Each time series feature from a single ICU stay is aggregated into hourly intervals (e.g., 0–1 hr, 1–2 hr).

Occasionally, a feature is measured multiple times within a single hour. In these rare cases, the measured values are aggregated. The choice of aggregation function depends on each feature and is determined by a medical expert. For example, urine output is summed, while Mean Arterial Pressure, GCSVerbal, GCSMotor, and GCSEyes use the minimum values. Other features use either minimum or maximum values based on their clinical relevance.

In instances without observations, missing values are filled with the nearest available observation. As suggested by clinicians, missing values are imputed with the last value within a fixed-length forward time window, and remaining gaps are filled with feature-wise median values. This approach ensures a more robust and complete dataset for analysis. During hourly data sampling, it was observed that 103 variables, including both laboratory and vital signs, exhibited an absence exceeding 93% for all patients. This can be attributed to the nature of medical testing, where each patient undergoes only a few tests based on their specific needs, and these tests are often scheduled sporadically over a considerable duration. Refer to (Fig. 3.1 and Fig. 3.2) for further details on this phenomenon.

7.1.2 Prediction tasks

This section outlines prediction tasks that address crucial concerns in critical care research. Among the most significant tasks are predicting mortality rates and the length of hospital stays, which have garnered considerable attention in the medical community [129, 133].

7.1.2.1 Mortalities tasks

This task is formulated as a binary classification problem, where the label represents the occurrence of a death event. Based on this, we propose the following tasks for mortality prediction as studied in [129].

1. **In-hospital mortality:** The aim of this task is to predict in-hospital mortality, which involves determining whether a patient will survive his hospitalization period or not.
2. **In-ICU mortality,** Predict whether the patient dies during his stay after being admitted to an ICU.
3. **Short-term mortality:** The objective of this task is to forecast the likelihood of a patient's death within a short period after his admission to the ICU. In this regard, we establish two distinct mortality prediction tasks: the 2-days and 3-days mortality tasks, where the patient passes away within two or three days of his ICU admission, respectively. To accomplish this, we use data from the first 24 hours of ICU admission to predict both 2-days and 3-days mortality. However, for 2-days mortality we only use the first 24 hours ICU data.
4. **Long-term mortality:** The objective of this task is to predict the likelihood of a patient's death long after they have been discharged from the hospital. Our prediction model focuses on two distinct timeframes, namely, 30-days and 1-year mortality. These timeframes encompass situations where the patient passes away within 30 days or a year after being discharged from the hospital. It is worth noting that we utilize only the first 24-hour data and the first 48-hour data to predict both the 30-days and 1-year mortality.

7.1.2.2 Length of stay task

In hospitals, the Length of Stay (LOS) for a patient serves as a crucial metric for assessing the severity of illness, planning care, and allocating resources [73]. We approach this task as a regression problem.

7.2 Hyperparameters & Evaluation metrics

This section discusses the selection and significance of hyperparameters, as well as the evaluation metrics used to assess model performance.

7.2.1 Model training

We utilize a linear layer for generating logits tailored to the task. Depending on the task type, various layer types are employed to achieve the desired outputs, with corresponding loss functions computed as follows:

- **Binary classification**

$$Loss(y_t, \hat{y}_t) = -(y_t \log(\hat{y}_t) + (1 - y_t) \log(1 - \hat{y}_t)) \tag{7.1}$$

  where y and $\hat{y}_t$ are the true and predicted labels.

- **Regression:** Suppose $(\boldsymbol{p}_t)_i$ is the predicted value and the ground truth value is $(\boldsymbol{y}_t)_i$ for stay $\boldsymbol{s}$, then MSE is computed as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \{(\boldsymbol{p}_t)_i - (\boldsymbol{y}_t)_i\}^2 \tag{7.2}$$

Our cohort consists of ICU patient data from the first 24 and 48 hours, which remains consistent across all tasks. Regardless of the architecture used, the batch size for classification tasks is fixed at 64, while for regression tasks, it is fixed at 128 during training. We utilize the Adam optimizer with a learning rate of 10^{-3} and a decay rate of 10^{-4}. The models are trained for 500 epochs, with early stopping implemented to prevent overfitting. To further mitigate overfitting, we incorporate a dropout rate of 0.3 during the training phase.

7.2.2 Evaluation metrics

We employ a 10-fold cross-validation methodology to evaluate the model's performance across various prediction tasks. In this approach, the dataset is randomly divided into ten folds, each comprising 10% of the data for assessing model performance (test set), while the remaining 90% serves as training data. During the training phase, 80% of the training data is utilized for model training, and the remaining 20% serves as the validation set to optimize the training process.

All performance metrics are calculated based on the test dataset. For regression tasks, we report Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Adjusted R^2. In classification tasks, we report the mean and standard error of performance scores across all ten (10) testing folds, including metrics such as AUROC (Area Under the Receiver Operating Characteristics), AUPRC (Area Under the Precision-Recall Curve), and F1-score.

7.2.2.1 Evaluating classifiers

A machine learning model's success is defined by measurements of the model's performance. Thus, there are several different ways of measuring the performance to see if your model is able to learn the patterns in the data. This measure is called a performance metric, and the choice of this performance metric is dependent on the target task the model is trying to predict.

Precision (Positive Predictive Value) measures the percentage of correctly identified positive samples, indicating the probability that a positive test result is accurate.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7.3}$$

Recall (Sensitivity or True Positive Rate) measures the proportion of actual positives correctly identified, reflecting the model's ability to find all positive samples.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7.4}$$

To evaluate a model's effectiveness, both precision and recall must be considered, as improving one often reduces the other, creating a trade-off. The F1-score, the harmonic mean of precision and recall, balances these metrics, making it suitable for tasks with unbalanced classes by accounting for both false positives and false negatives.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7.5}$$

7.3 Results analysis

In this section, we report the results of all the prediction algorithms on the MIMIC-III datasets and also the performance of prediction methods on the different feature sets used. We report the mean and standard deviation of AUROC, AUPRC, and F1-Score per model across all 10-testing folds for all mortalities tasks, while for length-of-stay (LOS) we report the RMSE, MAE and Adjusted R^2.

7.3.1 MWTA-LSTM

This subsection presents the results of the MWTA-LSTM model across various prediction tasks on the MIMIC-III and eICU datasets. It also includes a discussion of the diverse outcomes observed.

7.3.1.1 Mortalities tasks

In the domain of In-hospital and ICU Mortality Prediction, the outcomes are detailed in Table 7.2 and Table 7.3, presenting the results of diverse prediction algorithms applied to different Feature Sets of the MIMIC-III dataset for both 24-hour and 48-hour data. A comprehensive analysis of these tables reveals that our proposed framework consistently surpasses all other models in terms of F1-score for both time periods, demonstrating a substantial performance difference. These findings yield several noteworthy observations. Firstly, irrespective of the feature set employed, our architecture consistently achieves a higher F1-score compared to all other models. Moreover, the F1-score demonstrates a significant increase as more features are included as input variables, underscoring the effectiveness of our framework in leveraging additional information to enhance predictive performance. Secondly, with a larger dataset utilized over a 48-hour observation period, our model exhibits substantial performance improvements across various prediction tasks compared to state-of-the-art models, achieving an F1-score exceeding 10 %. This suggests that a prolonged

Table 7.2: In-hospital & ICU Mortality using 24 HRS DATA, MP denotes Max Pooling with (MWTA).

24 HOURS DATA							
TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	LSTM	0.88±0.0125	0.60±0.0275	0.56±0.0235	0.88±0.0140	0.61±0.0110	0.57±0.0235
	LSTM-WΔt	0.89±0.0034	0.65±0.0023	0.57±0.0135	0.90±0.0132	0.63±0.0161	0.58±0.0143
	Time LSTM[176]	0.88±0.0030	0.63±0.0020	0.57±0.0040	0.88±0.0041	0.61±0.0110	0.57±0.0090
	TLSTM[5]	0.88±0.0011	0.64±0.0006	0.57±0.0021	0.88±0.0059	0.62±0.0114	0.58±0.0085
	RETAIN[24]	0.86±0.0022	0.58±0.0082	0.54±0.0002	0.87±0.0002	0.58±0.0075	0.54±0.0017
	ATTAIN[171]	0.89±0.0050	0.65±0.0134	0.57±0.0084	0.90±0.0096	0.66±0.0215	0.60±0.0121
	nTAND[144]	0.86±0.0064	0.56±0.0182	0.53±0.0110	0.90±0.0064	0.68±0.0243	0.61±0.0152
	HiTANet[102]	0.88±0.0222	0.66±0.0382	0.60±0.0207	0.91±0.0502	0.68±0.0085	0.61±0.0317
	ConCare[104]	0.88±0.0732	0.63±0.0375	0.56±0.0512	0.90±0.0132	0.66±0.0121	0.60±0.0617
	MWTA-LSTM-1 MP	0.91±0.0025	0.69±0.0047	0.62±0.0014	0.92±0.0013	0.72±0.0001	0.64±0.0005
	MWTA-LSTM-2 MP	0.91±0.0047	0.70±0.0058	0.62±0.0084	0.93±0.0002	0.76±0.0003	0.67±0.0021
	MWTA-LSTM-2 ASP	0.90±0.0019	0.70±0.0010	**0.62±0.0001**	0.93±0.0018	0.76±0.0001	**0.69±0.0123**
ICU	LSTM	0.90±0.0254	0.60±0.0137	0.54±0.0250	0.92±0.0040	0.63±0.0080	0.57±0.0060
	LSTM-WΔt	0.92±0.0012	0.63±0.0037	0.58±0.0169	0.92±0.0060	0.65±0.0052	0.60±0.0040
	Time LSTM[176]	0.91±0.0011	0.64±0.0010	0.58±0.0050	0.91±0.0010	0.62±0.0040	0.58±0.0080
	TLSTM[5]	0.91±0.0020	0.64±0.0004	0.57±0.0005	0.92±0.0033	0.64±0.0072	0.59±0.0022
	RETAIN[24]	0.89±0.0032	0.58±0.0080	0.54±0.0098	0.90±0.0016	0.59±0.0013	0.55±0.0002
	ATTAIN[171]	0.90±0.0104	0.63±0.0234	0.56±0.0191	0.92±0.0073	0.64±0.0220	0.61±0.0201
	nTAND[144]	0.88±0.0076	0.55±0.0224	0.53±0.0186	0.90±0.0112	0.60±0.0251	0.58±0.0237
	HiTANet[102]	0.90±0.0022	0.60±0.0402	0.57±0.0102	0.91±0.0122	0.65±0.0325	0.60±0.0217
	ConCare[104]	0.90±0.0128	0.59±0.0352	0.56±0.0392	0.90±0.0087	0.60±0.0371	0.58±0.0217
	MWTA-LSTM-1 MP	0.93±0.0038	0.69±0.0131	0.62±0.0172	0.94±0.0001	0.72±0.0074	0.64±0.0074
	MWTA-LSTM-2 MP	0.93±0.0012	0.72±0.0034	0.64±0.0103	0.95±0.0005	0.75±0.0054	0.68±0.0004
	MWTA-LSTM-2 ASP	0.93±0.0041	0.73±0.0035	**0.64±0.0078**	0.96±0.0028	0.80±0.0076	**0.71±0.0076**

Table 7.3: In-hospital & ICU Mortality using 48 HRS DATA with (MWTA).

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	LSTM	0.90±0.0345	0.67±0.0205	0.59±0.0076	0.92±0.0303	0.69±0.0350	0.61±0.0451
	LSTM-WΔt	0.92±0.0270	0.69±0.0523	0.60±0.0212	0.93±0.0181	0.72±0.0330	0.62±0.0261
	Time LSTM[176]	0.91±0.0030	0.71±0.0034	0.63±0.0127	0.91±0.0050	0.71±0.0090	0.62±0.0098
	TLSTM[5]	0.91±0.0050	0.71±0.0033	0.63±0.0021	0.91±0.0046	0.71±0.0054	0.62±0.0041
	RETAIN[24]	0.89±0.0028	0.67±0.0032	0.60±0.0086	0.91±0.0052	0.67±0.0041	0.60±0.0079
	ATTAIN[171]	0.90±0.0075	0.68±0.0150	0.61±0.0125	0.92±0.0076	0.71±0.0226	0.63±0.0152
	nTAND[144]	0.89±0.0082	0.65±0.0103	0.59±0.0104	0.91±0.0057	0.70±0.0145	0.63±0.0121
	HiTANet[102]	0.89±0.0225	0.64±0.0332	0.60±0.0902	0.90±0.0252	0.70±0.0375	0.62±0.0027
	ConCare[104]	0.90±0.0322	0.63±0.0542	0.60±0.0632	0.90±0.0672	0.71±0.0375	0.63±0.0201
	MWTA-LSTM-1 MP	0.92±0.0013	0.76±0.0011	0.65±0.0060	0.94±0.0014	0.78±0.0084	0.70±0.0017
	MWTA-LSTM-2 MP	0.94±0.0007	0.78±0.0007	0.69±0.0030	0.94±0.0018	0.80±0.0019	0.72±0.0076
	MWTA-LSTM-2 ASP	0.94±0.0127	0.78±0.0047	**0.70±0.0080**	0.95±0.0036	0.81±0.0029	**0.73±0.0073**
ICU	LSTM	0.92±0.0025	0.67±0.0101	0.58±0.0516	0.93±0.0250	0.69±0.0370	0.60±0.0460
	LSTM-WΔt	0.95±0.0018	0.73±0.0319	0.65±0.0221	0.94±0.0341	0.72±0.0080	0.63±0.0170
	Time LSTM[176]	0.94±0.0001	0.73±0.0017	0.65±0.0002	0.94±0.0055	0.71±0.0081	0.63±0.0054
	TLSTM[5]	0.94±0.0023	0.74±0.0049	0.65±0.0030	0.94±0.0038	0.72±0.0075	0.63±0.0081
	RETAIN[24]	0.91±0.0022	0.65±0.0004	0.60±0.0006	0.92±0.0064	0.68±0.0048	0.60±0.0022
	ATTAIN[171]	0.92±0.0079	0.68±0.0185	0.63±0.0141	0.93±0.0087	0.68±0.0230	0.63±0.0143
	nTAND[144]	0.91±0.0080	0.65±0.0141	0.61±0.0130	0.94±0.0061	0.74±0.0190	0.66±0.0182
	HiTANet[102]	0.92±0.0087	0.69±0.0125	0.63±0.0502	0.93±0.0287	0.70±0.0175	0.63±0.0717
	ConCare[104]	0.92±0.0007	0.68±0.0032	0.62±0.0101	0.93±0.0074	0.69±0.0175	0.61±0.0213
	MWTA-LSTM-1 MP	0.95±0.0007	0.78±0.0035	0.70±0.0048	0.96±0.0018	0.79±0.0016	0.71±0.0072
	MWTA-LSTM-2 MP	0.96±0.0004	0.79±0.0022	0.70±0.0012	0.97±0.0025	0.83±0.0019	0.74±0.0046
	MWTA-LSTM-2 ASP	0.96±0.0044	0.81±0.0122	**0.72±0.0042**	0.97±0.0024	0.85±0.0028	**0.77±0.0055**

Table 7.4: In-hospital & ICU mortality task on eICU dataset using First 24 & 48 hours data with (MWTA).

MORTALITY TASKS							
MODELS	TASKS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
MWTA-LSTM-2 ASP	**In-Hospital**						
	In-hospital - 24HRS	0.91±0.0008	0.62±0.0032	0.57±0.0087	0.91±0.0023	0.68±0.0011	0.63±0.0030
	In-hospital - 48HRS	0.92±0.0008	0.67±0.0132	0.62±0.0237	0.94±0.0073	0.75±0.0131	0.68±0.0070
	ICU						
	ICU - 24HRS	0.93±0.0053	0.65±0.0024	0.61±0.0030	0.93±0.0008	0.71±0.0052	0.67±0.0050
	ICU - 48HRS	0.94±0.0088	0.74±0.0240	0.68±0.0251	0.96±0.0068	0.80±0.0040	0.73±0.0423

observation period, coupled with increased data availability, contributes to enhanced predictive capabilities.

Furthermore, our approach demonstrates a 3.2% improvement in the performance of MWTA-LSTM 2 ASP over MWTA-LSTM 2 without the ASP strategy. To ensure the generalizability of our approach, we've evaluated on eICU dataset, the results are shown in Table 7.4 and from the results, we can see that, our model always achieves a remarkable result when more features are used. This improvement can be attributed to the careful modeling of irregular timing, intervention effects, and the handling of outliers, which allows our model to better capture the underlying patterns and variability present in the patient data.

In the domain of Short-term and Long-term Mortality Prediction, Table 7.5 and Table 7.6 showcase the outcomes of these tasks, utilizing diverse feature sets from the MIMIC-III dataset for both 24-hour and 48-hour data. Notably, our framework still consistently achieves top-tier results across AUROC, AUPRC, and F1-score metrics, demonstrating its effectiveness in predicting both short-term and long-term mortality.

More specifically, for short-term mortality (2 and 3 days), our framework stands out with an impressive F1-score exceeding 10 %, while for long-term mortality tasks (30 days and One year) in the 24-hour data, it maintains an F1-score above 7 %. These robust results persist even when expanding the dataset to include the initial 48 hours of patient data.

In summary, The MWTA-LSTM 2 with ASP emerges as the superior model. It demonstrates exceptional performance in handling outliers and accurately capturing central tendencies and variability in patient data when compared to alternative models.

Moreover, we conducted p-value calculations on the testing folds to perform pairwise comparisons among all the models. In addition, the degrees of freedom and alpha are adjusted to 58 and 0.05, respectively. Therefore, there is 95% confidence that the conclusion of test is valid. The results obtained unequivocally demonstrate a significant difference between the models in terms of mortality prediction tasks. Importantly, our proposed framework, **MWTA-LSTM-2** with ASP or without ASP consistently outperform all other models, as evidenced by the averaged F1-score displayed in Fig. 7.1 and Fig. 7.2 for In-hospital & ICU Mortality. Below are two use cases that provide explanations for the figures related to the ICU Mortality task.

The paired samples t-test comparing **ATTAIN** to **MWTA-LSTM-ASP** yielded a highly significant result, with a P-value of 7.7e-09 and a t-statistic of -20, as depicted in Fig. 7.1(a) and Fig. 7.1(b). The remarkably small p-value strongly rejects the null hypothesis, providing robust evidence of a substantial difference between the two models. The negative t-statistic suggests that the mean of **ATTAIN** is significantly lower than that of **MWTA-LSTM-ASP**. In summary, the t-test results affirm a statistically significant distinction between **ATTAIN** and **MWTA-LSTM-ASP**, with the negative t-statistic emphasizing the superiority of **MWTA-LSTM-ASP** over **ATTAIN**.

Additionally, the paired samples t-test comparing **MWTA-LSTM-2 MP** to **MWTA-LSTM-ASP** revealed a highly significant result, with a P-value of 0.00091 and a t-statistic of -4.9, as illustrated in Fig. 7.1(a) and Fig. 7.1(b). The small p-value indicates compelling evidence against the null hypothesis, signifying a substantial difference between the two models. The negative t-statistic further suggests that the mean of **MWTA-LSTM-2 MP** is lower than that of **MWTA-LSTM-ASP**. In summary, the t-test results confirm a statistically significant distinction between **MWTA-LSTM-2** and **MWTA-LSTM-ASP**, providing sub-

Table 7.5: Short-term & Long-term Mortality using 24 HRS DATA with (MWTA).

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
2-DAYS	LSTM	0.96±0.0090	0.70±0.0044	0.65±0.0110	0.96±0.0101	0.72±0.0160	0.67±0.0180
	LSTM-WΔt	0.97±0.0023	0.77±0.0125	0.73±0.0213	0.95±0.0270	0.76±0.0410	0.72±0.0221
	Time LSTM[176]	0.96±0.0128	0.76±0.0114	0.70±0.0087	0.95±0.0038	0.71±0.0008	0.68±0.0088
	TLSTM[5]	0.97±0.0004	0.78±0.0070	0.73±0.0190	0.94±0.0010	0.68±0.0086	0.62±0.0039
	RETAIN[24]	0.94±0.0004	0.64±0.0040	0.59±0.0025	0.94±0.0037	0.67±0.0147	0.63±0.0141
	ATTAIN[171]	0.94±0.0126	0.72±0.0254	0.67±0.0167	0.94±0.0149	0.76±0.0295	0.71±0.0211
	nTAND[144]	0.93±0.0022	0.64±0.0260	0.63±0.0261	0.95±0.0085	0.70±0.0257	0.67±0.0251
	HiTANet[102]	0.94±0.0208	0.70±0.0112	0.66±0.0402	0.95±0.0052	0.75±0.0055	0.68±0.0237
	ConCare[104]	0.94±0.0302	0.70±0.0082	0.65±0.0002	0.95±0.0062	0.73±0.0075	0.69±0.0017
	MWTA-LSTM-1 MP	0.96±0.0101	0.81±0.0100	0.77±0.0255	0.97±0.0031	0.85±0.0090	0.79±0.0146
	MWTA-LSTM-2 MP	0.97±0.0030	0.84±0.0019	0.77±0.0025	0.98±0.0017	0.87±0.0038	0.83±0.0201
	MWTA-LSTM-2 ASP	0.97±0.0037	0.86±0.0049	**0.79±0.0023**	0.98±0.0042	0.90±0.0022	**0.86±0.0087**
3-DAYS	LSTM	0.94±0.0016	0.67±0.0205	0.60±0.0121	0.95±0.0032	0.68±0.0110	0.63±0.0083
	LSTM-WΔt	0.94±0.0024	0.69±0.0375	0.62±0.0141	0.97±0.0080	0.72±0.0730	0.65±0.0217
	Time LSTM[176]	0.93±0.0081	0.68±0.0119	0.64±0.0184	0.93±0.0005	0.64±0.0111	0.61±0.0007
	TLSTM[5]	0.94±0.0044	0.71±0.0022	0.66±0.0124	0.92±0.0015	0.60±0.0005	0.56±0.0056
	RETAIN[24]	0.92±0.0019	0.59±0.0102	0.56±0.0262	0.92±0.0001	0.63±0.0017	0.58±0.0143
	ATTAIN[171]	0.92±0.0135	0.67±0.0392	0.64±0.0369	0.94±0.0087	0.69±0.0273	0.65±0.0305
	nTAND[144]	0.90±0.0154	0.63±0.0300	0.59±0.0233	0.92±0.0146	0.66±0.0242	0.60±0.0210
	HiTANet[102]	0.91±0.0642	0.68±0.0385	0.64±0.0502	0.93±0.0432	0.70±0.0375	0.66±0.0017
	ConCare[104]	0.92±0.0267	0.67±0.0102	0.64±0.0252	0.93±0.0052	0.68±0.0035	0.65±0.0237
	MWTA-LSTM-1 MP	0.96±0.0012	0.75±0.0059	0.70±0.0115	0.96±0.0030	0.79±0.0008	0.75±0.0119
	MWTA-LSTM-2 MP	0.96±0.0027	0.78±0.0033	0.73±0.0061	0.98±0.0048	0.85±0.0016	0.78±0.0161
	MWTA-LSTM-2 ASP	0.96±0.0012	0.80±0.0030	**0.76±0.0080**	0.98±0.0020	0.86±0.0054	**0.80±0.0072**
30-DAYS	LSTM	0.86±0.0016	0.63±0.0050	0.58±0.0060	0.89±0.0042	0.67±0.0080	0.60±0.0105
	LSTM-WΔt	0.89±0.0012	0.68±0.0024	0.61±0.0010	0.89±0.0011	0.68±0.0052	0.61±0.0035
	Time LSTM[176]	0.87±0.0065	0.64±0.0060	0.58±0.0062	0.88±0.0013	0.64±0.0051	0.60±0.0045
	TLSTM[5]	0.88±0.0016	0.64±0.0033	0.59±0.0004	0.87±0.0031	0.64±0.0086	0.60±0.0029
	RETAIN[24]	0.86±0.0010	0.62±0.0011	0.57±0.0019	0.86±0.0014	0.61±0.0014	0.57±0.0029
	ATTAIN[171]	0.84±0.0047	0.66±0.0123	0.58±0.0098	0.89±0.0100	0.67±0.0192	0.61±0.0177
	nTAND[144]	0.85±0.0057	0.60±0.0123	0.56±0.0121	0.88±0.0071	0.66±0.0250	0.60±0.0120
	HiTANet[102]	0.87±0.0374	0.65±0.0262	0.60±0.0052	0.90±0.0502	0.68±0.0135	0.61±0.0417
	ConCare[104]	0.87±0.0542	0.63±0.0048	0.58±0.0033	0.89±0.0371	0.66±0.0218	0.60±0.0073
	MWTA-LSTM-1 MP	0.89±0.0067	0.68±0.0114	0.61±0.0081	0.91±0.0043	0.73±0.0049	0.65±0.0016
	MWTA-LSTM-2 MP	0.90±0.0007	0.69±0.0068	0.62±0.0051	0.93±0.0012	0.77±0.0019	0.69±0.0010
	MWTA-LSTM-2 ASP	0.91±0.0057	0.70±0.0168	0.62±0.0151	0.93±0.0004	0.79±0.0021	**0.70±0.0024**
1-YEAR	LSTM	0.83±0.0050	0.68±0.0040	0.63±0.0100	0.85±0.0010	0.72±0.0030	0.65±0.0040
	LSTM-WΔt	0.84±0.0040	0.70±0.0070	0.64±0.0014	0.84±0.0040	0.71±0.0060	0.65±0.0007
	Time LSTM[176]	0.82±0.0018	0.69±0.0017	0.64±0.0068	0.85±0.0032	0.72±0.0036	0.66±0.0040
	TLSTM[5]	0.84±0.0004	0.71±0.0007	0.64±0.0004	0.83±0.0006	0.69±0.0002	0.64±0.0005
	RETAIN[24]	0.82±0.0072	0.68±0.0085	0.62±0.0107	0.83±0.0033	0.68±0.0054	0.64±0.0011
	ATTAIN[171]	0.83±0.0080	0.60±0.0132	0.60±0.0101	0.84±0.0067	0.71±0.0116	0.65±0.0078
	nTAND[144]	0.83±0.0068	0.67±0.0130	0.63±0.0080	0.85±0.0146	0.68±0.0121	0.63±0.0101
	HiTANet[102]	0.84±0.0331	0.68±0.0044	0.62±0.0202	0.85±0.0411	0.73±0.0234	0.64±0.0037
	ConCare[104]	0.84±0.0642	0.66±0.0044	0.61±0.0308	0.86±0.0574	0.71±0.0370	0.62±0.0017
	MWTA-LSTM-1 MP	0.84±0.0040	0.72±0.0010	0.65±0.0011	0.87±0.0032	0.77±0.0069	0.69±0.0005
	MWTA-LSTM-2 MP	0.84±0.0011	0.72±0.0005	0.65±0.0008	0.88±0.0014	0.78±0.0021	0.70±0.0035
	MWTA-LSTM-2 ASP	0.85±0.0211	0.74±0.0064	**0.66±0.0208**	0.90±0.0025	0.81±0.0004	**0.72±0.0013**

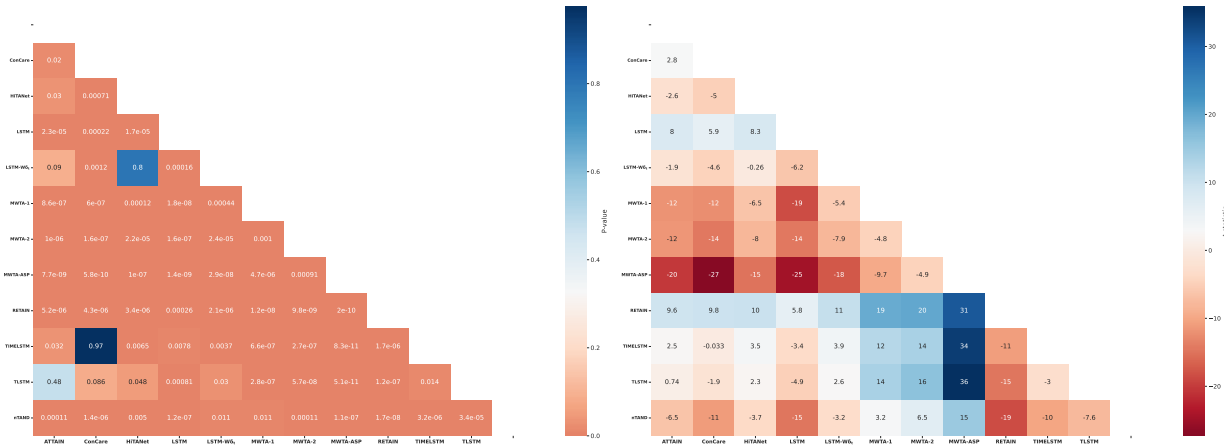
stantial evidence that **MWTA-LSTM-ASP** significantly outperforms the former. Overall, similar analyses were conducted for all other baseline models, yielding consistent results.

7.3.1.2 Length of stay (LOS)

In regards to the Length of Stay (LOS) task, we evaluate the accuracy of predicted values compared to actual values in the test set using two metrics: mean absolute error (MAE) and root mean squared error (RMSE). The minimum, and maximum LOS for the entire population recorded were 1.2 and 173.07 days, respectively. The error values per day, measured by both MAE and RMSE, for all models, are presented in Table 7.7. The best performance was achieved with an MAE of 1.21 days. Among all the models, the MWTA-LSTM 2 ASP model consistently outperforms the others across all metrics, as shown in Table 7.7. Our model

Table 7.6: Short & Long Term Mortality using 48 HRS DATA with (MWTA).

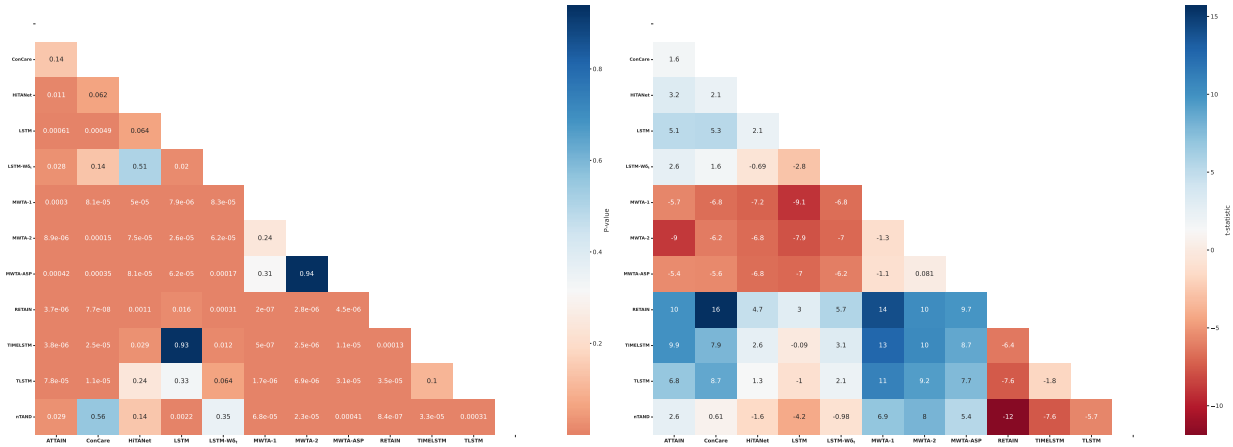
TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
3-DAYS	LSTM	0.96±0.0624	0.73±0.0320	0.70±0.0251	0.96±0.0660	0.77±0.0220	0.73±0.0217
	LSTM-WΔt	0.94±0.0423	0.78±0.0710	0.76±0.0202	0.96±0.0040	0.84±0.0251	0.79±0.0513
	Time LSTM[176]	0.97±0.0047	0.85±0.0019	0.80±0.0013	0.96±0.0005	0.79±0.0032	0.73±0.0046
	TLSTM[5]	0.97±0.0010	0.86±0.0020	0.82±0.0040	0.96±0.0015	0.80±0.0060	0.74±0.0071
	RETAIN[24]	0.92±0.0022	0.73±0.0004	0.70±0.0006	0.95±0.0011	0.77±0.0042	0.73±0.0173
	ATTAIN[171]	0.95±0.0342	0.76±0.0232	0.72±0.0402	0.96±0.0203	0.82±0.0180	0.78±0.0221
	nTAND[144]	0.94±0.0100	0.72±0.0115	0.68±0.0128	0.95±0.0013	0.80±0.0232	0.75±0.0185
	HiTANet[102]	0.95±0.0351	0.80±0.0182	0.77±0.0057	0.97±0.0002	0.84±0.0375	0.80±0.0617
	ConCare[104]	0.95±0.0047	0.78±0.0483	0.76±0.0607	0.96±0.0532	0.84±0.0065	0.79±0.0042
	MWTA-LSTM-1 MP	0.98±0.0018	0.90±0.0029	0.85±0.0036	0.99±0.0043	0.91±0.0057	0.86±0.0011
	MWTA-LSTM-2 MP	0.98±0.0056	0.92±0.0100	0.87±0.0053	0.99±0.0033	0.94±0.0138	0.89±0.0219
	MWTA-LSTM-2 ASP	0.98±0.0086	0.93±0.0122	0.87±0.0153	0.99±0.0011	0.95±0.0064	**0.92±0.0103**
30-DAYS	LSTM	0.87±0.0344	0.68±0.0228	0.59±0.0263	0.89±0.0230	0.70±0.0151	0.60±0.0812
	LSTM-WΔt	0.90±0.0660	0.69±0.0177	0.61±0.0380	0.91±0.0023	0.72±0.0219	0.63±0.0133
	Time LSTM[176]	0.89±0.0003	0.70±0.0042	0.62±0.0001	0.89±0.0014	0.68±0.0029	0.62±0.0004
	TLSTM[5]	0.89±0.0014	0.70±0.0017	0.63±0.0047	0.89±0.0014	0.70±0.0017	0.63±0.0047
	RETAIN[24]	0.87±0.0010	0.65±0.0004	0.60±0.0006	0.88±0.0002	0.67±0.0044	0.61±0.0070
	ATTAIN[171]	0.88±0.0022	0.67±0.0082	0.60±0.0002	0.90±0.0067	0.68±0.0134	0.61±0.0094
	nTAND[144]	0.87±0.0077	0.64±0.0120	0.59±0.0133	0.89±0.0043	0.70±0.0090	0.62±0.0103
	HiTANet[102]	0.88±0.0164	0.70±0.0351	0.60±0.0067	0.90±0.0502	0.73±0.0025	0.62±0.0017
	ConCare[104]	0.88±0.0052	0.70±0.0182	0.61±0.0602	0.89±0.0288	0.72±0.0375	0.62±0.0038
	MWTA-LSTM-1 MP	0.90±0.0001	0.74±0.0050	0.65±0.0057	0.92±0.0012	0.77±0.0016	0.68±0.0036
	MWTA-LSTM-2 MP	0.91±0.0027	0.75±0.0053	0.67±0.0015	0.93±0.0033	0.82±0.0036	0.72±0.0079
	MWTA-LSTM-2 ASP	0.92±0.0047	0.77±0.0083	**0.68±0.0115**	0.94±0.0021	0.82±0.0002	**0.73±0.0036**
1-YEAR	LSTM	0.84±0.0010	0.69±0.0017	0.60±0.0050	0.85±0.0043	0.72±0.0061	0.61±0.0130
	LSTM-WΔt	0.84±0.0051	0.71±0.0057	0.61±0.0170	0.86±0.0017	0.73±0.0002	0.63±0.0012
	Time LSTM[176]	0.84±0.0060	0.71±0.0080	0.64±0.0086	0.83±0.0059	0.70±0.0020	0.65±0.0088
	TLSTM[5]	0.85±0.0032	0.74±0.0060	0.66±0.0065	0.84±0.0067	0.71±0.0076	0.66±0.0090
	RETAIN[24]	0.81±0.0010	0.68±0.0086	0.62±0.0107	0.83±0.0058	0.69±0.0063	0.65±0.0014
	ATTAIN[171]	0.84±0.0070	0.65±0.0112	0.61±0.0097	0.85±0.0051	0.72±0.0078	0.64±0.0074
	nTAND[144]	0.83±0.0053	0.69±0.0082	0.63±0.0070	0.86±0.0047	0.73±0.0835	0.64±0.0058
	HiTANet[102]	0.85±0.0301	0.69±0.0036	0.60±0.0062	0.85±0.0352	0.71±0.0037	0.62±0.0036
	ConCare[104]	0.84±0.0092	0.68±0.0045	0.62±0.0802	0.85±0.0002	0.71±0.0175	0.63±0.0603
	MWTA-LSTM-1 MP	0.85±0.0036	0.74±0.0036	0.66±0.0129	0.88±0.0006	0.78±0.0004	0.70±0.0033
	MWTA-LSTM-2 MP	0.86±0.0016	0.76±0.0049	0.67±0.0052	0.89±0.0009	0.80±0.0001	0.71±0.0075
	MWTA-LSTM-2 ASP	0.86±0.0006	0.77±0.0069	0.67±0.0152	0.89±0.0014	0.80±0.0041	**0.72±0.0015**



(a) P-value (b) t-statistic

Figure 7.1: Pairwise comparisons p-value < 0.05, conducted on F1-scores across the 10 testing folds using Feature Set B over a 48-hour period, reveal the significance of the models in predicting ICU Mortality. A lower p-value indicates greater statistical significance.

demonstrates superior performance compared to state-of-the-art approaches and two other proposed models. Regardless of the features used (set A or B) and the range of observation data utilized as inputs, our model's predictions for RMSE range from 3 to 4 days when compared to the actual data.



(a) P-value (b) t-statistic

Figure 7.2: Pairwise comparisons p-value < 0.05, conducted on F1-scores across the 10 testing folds using Feature Set B over a 48-hour period, reveal the significance of the models in predicting HOSPITAL Mortality. A lower p-value indicates greater statistical significance.

Table 7.7: Length of Stay prediction with (MWTA).

DATA USED	MODELS	SAPS II FEATURES		ALL FEATURES	
		RMSE	MAE	RMSE	MAE
24 HRS DATA	**LSTM**	5.86±0.3364	2.51±0.0986	5.66±0.3359	2.48±0.1226
	LSTM-WΔt	5.70±0.3057	2.35±0.0807	5.56±0.2979	2.33±0.1393
	Time LSTM[176]	5.39±0.1776	2.79±0.0776	4.45±0.1748	2.90±0.0748
	TLSTM[5]	5.82±0.3457	2.52±0.0727	5.80±0.4579	2.53±0.0393
	RETAIN[24]	6.05±0.2282	3.26±0.0691	6.93±0.2317	3.66±0.0878
	ATTAIN[171]	5.75±0.2432	2.86±0.0531	5.80±0.2101	2.82±0.0188
	nTAND[144]	5.85±0.2132	2.92±0.0567	5.88±0.2212	2.80±0.0537
	HiTANet[102]	5.90±0.2663	3.01±0.0231	5.66±0.2248	2.82±0.0673
	ConCare[104]	6.01±0.2312	3.14±0.0582	5.85±0.2254	3.02±0.0677
	MWTA-LSTM-1 MP	4.87±0.1635	1.92±0.0284	4.85±0.1678	1.82±0.0235
	MWTA-LSTM-2 MP	4.73±0.1592	2.12±0.0137	4.79±0.1252	2.12±0.0303
	MWTA-LSTM-2 ASP	**4.57±0.1308**	**2.08±0.0387**	**3.68±0.1712**	**1.57±0.0128**
48 HRS DATA	**LSTM**	5.42±0.2300	2.00±0.1005	5.19±0.3688	2.11±0.1935
	LSTM-WΔt	5.32±0.3294	1.91±0.0916	5.18±0.3615	2.07±0.1364
	Time LSTM[176]	4.86±0.2099	2.11±0.0446	5.07±0.2161	2.44±0.0437
	TLSTM[5]	4.88±0.1529	2.23±0.0442	5.25±0.2331	2.45±0.0723
	RETAIN[24]	6.19±0.2902	3.29±0.0432	6.59±0.5105	3.39±0.0638
	ATTAIN[171]	5.60±0.2033	2.26±0.0251	5.53±0.2413	2.13±0.0547
	nTAND[144]	5.67±0.2034	2.39±0.0298	5.73±0.2263	2.15±0.0772
	HiTANet[102]	5.71±0.2437	2.33±0.0572	5.49±0.2134	2.24±0.0727
	ConCare[104]	5.80±0.2044	2.56±0.0369	5.74±0.2437	2.49±0.0956
	MWTA-LSTM-1 MP	4.70±0.1735	1.88±0.0578	4.68±0.1833	1.75±0.0152
	MWTA-LSTM-2 MP	4.33±0.1835	1.74±0.0378	4.43±0.1632	1.72±0.0352
	MWTA-LSTM-2 ASP	**4.03±0.1722**	**1.37±0.0216**	**3.27±0.1038**	**1.21±0.0186**

When analyzing the performance on eICU and MIMIC-III datasets, we observe a similar pattern of results. However, there are noticeable variations in the metric magnitudes, which can be attributed to the differences in their LOS distributions. Both datasets exhibit a positive skew, but the skew is more severe in the eICU dataset than the MIMIC-III dataset as shown in Fig. 7.3. This skew has a disproportionate impact on the absolute error, captured by the MSE and MAE metrics since the majority of patients spend fewer than 4 days. Consequently, predicting LOS becomes a challenging task due to this pattern.

Considering all the results obtained from the prediction task, we draw the following observations: our architecture performs significantly better when utilizing more features for prediction. This suggests that our model can effectively learn feature representations

Table 7.8: Length of Stay prediction on eICU dataset with (MWTA).

DATA USED	MODELS	SAPS II FEATURES		ALL FEATURES	
		RMSE	MAE	RMSE	MAE
24 HRS DATA	MWTA-LSTM-2 ASP	6.01±0.1536	3.16±0.0632	5.89±0.1926	3.06±0.0342
48 HRS DATA	MWTA-LSTM-2 ASP	5.88±0.1764	2.92±0.0345	5.84±0.1732	2.69±0.0126

from multiple data modalities and effectively handle the irregular timing patterns present in electronic health record (EHR) data.

(a) MIMIC III (b) eICU

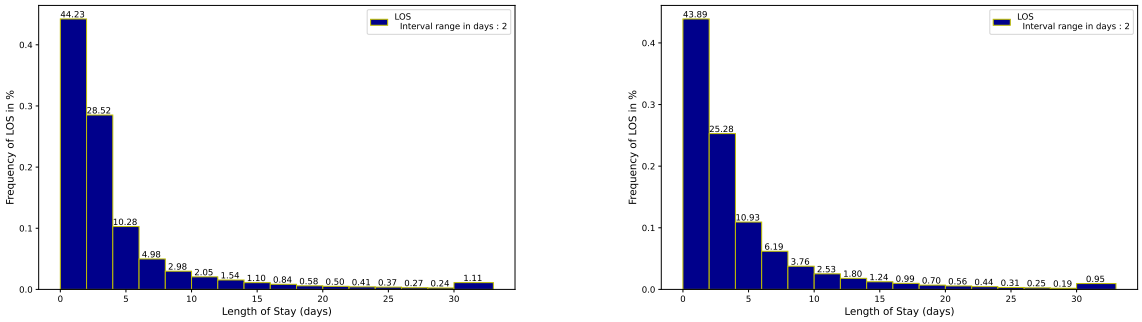


Figure 7.3: The distribution of lengths of stay in days for the two datasets using an interval of 2 days between bars except for the final bar, which contains all lengths of stay between the 30 days till the max LOS

7.3.1.3 Time series forecasting

To ensure its applicability in different settings, thorough evaluations were conducted using three diverse datasets in the field of time series forecasting. The consistent superior performance of MWTA-LSTM over state-of-the-art models, as highlighted in the study by [130], underscores its capabilities. For a comprehensive comparison and further details, please refer to Table 7.9.

The SML 2010 dataset, a publicly available dataset, is utilized for predicting indoor temperatures. It is derived from a monitoring system installed in a residential house. The target series in this dataset is the room temperature, and we have selected 16 relevant driving series that encompass approximately 40 days of monitoring data. The data points were recorded at one-minute intervals and then smoothed using 15-minute averages.

In the NASDAQ 100 Stock dataset [130][1], we have collected the stock prices of 81 major corporations listed under NASDAQ 100, which serve as the driving time series. The target series in this dataset is the index value of the NASDAQ 100. The data was collected on a minute-by-minute basis and covers a period of 105 days from July 26, 2016, to December 22, 2016. Each day encompasses 390 data points representing the market's opening and

[1] https://cseweb.ucsd.edu/~yaq007/NASDAQ100_stock_data.html

Table 7.9: Time series forecasting results with (MWTA).

TIME SERIES FORECASTING							
HORIZON	DATASET	MWTA-LSTM 2 ASP			DA-RNN [130]		
		RMSE	MAE	R2	RMSE	MAE	R2
1	SML 2010	0.041±0.0019	0.030±0.0010	0.94±0.0056	0.038±0.0039	0.027±0.0023	0.95±0.0066
1	NASDAQ 100	0.085±0.0004	0.079±0.0004	0.94±0.0011	0.322±0.0051	0.250± 0.0025	0.89±0.0156
24	IHEPC	0.023±0.0002	0.017±0.0001	0.97±0.0002	0.054±0.0062	0.027±0.0051	0.93±0.0085

closing, except for November 25 (210 data points) and December 22 (which has 180 data points).

The Individual Household Electric Power Consumption (IHEPC) dataset [59] consists of 2.07 million measurements of electric power consumption from a single house located in Sceaux, France (7 km from Paris). The measurements were taken every minute between December 2006 and November 2010, covering a total of 47 months. Our focus in this study is on predicting the "Global active power" parameter. Although there is a small percentage (1.25%) of missing measurements, all the available data points are accompanied by timestamps. To ensure a standardized approach, we have resampled the dataset to have a sampling rate of 1 hour. The prediction task involves forecasting the electric load for the next day, specifically 24 timesteps ahead.

7.3.2 AMITA

This subsection presents the benchmarking results of AMITA on various prediction tasks across both the MIMIC-III and eICU datasets, accompanied by an in-depth discussion of the diverse outcomes observed.

7.3.2.1 Mortalities tasks

In the context of In-hospital and ICU Mortality Prediction, Table 7.10 and Table 7.12 present the results of various prediction algorithms applied to different Feature Sets of the MIMIC-III dataset for both 24-hours and 48-hours data. Upon scrutinizing these tables, it becomes evident that our proposed framework consistently outperforms all other models in terms of F1-score for both time periods, demonstrating a significant advantage. Several notable observations arise from these results. Firstly, irrespective of the feature set employed, our architecture consistently achieves a higher F1-score compared to all other models. Furthermore, the F1-score exhibits a significant increase as more features are included as input variables, highlighting the effectiveness of our framework in leveraging additional information to enhance predictive performance. Secondly, when a larger set of features is utilized over an extended observation period 48-hours, our model exhibits substantial performance improvements across various prediction tasks compared to state-of-the-art models, with an F1-score exceeding 8% with different pooling mechanisms and 3% without for both ICU and hospital mortality tasks. This suggests that a longer observation period, coupled with increased data availability, contributes to enhanced predictive capabilities.

Table 7.10: In-hospital & ICU Mortality using 24 HRS DATA (AMITA).

24 HOURS DATA							
TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	LSTM	0.88±0.0125	0.60±0.0275	0.56±0.0235	0.88±0.0140	0.61±0.0110	0.57±0.0235
	LSTM-WΔt	0.89±0.0034	0.65±0.0023	0.57±0.0135	0.90±0.0132	0.63±0.0161	0.58±0.0143
	Time LSTM[176]	0.88±0.0030	0.63±0.0020	0.57±0.0040	0.88±0.0041	0.61±0.0110	0.57±0.0090
	TLSTM[5]	0.88±0.0011	0.64±0.0006	0.57±0.0021	0.88±0.0059	0.62±0.0114	0.58±0.0085
	RETAIN[24]	0.86±0.0022	0.58±0.0082	0.54±0.0002	0.87±0.0002	0.58±0.0075	0.54±0.0017
	ATTAIN[171]	0.89±0.0050	0.65±0.0134	0.57±0.0084	0.90±0.0096	0.66±0.0215	0.60±0.0121
	nTAND[144]	0.86±0.0064	0.56±0.0182	0.53±0.0110	0.90±0.0064	0.68±0.0243	0.61±0.0152
	HiTANet[102]	0.88±0.0222	0.66±0.0382	0.60±0.0207	0.91±0.0502	0.68±0.0085	0.61±0.0317
	ConCare[104]	0.88±0.0732	0.63±0.0375	0.56±0.0512	0.90±0.0132	0.66±0.0121	0.60±0.0617
	AMITA	0.89±0.0073	0.66±0.0191	0.60±0.0137	0.92±0.0067	0.72±0.0159	0.65±0.0163
	AMITA$_{\text{sum}}$	0.90±0.0015	0.67±0.0030	0.60±0.0046	0.93±0.0002	0.76±0.0003	**0.67±0.0021**
	AMITA$_{\text{max}}$	0.90±0.0007	0.68±0.0029	**0.62±0.0022**	0.93±0.0014	0.75±0.0033	0.67±0.0037
	AMITA$_{\text{mean}}$	0.90±0.0022	0.68±0.0007	0.61±0.0027	0.92±0.0017	0.74±0.0010	0.66±0.0011
	AMITA$_{\text{global mean}}$	0.90±0.0003	0.67±0.0002	0.61±0.0005	0.93±0.0013	0.74±0.0013	0.67±0.0024
ICU	LSTM	0.90±0.0254	0.60±0.0137	0.54±0.0250	0.92±0.0040	0.63±0.0080	0.57±0.0060
	LSTM-WΔt	0.92±0.0012	0.63±0.0037	0.58±0.0169	0.92±0.0060	0.65±0.0052	0.60±0.0040
	Time LSTM[176]	0.91±0.0011	0.64±0.0010	0.58±0.0050	0.91±0.0010	0.62±0.0040	0.58±0.0080
	TLSTM[5]	0.91±0.0020	0.64±0.0004	0.57±0.0005	0.92±0.0033	0.64±0.0072	0.59±0.0022
	RETAIN[24]	0.89±0.0032	0.58±0.0080	0.54±0.0098	0.90±0.0016	0.59±0.0013	0.55±0.0002
	ATTAIN[171]	0.90±0.0104	0.63±0.0234	0.56±0.0191	0.92±0.0073	0.64±0.0220	0.61±0.0201
	nTAND[144]	0.88±0.0076	0.55±0.0224	0.53±0.0186	0.90±0.0112	0.60±0.0251	0.58±0.0237
	HiTANet[102]	0.90±0.0022	0.60±0.0402	0.57±0.0102	0.91±0.0122	0.65±0.0325	0.60±0.0217
	ConCare[104]	0.90±0.0128	0.59±0.0352	0.56±0.0392	0.90±0.0087	0.60±0.0371	0.58±0.0217
	AMITA	0.92±0.0075	0.66±0.0215	0.60±0.0174	0.94±0.0045	0.74±0.0183	0.65±0.0164
	AMITA$_{\text{sum}}$	0.93±0.0021	0.70±0.0049	0.63±0.0035	0.93±0.0002	0.76±0.0003	0.67±0.0021
	AMITA$_{\text{max}}$	0.93±0.0041	0.70±0.0060	0.63±0.0010	0.95±0.0008	0.76±0.0046	0.69±0.0005
	AMITA$_{\text{mean}}$	0.93±0.0027	0.70±0.0042	0.63±0.0039	0.95±0.0023	0.77±0.0055	0.69±0.0017
	AMITA$_{\text{global mean}}$	0.93±0.0030	0.71±0.0044	**0.64±0.0089**	0.95±0.0013	0.77±0.0003	**0.70±0.0019**

Table 7.11: In-hospital & ICU mortality task on eICU dataset using First 24 HRS DATA (AMITA)

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	AMITA	0.90±0.0145	0.60±0.0065	0.56±0.0176	0.92±0.0022	0.65±0.0150	0.60±0.0151
	AMITA$_{\text{sum}}$	0.91±0.0070	0.62±0.0033	0.57±0.0112	0.92±0.0221	0.66±0.0033	0.62±0.0261
	AMITA$_{\text{max}}$	0.91±0.0030	0.63±0.0134	**0.60±0.0127**	0.93±0.0055	0.68±0.0090	**0.63±0.0098**
	AMITA$_{\text{mean}}$	0.91±0.0055	0.62±0.0033	0.58±0.0021	0.92±0.0146	0.67±0.0054	0.62±0.0041
	AMITA$_{\text{global mean}}$	0.91±0.0019	0.62±0.0015	0.58±0.0141	0.93±0.0266	0.67±0.0215	0.62±0.0121
ICU	AMITA	0.92±0.0022	0.64±0.0151	0.60±0.0516	0.93±0.0050	0.68±0.0170	0.64±0.0260
	AMITA$_{\text{sum}}$	0.93±0.0118	0.67±0.0119	0.62±0.0222	0.95±0.0041	0.70±0.0150	0.66±0.0071
	AMITA$_{\text{max}}$	0.93±0.0211	0.68±0.0217	**0.63±0.0102**	0.95±0.0055	0.71±0.0081	0.67±0.0154
	AMITA$_{\text{mean}}$	0.93±0.0123	0.67±0.0049	0.63±0.0130	0.95±0.0038	0.70±0.0075	0.66±0.0181
	AMITA$_{\text{global mean}}$	0.93±0.0219	0.67±0.0110	0.63±0.0251	0.95±0.0096	0.71±0.0215	**0.67±0.0121**

Moreover, our approach with different pooling mechanisms gains a slight improvement from 2-3% in performance over AMITA. To ensure the generalizability of our approach, we've evaluated it on the eICU dataset, and the results are shown in Table 7.11 and Table 7.13. From the results, it is evident that our model consistently achieves remarkable results when more features are used. This improvement can be attributed to the careful modeling of irregular timing and intervention effects, allowing our model to better capture the underlying patterns and variability present in the patient data.

Shifting the focus to Short-term and Long-term Mortality Prediction, the results showcased in Table 7.14 and Table 7.15 unveil the outcomes of our mortality prediction task, employing diverse feature sets from the MIMIC-III dataset for both 24-hours and 48-hours intervals. Significantly, our framework consistently achieves superior results, measured by AUROC, AUPRC, and F1-score.

In the context of short-term mortality (2 and 3 days), our framework excels with an F1-score exceeding 8%, while for long-term mortality tasks (30 days and One year) in the 24-hours data, it maintains an F1-score above 4%. These results are consistently observed

Table 7.12: In-hospital & ICU Mortality using 48 HRS DATA (AMITA)

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	**LSTM**	0.90±0.0345	0.67±0.0205	0.59±0.0076	0.92±0.0303	0.69±0.0350	0.61±0.0451
	LSTM-WΔt	0.92±0.0270	0.69±0.0523	0.60±0.0212	0.93±0.0181	0.72±0.0330	0.62±0.0261
	Time LSTM[176]	0.91±0.0030	0.71±0.0034	0.63±0.0127	0.91±0.0050	0.71±0.0090	0.62±0.0098
	TLSTM[5]	0.91±0.0050	0.71±0.0033	0.63±0.0021	0.91±0.0046	0.71±0.0054	0.62±0.0041
	RETAIN[24]	0.89±0.0028	0.67±0.0032	0.60±0.0086	0.91±0.0052	0.67±0.0041	0.60±0.0079
	ATTAIN[171]	0.90±0.0075	0.68±0.0150	0.61±0.0125	0.92±0.0076	0.71±0.0226	0.63±0.0152
	nTAND[144]	0.89±0.0082	0.65±0.0103	0.59±0.0104	0.91±0.0057	0.70±0.0145	0.63±0.0121
	HiTANet[102]	0.89±0.0225	0.64±0.0332	0.60±0.0902	0.90±0.0252	0.70±0.0375	0.62±0.0027
	ConCare[104]	0.90±0.0322	0.63±0.0542	0.60±0.0632	0.90±0.0672	0.71±0.0375	0.63±0.0201
	AMITA	0.92±0.0225	0.73±0.0147	0.65±0.0014	0.93±0.0066	0.78±0.0132	0.67±0.0122
	AMITA$_{\text{sum}}$	0.93±0.0047	0.77±0.0056	0.67±0.0084	0.93±0.0002	0.79±0.0123	0.69±0.0021
	AMITA$_{\text{max}}$	0.94±0.0029	0.78±0.0023	**0.68±0.0401**	0.94±0.0017	0.80±0.0015	**0.71±0.0029**
	AMITA$_{\text{mean}}$	0.94±0.0147	0.77±0.0038	0.67±0.0084	0.94±0.0002	0.79±0.0003	0.70±0.0141
	AMITA$_{\text{global mean}}$	0.94±0.0119	0.77±0.0012	0.67±0.0201	0.95±0.0096	0.79±0.0215	0.70±0.0131
ICU	**LSTM**	0.92±0.0025	0.67±0.0101	0.58±0.0516	0.93±0.0250	0.69±0.0370	0.60±0.0460
	LSTM-WΔt	0.95±0.0018	0.73±0.0319	0.65±0.0221	0.94±0.0341	0.72±0.0080	0.63±0.0170
	Time LSTM[176]	0.94±0.0001	0.73±0.0017	0.65±0.0002	0.94±0.0055	0.71±0.0081	0.63±0.0054
	TLSTM[5]	0.94±0.0023	0.74±0.0049	0.65±0.0030	0.94±0.0038	0.72±0.0075	0.63±0.0081
	RETAIN[24]	0.91±0.0022	0.65±0.0004	0.60±0.0006	0.92±0.0064	0.68±0.0048	0.60±0.0022
	ATTAIN[171]	0.92±0.0079	0.68±0.0185	0.63±0.0141	0.93±0.0087	0.68±0.0230	0.63±0.0143
	nTAND[144]	0.91±0.0080	0.65±0.0141	0.61±0.0130	0.94±0.0061	0.74±0.0190	0.66±0.0182
	HiTANet[102]	0.92±0.0087	0.69±0.0125	0.63±0.0502	0.93±0.0287	0.70±0.0175	0.63±0.0717
	ConCare[104]	0.92±0.0007	0.68±0.0032	0.62±0.0101	0.93±0.0074	0.69±0.0175	0.61±0.0213
	AMITA	0.94±0.0155	0.75±0.0147	0.67±0.0114	0.95±0.0042	0.77±0.0175	0.70±0.0181
	AMITA$_{\text{sum}}$	0.95±0.0237	0.77±0.0258	0.68±0.0184	0.96±0.0102	0.80±0.0623	0.71±0.0261
	AMITA$_{\text{max}}$	0.95±0.0119	0.78±0.0014	0.69±0.0201	0.97±0.0004	0.82±0.0010	**0.74±0.0019**
	AMITA$_{\text{mean}}$	0.95±0.0447	0.78±0.0258	0.68±0.0084	0.97±0.0002	0.80±0.0033	0.72±0.0351
	AMITA$_{\text{global mean}}$	0.96±0.0005	0.78±0.0331	**0.69±0.0131**	0.97±0.0096	0.81±0.0215	0.73±0.0071

Table 7.13: In-hospital & ICU mortality task on eICU dataset using First 48 HRS DATA (AMITA)

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
In-Hospital	**AMITA**	0.91±0.0215	0.64±0.0065	0.61±0.0085	0.93±0.0311	0.70±0.0052	0.65±0.0151
	AMITA$_{\text{sum}}$	0.93±0.0120	0.67±0.0113	0.63±0.0012	0.93±0.0331	0.71±0.0033	0.66±0.0261
	AMITA$_{\text{max}}$	0.93±0.0144	0.68±0.0094	**0.65±0.0177**	0.93±0.0252	0.72±0.0098	0.67±0.0168
	AMITA$_{\text{mean}}$	0.93±0.0150	0.67±0.0049	0.64±0.0267	0.93±0.0266	0.71±0.0104	0.67±0.0271
	AMITA$_{\text{global mean}}$	0.93±0.0119	0.68±0.0077	0.65±0.0528	0.93±0.0396	0.72±0.0215	**0.67±0.0133**
ICU	**AMITA**	0.93±0.0270	0.68±0.0251	0.64±0.0136	0.95±0.0057	0.72±0.0170	0.67±0.0260
	AMITA$_{\text{sum}}$	0.94±0.0218	0.70±0.0119	0.67±0.0121	0.95±0.0341	0.75±0.0150	0.69±0.0077
	AMITA$_{\text{max}}$	0.94±0.0071	0.71±0.0017	0.68±0.0102	0.95±0.0079	0.77±0.0086	**0.71±0.0153**
	AMITA$_{\text{mean}}$	0.94±0.0093	0.71±0.0339	0.68±0.0137	0.95±0.0088	0.76±0.0077	0.70±0.0281
	AMITA$_{\text{global mean}}$	0.95±0.0219	0.71±0.0010	**0.68±0.0071**	0.95±0.0096	0.77±0.0215	0.70±0.0121

when expanding the dataset to include the first 48 hours of patient data.

Drawing parallels to the scenario of In-hospital and ICU Mortality Prediction, our approach, **AMITA** and its variants incorporating pooling mechanisms outperform all other prediction algorithms in both short-term and long-term mortality prediction tasks.

Furthermore, we conducted p-value calculations on the testing folds to perform pairwise comparisons among all the models. In addition, the degrees of freedom and alpha are adjusted to 58 and 0.05, respectively. Therefore, there is 95% confidence that the conclusion of test is valid. The results obtained unequivocally demonstrate a significant difference between the models in terms of mortality prediction tasks. Importantly, our proposed framework, **AMITA** and its other pooling versions consistently outperform all other models, as evidenced by the averaged F1-score displayed in Fig. 7.4 and Fig. 7.5 for ICU Mortality & In-hospital tasks. Two use cases, elucidating the content of the respective figures, are provided below for the ICU Mortality task:

Comparison between **AMITA** vs **AMITA$_{\text{global mean}}$** using a paired samples t-test yielded

Table 7.14: Short-term & Long-term Mortality using 24 HRS DATA (AMITA).

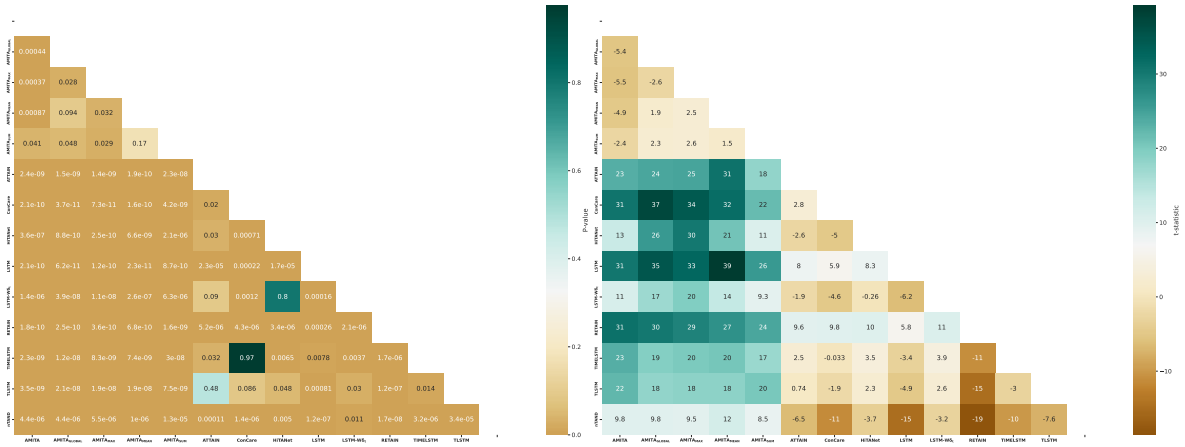
TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
2-DAYS	**LSTM**	0.96±0.0090	0.70±0.0044	0.65±0.0110	0.96±0.0101	0.72±0.0160	0.67±0.0180
	LSTM-WΔt	0.97±0.0023	0.77±0.0125	0.73±0.0213	0.95±0.0270	0.76±0.0410	0.72±0.0221
	Time LSTM[176]	0.96±0.0128	0.76±0.0114	0.70±0.0087	0.95±0.0038	0.71±0.0008	0.68±0.0088
	TLSTM[5]	0.97±0.0004	0.78±0.0070	0.73±0.0190	0.94±0.0010	0.68±0.0086	0.62±0.0039
	RETAIN[24]	0.94±0.0004	0.64±0.0040	0.59±0.0025	0.94±0.0037	0.67±0.0147	0.63±0.0141
	ATTAIN[171]	0.94±0.0126	0.72±0.0254	0.67±0.0167	0.94±0.0149	0.76±0.0295	0.71±0.0211
	nTAND[144]	0.93±0.0022	0.64±0.0260	0.63±0.0261	0.95±0.0085	0.70±0.0257	0.67±0.0251
	HiTANet[102]	0.94±0.0208	0.70±0.0112	0.66±0.0402	0.95±0.0052	0.75±0.0055	0.68±0.0237
	ConCare[104]	0.94±0.0302	0.70±0.0082	0.65±0.0002	0.95±0.0062	0.73±0.0075	0.69±0.0017
	AMITA	0.97±0.0062	0.81±0.0244	0.76±0.0241	0.98±0.0051	0.88±0.0290	0.82±0.0315
	AMITA$_{\text{sum}}$	0.98±0.0004	0.83±0.0036	0.79±0.0019	0.98±0.0022	0.88±0.0073	0.84±0.0121
	AMITA$_{\text{max}}$	0.98±0.0019	0.84±0.0013	0.79±0.0051	0.98±0.0036	0.90±0.0002	0.84±0.0001
	AMITA$_{\text{mean}}$	0.97±0.0019	0.83±0.0073	0.78±0.0006	0.98±0.0052	0.89±0.0063	0.85±0.0021
	AMITA$_{\text{global mean}}$	0.98±0.0019	0.84±0.0013	**0.79±0.0001**	0.98±0.0020	0.90±0.0001	**0.85±0.0086**
3-DAYS	**LSTM**	0.94±0.0016	0.67±0.0205	0.60±0.0121	0.95±0.0032	0.68±0.0110	0.63±0.0083
	LSTM-WΔt	0.94±0.0024	0.69±0.0375	0.62±0.0141	0.97±0.0080	0.72±0.0730	0.65±0.0217
	Time LSTM[176]	0.93±0.0081	0.68±0.0119	0.64±0.0184	0.93±0.0005	0.64±0.0111	0.61±0.0007
	TLSTM[5]	0.94±0.0044	0.71±0.0022	0.66±0.0124	0.92±0.0015	0.60±0.0005	0.56±0.0056
	RETAIN[24]	0.92±0.0019	0.59±0.0102	0.56±0.0262	0.92±0.0001	0.63±0.0017	0.58±0.0143
	ATTAIN[171]	0.92±0.0135	0.67±0.0392	0.64±0.0369	0.94±0.0087	0.69±0.0273	0.65±0.0305
	nTAND[144]	0.90±0.0154	0.63±0.0300	0.59±0.0233	0.92±0.0146	0.66±0.0242	0.60±0.0210
	HiTANet[102]	0.91±0.0642	0.68±0.0385	0.64±0.0502	0.93±0.0432	0.70±0.0375	0.66±0.0017
	ConCare[104]	0.92±0.0267	0.67±0.0102	0.64±0.0252	0.93±0.0052	0.68±0.0035	0.65±0.0237
	AMITA	0.95±0.0087	0.76±0.0219	0.70±0.0167	0.97±0.0103	0.81±0.0312	0.75±0.0255
	AMITA$_{\text{sum}}$	0.96±0.0035	0.78±0.0034	0.72±0.0045	0.97±0.0026	0.83±0.0033	0.75±0.0021
	AMITA$_{\text{max}}$	0.96±0.0042	0.79±0.0087	0.73±0.0040	0.98±0.0012	0.84±0.0053	0.77±0.0037
	AMITA$_{\text{mean}}$	0.96±0.0060	0.78±0.0046	0.73±0.0028	0.98±0.0014	0.85±0.0033	0.77±0.0021
	AMITA$_{\text{global mean}}$	0.96±0.0065	0.79±0.0108	**0.74±0.0137**	0.98±0.0020	0.86±0.0011	**0.79±0.0086**
30-DAYS	**LSTM**	0.86±0.0016	0.63±0.0050	0.58±0.0060	0.89±0.0042	0.67±0.0080	0.60±0.0105
	LSTM-WΔt	0.89±0.0012	0.68±0.0024	0.61±0.0010	0.89±0.0011	0.68±0.0052	0.61±0.0035
	Time LSTM[176]	0.87±0.0065	0.64±0.0060	0.58±0.0062	0.88±0.0013	0.64±0.0051	0.60±0.0045
	TLSTM[5]	0.88±0.0016	0.64±0.0033	0.59±0.0004	0.87±0.0031	0.64±0.0086	0.60±0.0029
	RETAIN[24]	0.86±0.0010	0.62±0.0011	0.57±0.0019	0.86±0.0014	0.61±0.0014	0.57±0.0029
	ATTAIN[171]	0.84±0.0047	0.66±0.0123	0.58±0.0098	0.89±0.0100	0.67±0.0192	0.61±0.0177
	nTAND[144]	0.85±0.0057	0.60±0.0123	0.56±0.0121	0.88±0.0071	0.66±0.0250	0.60±0.0120
	HiTANet[102]	0.87±0.0374	0.65±0.0262	0.60±0.0052	0.90±0.0502	0.68±0.0135	0.61±0.0417
	ConCare[104]	0.87±0.0542	0.63±0.0048	0.58±0.0033	0.89±0.0371	0.66±0.0218	0.60±0.0073
	AMITA	0.89±0.0035	0.68±0.0064	0.62±0.0006	0.90±0.0075	0.72±0.0182	0.64±0.0159
	AMITA$_{\text{sum}}$	0.88±0.0041	0.68±0.0073	0.61±0.0019	0.90±0.0012	0.73±0.0033	0.66±0.0021
	AMITA$_{\text{max}}$	0.90±0.0075	0.72±0.0182	**0.64±0.0159**	0.91±0.0045	0.74±0.0055	**0.67±0.0098**
	AMITA$_{\text{mean}}$	0.89±0.0041	0.69±0.0055	0.62±0.0017	0.91±0.0002	0.75±0.0003	0.67±0.0121
	AMITA$_{\text{global mean}}$	0.90±0.0009	0.70±0.0014	0.62±0.0018	0.91±0.0036	0.75±0.0215	0.67±0.0121
1-YEAR	**LSTM**	0.83±0.0050	0.68±0.0040	0.63±0.0100	0.85±0.0010	0.72±0.0030	0.65±0.0040
	LSTM-WΔt	0.84±0.0040	0.70±0.0070	0.64±0.0014	0.84±0.0040	0.71±0.0060	0.65±0.0007
	Time LSTM[176]	0.82±0.0018	0.69±0.0017	0.64±0.0068	0.85±0.0032	0.72±0.0036	0.66±0.0040
	TLSTM[5]	0.84±0.0004	0.71±0.0007	0.64±0.0004	0.83±0.0006	0.69±0.0002	0.64±0.0005
	RETAIN[24]	0.82±0.0072	0.68±0.0085	0.62±0.0107	0.83±0.0033	0.68±0.0054	0.64±0.0011
	ATTAIN[171]	0.83±0.0080	0.60±0.0132	0.60±0.0101	0.84±0.0067	0.71±0.0116	0.65±0.0078
	nTAND[144]	0.83±0.0068	0.67±0.0130	0.63±0.0080	0.85±0.0146	0.68±0.0121	0.63±0.0101
	HiTANet[102]	0.84±0.0331	0.68±0.0044	0.62±0.0202	0.85±0.0411	0.73±0.0234	0.64±0.0037
	ConCare[104]	0.84±0.0642	0.66±0.0044	0.61±0.0308	0.86±0.0574	0.71±0.0370	0.62±0.0017
	AMITA	0.84±0.0050	0.70±0.0083	0.64±0.0094	0.86±0.0055	0.75±0.0111	0.68±0.0100
	AMITA$_{\text{sum}}$	0.83±0.0006	0.70±0.0004	0.64±0.0012	0.87±0.0002	0.76±0.0003	0.68±0.0151
	AMITA$_{\text{max}}$	0.84±0.0019	0.71±0.0001	**0.66±0.0007**	0.87±0.0014	0.76±0.0017	0.69±0.0004
	AMITA$_{\text{mean}}$	0.84±0.0009	0.72±0.0002	0.65±0.0010	0.88±0.0002	0.76±0.0003	0.68±0.0021
	AMITA$_{\text{global mean}}$	0.85±0.0020	0.72±0.0023	0.66±0.0010	0.88±0.0136	0.77±0.0215	**0.70±0.0121**

a highly significant result P-value of 4.4e-04 as shown in Fig. 7.4(a), t-statistic of -5.4e+00 as depicted in Fig. 7.4(b). The small p-value provides robust evidence against the null hypothesis, indicating a substantial difference between the two models. The negative t-statistic further implies that the mean of **AMITA** is significantly lower than that of **AMITA**$_{\text{global mean}}$. In summary, the t-test results affirm a statistically significant distinction between **AMITA** and **AMITA**$_{\text{global mean}}$, with the negative t-statistic underscoring a lower mean in **AMITA**. Therefore, the small p-value reinforces the rejection of the null hypothesis, highlighting the superiority of **AMITA**$_{\text{global mean}}$.

Additionally, the paired samples t-test comparing **AMITA** to **ATTAIN** revealed a highly significant result with a P-value of 2.4e-09, as illustrated in Fig. 7.4(a), accompanied by a t-statistic of 23, as shown in Fig. 7.4(b). The exceptionally small p-value provides compelling evidence against the null hypothesis, signifying a substantial difference between

Table 7.15: Short & Long Term Mortality using 48 HRS DATA (AMITA)

TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
		AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
3-DAYS	**LSTM**	0.96±0.0624	0.73±0.0320	0.70±0.0251	0.96±0.0660	0.77±0.0220	0.73±0.0217
	LSTM-WΔt	0.94±0.0423	0.78±0.0710	0.76±0.0202	0.96±0.0040	0.84±0.0251	0.79±0.0513
	Time LSTM[176]	0.97±0.0047	0.85±0.0019	0.80±0.0013	0.96±0.0005	0.79±0.0032	0.73±0.0046
	TLSTM[5]	0.97±0.0010	0.86±0.0020	0.82±0.0040	0.96±0.0015	0.80±0.0060	0.74±0.0071
	RETAIN[24]	0.92±0.0022	0.73±0.0004	0.70±0.0006	0.95±0.0011	0.77±0.0042	0.73±0.0173
	ATTAIN[171]	0.95±0.0342	0.76±0.0232	0.72±0.0402	0.96±0.0203	0.82±0.0180	0.78±0.0221
	nTAND[144]	0.94±0.0100	0.72±0.0115	0.68±0.0128	0.95±0.0013	0.80±0.0232	0.75±0.0185
	HiTANet[102]	0.95±0.0351	0.80±0.0182	0.77±0.0057	0.97±0.0002	0.84±0.0375	0.80±0.0617
	ConCare[104]	0.95±0.0047	0.78±0.0483	0.76±0.0607	0.96±0.0532	0.84±0.0065	0.79±0.0042
	AMITA	0.97±0.0125	0.86±0.0037	0.83±0.0114	0.97±0.0013	0.88±0.0311	0.85±0.0205
	AMITA$_{\text{sum}}$	0.98±0.0047	0.88±0.0038	0.84±0.0284	0.99±0.0202	0.90±0.0313	0.87±0.0065
	AMITA$_{\text{max}}$	0.98±0.0314	0.90±0.0110	0.85±0.0098	0.99±0.0118	0.92±0.0311	0.89±0.0133
	AMITA$_{\text{mean}}$	0.97±0.0037	0.90±0.0148	0.84±0.0044	0.99±0.0402	0.91±0.0033	0.89±0.0052
	AMITA$_{\text{global mean}}$	0.98±0.0139	0.90±0.0011	**0.85±0.0001**	0.99±0.0096	0.92±0.0215	**0.90±0.0121**
30-DAYS	**LSTM**	0.87±0.0344	0.68±0.0228	0.59±0.0263	0.89±0.0230	0.70±0.0151	0.60±0.0812
	LSTM-WΔt	0.90±0.0660	0.69±0.0177	0.61±0.0380	0.91±0.0023	0.72±0.0219	0.63±0.0133
	Time LSTM[176]	0.89±0.0003	0.70±0.0042	0.62±0.0001	0.89±0.0014	0.68±0.0029	0.62±0.0004
	TLSTM[5]	0.89±0.0014	0.70±0.0017	0.63±0.0047	0.89±0.0014	0.70±0.0017	0.63±0.0047
	RETAIN[24]	0.87±0.0010	0.65±0.0004	0.60±0.0006	0.88±0.0002	0.67±0.0044	0.61±0.0070
	ATTAIN[171]	0.88±0.0022	0.67±0.0082	0.60±0.0002	0.90±0.0067	0.68±0.0134	0.61±0.0094
	nTAND[144]	0.87±0.0077	0.64±0.0120	0.59±0.0133	0.89±0.0043	0.70±0.0090	0.62±0.0103
	HiTANet[102]	0.88±0.0164	0.70±0.0351	0.60±0.0067	0.90±0.0502	0.73±0.0025	0.62±0.0017
	ConCare[104]	0.88±0.0052	0.70±0.0182	0.61±0.0602	0.89±0.0288	0.72±0.0375	0.62±0.0038
	AMITA	0.89±0.0125	0.70±0.0047	0.63±0.0014	0.92±0.0013	0.75±0.0001	0.66±0.0055
	AMITA$_{\text{sum}}$	0.90±0.0247	0.71±0.0258	0.63±0.0384	0.93±0.0002	0.78±0.0003	0.68±0.0021
	AMITA$_{\text{max}}$	0.90±0.0019	0.73±0.0010	**0.65±0.0011**	0.93±0.0018	0.78±0.0001	0.69±0.0123
	AMITA$_{\text{mean}}$	0.90±0.0047	0.72±0.0158	0.64±0.0084	0.93±0.0102	0.78±0.0133	0.68±0.0431
	AMITA$_{\text{global mean}}$	0.90±0.0033	0.72±0.0094	0.64±0.0079	0.93±0.0136	0.79±0.0215	**0.70±0.0132**
1-YEAR	**LSTM**	0.84±0.0010	0.69±0.0017	0.60±0.0050	0.85±0.0043	0.72±0.0061	0.61±0.0130
	LSTM-WΔt	0.84±0.0051	0.71±0.0057	0.61±0.0170	0.86±0.0017	0.73±0.0002	0.63±0.0012
	Time LSTM[176]	0.84±0.0060	0.71±0.0080	0.64±0.0086	0.83±0.0059	0.70±0.0020	0.65±0.0088
	TLSTM[5]	0.85±0.0032	0.74±0.0060	0.66±0.0065	0.84±0.0067	0.71±0.0076	0.66±0.0090
	RETAIN[24]	0.81±0.0010	0.68±0.0086	0.62±0.0107	0.83±0.0058	0.69±0.0063	0.65±0.0014
	ATTAIN[171]	0.84±0.0070	0.65±0.0112	0.61±0.0097	0.85±0.0051	0.72±0.0078	0.64±0.0074
	nTAND[144]	0.83±0.0053	0.69±0.0082	0.63±0.0070	0.86±0.0047	0.73±0.0835	0.64±0.0058
	HiTANet[102]	0.85±0.0301	0.69±0.0036	0.60±0.0062	0.85±0.0352	0.71±0.0037	0.62±0.0036
	ConCare[104]	0.84±0.0092	0.68±0.0045	0.62±0.0802	0.85±0.0002	0.71±0.0175	0.63±0.0603
	AMITA	0.85±0.0065	0.73±0.0047	0.65±0.0014	0.87±0.0113	0.76±0.0091	0.67±0.0005
	AMITA$_{\text{sum}}$	0.86±0.0037	0.75±0.0058	0.66±0.0084	0.88±0.0002	0.77±0.0003	0.67±0.0173
	AMITA$_{\text{max}}$	0.85±0.0119	0.75±0.0310	0.66±0.0033	0.89±0.0018	0.78±0.0001	0.69±0.0123
	AMITA$_{\text{mean}}$	0.85±0.0247	0.74±0.0058	0.65±0.0084	0.89±0.0402	0.78±0.0033	0.68±0.0021
	AMITA$_{\text{global mean}}$	0.86±0.0019	0.75±0.0010	**0.66±0.0001**	0.89±0.0096	0.79±0.0215	**0.69±0.0121**

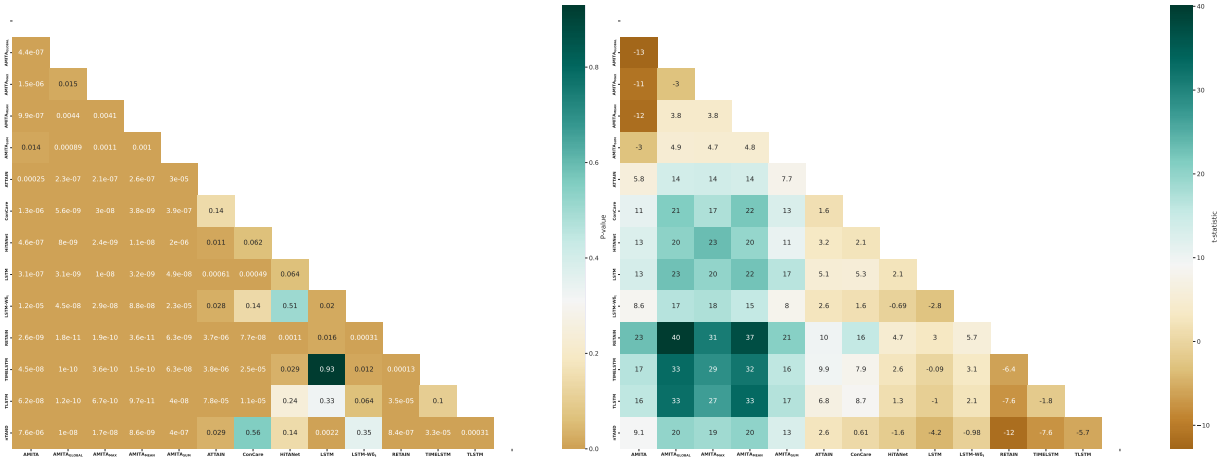


(a) P-value (b) t-statistic

Figure 7.4: Pairwise comparisons p-value < 0.05, conducted on F1-scores across the 10 testing folds using Feature Set B over the first 48-hours data, reveal the significance of the models in predicting ICU Mortality. A lower p-value indicates greater statistical significance.

the two models. The positive t-statistic further indicates that the mean of our model is significantly greater than that of **ATTAIN**. In summary, the t-test results confirm a statistically significant distinction between **AMITA** and **ATTAIN**, offering substantial evidence

that our proposed model significantly outperforms the latter. Overall, similar analyses were conducted for all other baseline models, yielding consistent results.



(a) P-value (b) t-statistic

Figure 7.5: Pairwise comparisons p-value < 0.05, conducted on F1-scores across the 10 testing folds using Feature Set B over the first 48-hours data, reveal the significance of the models in predicting HOSPITAL Mortality. A lower p-value indicates greater statistical significance.

Table 7.16: Length of Stay prediction in DAYS (AMITA)

DATA USED	MODELS	SAPS II FEATURES			ALL FEATURES		
		RMSE	MAE	Adjusted R^2	RMSE	MAE	Adjusted R^2
24 HOURS DATA	**LSTM**	5.86±0.3364	2.51±0.0986	0.14±0.0265	5.66±0.3359	2.48±0.1226	0.15±0.0218
	LSTM-WΔt	5.70±0.3057	2.35±0.0807	0.17±0.0241	5.57±0.2979	2.33±0.1393	0.17±0.0123
	Time LSTM[176]	5.39±0.1776	2.79±0.0776	0.14±0.0278	4.45±0.1748	2.40±0.0748	0.17±0.0118
	TLSTM[5]	5.82±0.3457	2.52±0.0727	0.15±0.0265	5.80±0.4579	2.53±0.0393	0.16±0.0231
	RETAIN[24]	6.05±0.2282	3.26±0.0691	0.10±0.0209	6.93±0.2317	3.66±0.0878	0.10±0.0225
	ATTAIN[171]	5.75±0.2432	2.86±0.0531	0.13±0.0213	5.80±0.2101	2.82±0.0188	0.19±0.0275
	nTAND[144]	5.85±0.2132	2.92±0.0567	0.14±0.0167	5.88±0.2212	2.80±0.0537	0.20±0.0195
	HiTANet[102]	5.90±0.2663	3.01±0.0231	0.12±0.0113	5.66±0.2248	2.82±0.0673	0.19±0.0135
	ConCare[104]	5.71±0.2312	3.14±0.0582	0.12±0.0244	5.65±0.2254	3.02±0.0677	0.15±0.0212
	AMITA	1.35±0.0183	1.05±0.0151	0.48±0.0143	1.27±0.0131	0.95±0.0132	0.50±0.0108
	AMITA$_{\text{sum}}$	1.32±0.0155	0.97±0.0107	0.50±0.0113	1.25±0.0122	0.90±0.0093	0.51±0.0085
	AMITA$_{\text{max}}$	**1.23±0.0134**	**0.90±0.0112**	**0.51±0.0125**	**1.17±0.0117**	**0.86±0.0075**	**0.53±0.0092**
	AMITA$_{\text{mean}}$	1.27±0.0177	0.91±0.0146	0.51±0.0168	1.18±0.0131	0.88±0.0086	0.52±0.0101
	AMITA$_{\text{global mean}}$	1.29±0.0132	0.92±0.0102	0.50±0.0191	1.18±0.0107	0.88±0.0057	0.52±0.0134
48 HOURS DATA	**LSTM**	5.42±0.2300	2.30±0.0275	0.28±0.0285	5.19±0.3688	2.11±0.1935	0.35±0.0204
	LSTM-WΔt	5.32±0.3294	1.91±0.0916	0.31±0.0268	5.18±0.3615	2.07±0.1364	0.36±0.0254
	Time LSTM[176]	4.86±0.2099	2.11±0.0446	0.35±0.0233	5.07±0.2161	2.44±0.0437	0.38±0.0231
	TLSTM[5]	4.88±0.1529	2.23±0.0442	0.35±0.0216	5.25±0.2331	2.45±0.0723	0.31±0.0248
	RETAIN[24]	6.19±0.2902	3.29±0.0432	0.12±0.0271	6.59±0.5105	3.39±0.0638	0.15±0.0251
	ATTAIN[171]	5.60±0.2033	2.26±0.0251	0.34±0.0236	5.53±0.2413	2.03±0.0547	0.37±0.0217
	nTAND[144]	5.67±0.2034	2.39±0.0298	0.33±0.0243	5.73±0.2263	2.15±0.0772	0.35±0.0259
	HiTANet[102]	5.71±0.2437	2.33±0.0572	0.34±0.0216	5.49±0.2134	2.24±0.0727	0.34±0.0123
	ConCare[104]	5.80±0.2044	2.56±0.0369	0.34±0.0261	5.74±0.2437	2.49±0.0956	0.35±0.0258
	AMITA	0.92±0.0221	0.61±0.0192	0.70±0.0268	0.91±0.0253	0.60±0.0175	0.70±0.0114
	AMITA$_{\text{sum}}$	0.92±0.0215	0.61±0.0187	0.70±0.0271	0.90±0.0179	0.58±0.0138	0.72±0.0127
	AMITA$_{\text{max}}$	**0.89±0.0207**	**0.57±0.0168**	**0.72±0.0103**	**0.87±0.0192**	**0.55±0.0112**	**0.74±0.0088**
	AMITA$_{\text{mean}}$	0.90±0.0152	0.60±0.0065	0.72±0.0116	0.89±0.0142	0.58±0.0080	0.73±0.0079
	AMITA$_{\text{global mean}}$	0.92±0.0153	0.61±0.0134	0.71±0.0079	0.90±0.0114	0.58±0.0078	0.73±0.0065

Table 7.17: Length of Stay prediction in DAYS on eICU dataset (AMITA)

DATA USED	MODELS	SAPS II FEATURES			ALL FEATURES		
		RMSE	MAE	Adjusted R^2	RMSE	MAE	Adjusted R^2
24 HOURS DATA	**AMITA**	1.45±0.0183	1.10±0.0195	0.49±0.0272	1.31±0.0142	0.94±0.0188	0.49±0.0236
	AMITA$_{\text{sum}}$	1.37±0.0205	0.97±0.0241	0.49±0.0215	1.35±0.0204	0.94±0.0188	0.50±0.0282
	AMITA$_{\text{max}}$	**1.18±0.0261**	**0.87±0.0131**	**0.53±0.0168**	**1.13±0.0159**	**0.85±0.0137**	**0.54±0.0172**
	AMITA$_{\text{mean}}$	1.31±0.0204	0.90±0.0242	0.50±0.0104	1.23±0.0147	0.89±0.0181	0.51±0.0256
	AMITA$_{\text{global mean}}$	1.26±0.0168	0.90±0.0122	0.50±0.0291	1.24±0.0177	0.90±0.0087	0.50±0.0214
48 HOURS DATA	**AMITA**	0.93±0.0273	0.65±0.0204	0.70±0.0239	0.91±0.0207	0.62±0.0136	0.71±0.0144
	AMITA$_{\text{sum}}$	0.90±0.0290	0.62±0.0145	0.71±0.0234	0.85±0.0188	0.57±0.0102	0.74±0.0105
	AMITA$_{\text{max}}$	**0.86±0.0266**	**0.58±0.0138**	**0.74±0.0108**	**0.81±0.0163**	**0.51±0.0137**	**0.77±0.0063**
	AMITA$_{\text{mean}}$	0.91±0.0210	0.63±0.0275	0.70±0.0262	0.87±0.0129	0.60±0.0171	0.74±0.0087
	AMITA$_{\text{global mean}}$	0.91±0.0213	0.63±0.0127	0.70±0.0135	0.91±0.0184	0.63±0.0133	0.71±0.0118

7.3.2.2 Length of stay (LOS)

In the context of the Length of Stay (LOS) prediction task, we evaluate the precision of our predictions against actual LOS data from the test set, employing three pivotal metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Adjusted R^2. The dataset spans a wide range of LOS, from a minimum of 1.2 to a maximum of 173.07 days. Detailed error metrics, including MAE, RMSE and Adjusted R^2 across various models, are summarized in Table 7.16. Remarkably, our best performance yielded RMSE and MAE values of less than 1 day, specifically 20.88 and 13.2 hours, respectively, compared to the ground truth data, with an Adjusted R^2 of 74%.

Among the models evaluated, the **AMITA**$_{\text{max}}$ model stands out for its consistently superior performance across all metrics, as demonstrated in Table 7.16. This model exhibits unparalleled accuracy in comparison to the current leading approaches. Its predictions for RMSE are notably less than one day, irrespective of the feature set (A or B) when an extended observation period 48-hours data span used as inputs.

Analyzing results from both eICU and MIMIC-III datasets unveils a consistent trend, albeit with slight variations in metric values attributable to their distinct LOS distribution profiles. Each dataset exhibits a rightward skew in LOS distribution, however, this skewness is more pronounced in the eICU dataset compared to the MIMIC-III dataset, as illustrated in Fig. 7.3. Such skewness significantly influences the calculation of error metrics like MSE and MAE, especially given that the majority of stays are shorter than 4 days. This distribution poses a challenge in LOS prediction, highlighting the complexities involved.

Reflecting on the entire prediction task, we observe that our model exhibits notably improved performance when leveraging a more extensive set of features for prediction. This underscores our model's adeptness in learning feature representations from multiple data modalities and effectively managing the irregular timing patterns inherent in EHR data.

7.4 Ablation studies

This section presents a systematic series of ablation studies conducted to evaluate the contribution of individual components and configurations within our proposed machine learning model. Ablation studies, by systematically removing or modifying components of a

model, help in understanding the role and importance of each component in the model's performance. These studies are critical for validating the design choices made during the development of the model. We initiated the ablation studies by establishing a baseline model incorporating all proposed features and optimizations. Subsequent iterations involved modifications including the removal of specific components, alterations in model architecture. Each variant was evaluated against a consistent dataset, with performance metrics such as accuracy, precision, recall, and F1-score documented for comparison.

7.4.1 MWTA-LSTM

To evaluate the impact of auxiliary components in our model, we conducted an ablation study, systematically removing these components individually or collectively and comparing the resulting variants with the original **MWTA-LSTM**.

The experimental results for ICU and Hospital mortality prediction tasks on MIMIC-III are presented in Table 7.18, where **"w/o"** denotes "without." The performance of **MWTA-LSTM 2 w/o all** experiences a significant decline when all auxiliary components are removed, underscoring the importance of each considered auxiliary component in capturing intricate patterns in patients' EHR records. Similar trends are observed for **MWTA-LSTM 1 w/o all**, highlighting the importance of each individual component for our framework. Despite this, our proposed method continues to outperform the baseline models in predicting these tasks, as evidenced in Table 7.2 and Table 7.3.

Table 7.18: In-hospital & ICU mortality task on MIMIC dataset using First 24 & 48 HRS data (MWTA-LSTM)

DATA	TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
			AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
24 HRS DATA	In-Hospital	**MWTA-LSTM-2 ASP**	**0.90±0.0019**	**0.70±0.0010**	**0.62±0.0001**	**0.93±0.0018**	**0.76±0.0001**	**0.69±0.0123**
		MWTA-LSTM-2 w/o ASP	0.91±0.0047	0.70±0.0058	0.62±0.0084	0.93±0.0002	0.76±0.0003	0.67±0.0021
		MWTA-LSTM-2 w/o attention	0.90±0.0088	0.68±0.0108	0.60±0.0054	0.91±0.0079	0.73±0.0233	0.65±0.0067
		MWTA-LSTM-2 w/o gating	0.91±0.0014	0.68±0.0126	0.60±0.0204	0.92±0.0132	0.74±0.0071	0.66±0.0151
		MWTA-LSTM-2 w/o all	0.89±0.0017	0.68±0.0075	0.60±0.0084	0.91±0.0097	0.70±0.0120	0.62±0.0100
		MWTA-LSTM-1 w/o all	0.89±0.0029	0.67±0.0115	0.60±0.0234	0.90±0.0109	0.68±0.0110	0.61±0.0205
	ICU	**MWTA-LSTM-2 ASP**	**0.93±0.0041**	**0.73±0.0035**	**0.64±0.0078**	**0.96±0.0028**	**0.80±0.0076**	**0.71±0.0076**
		MWTA-LSTM-2 w/o ASP	0.93±0.0012	0.72±0.0034	0.64±0.0103	0.95±0.0005	0.75±0.0054	0.68±0.0004
		MWTA-LSTM-2 w/o attention	0.92±0.0054	0.70±0.0037	0.62±0.0093	0.93±0.0225	0.74±0.0144	0.67±0.0109
		MWTA-LSTM-2 w/o gating	0.92±0.0158	0.70±0.0078	0.61±0.0209	0.93±0.0165	0.73±0.0077	0.67±0.0089
		MWTA-LSTM-2 w/o all	0.92±0.0009	0.65±0.0070	0.61±0.0151	0.94±0.0027	0.72±0.0107	0.64±0.0104
		MWTA-LSTM-1 w/o all	0.91±0.0129	0.64±0.0037	0.60±0.0125	0.94±0.0114	0.71±0.0173	0.63±0.0154
48 HRS DATA	In-Hospital	**MWTA-LSTM-2 ASP**	**0.94±0.0127**	**0.78±0.0047**	**0.70±0.0080**	**0.95±0.0036**	**0.81±0.0029**	**0.73±0.0073**
		MWTA-LSTM-2 w/o ASP	0.94±0.0007	0.78±0.0007	0.69±0.0030	0.94±0.0018	0.80±0.0019	0.72±0.0076
		MWTA-LSTM-2 w/o attention	0.92±0.0112	0.75±0.0305	0.67±0.0070	0.93±0.0038	0.77±0.0051	0.69±0.0206
		MWTA-LSTM-2 w/o gating	0.93±0.0105	0.76±0.0058	0.69±0.0054	0.93±0.0021	0.78±0.0314	0.71±0.0068
		MWTA-LSTM-2 w/o all	0.93±0.0223	0.73±0.0089	0.64±0.0073	0.94±0.0103	0.75±0.0107	0.65±0.0061
		MWTA-LSTM-1 w/o all	0.92±0.0179	0.72±0.0025	0.62±0.0175	0.93±0.0193	0.73±0.0097	0.63±0.0088
	ICU	**MWTA-LSTM-2 ASP**	**0.96±0.0044**	**0.81±0.0122**	**0.72±0.0042**	**0.97±0.0024**	**0.85±0.0028**	**0.77±0.0055**
		MWTA-LSTM-2 w/o ASP	0.96±0.0004	0.79±0.0022	0.70±0.0012	0.97±0.0025	0.83±0.0019	0.74±0.0046
		MWTA-LSTM-2 w/o attention	0.94±0.0208	0.76±0.0075	0.68±0.0117	0.95±0.0057	0.80±0.0107	0.71±0.0312
		MWTA-LSTM-2 w/o gating	0.94±0.0207	0.76±0.0502	0.69±0.0219	0.96±0.0176	0.80±0.0067	0.72±0.0067
		MWTA-LSTM-2 w/o all	0.94±0.0023	0.74±0.0401	0.66±0.0310	0.95±0.0109	0.77±0.0032	0.68±0.0063
		MWTA-LSTM-1 w/o all	0.92±0.0136	0.73±0.0171	0.65±0.0151	0.93±0.0265	0.75±0.0133	0.66±0.0106

7.4.2 AMITA

To assess the influence of auxiliary components in our model, we conducted an ablation study, systematically removing these components either individually or collectively and comparing the resulting variants with the original **AMITA**. The experimental outcomes for ICU and Hospital mortality prediction tasks on the MIMIC-III cohort are outlined in Table 7.19, where **"w/o"** denotes "without." The performance of **AMITA w/o all** experiences a significant decline when all auxiliary components are removed, underscoring the importance of each considered auxiliary component in capturing intricate patterns in patients' EHR records. Notably, **AMITA w/o all** reverts to a basic LSTM, albeit with the cell state adjusted for irregular timing, exhibiting the least favorable performance among the variants. Despite this, it still outperforms the baseline models in predicting these tasks, as evidenced in Table 7.10 and Table 7.12.

Concerning **AMITA w/o last**, the variant without the last observation, as expressed in (Eq. 6.7), exhibits a slight decrease in performance compared to **AMITA**, yet it consistently outperforms the baselines with an F1-score exceeding 2% for all tasks. For the **AMITA w/o f_t^{new}**, the variant without the two decay terms, as expressed in (Eq. 6.11), experiences a 2% decrease in F1-score compared to **AMITA** for both tasks, underscoring their importance. Regarding **AMITA w/o time gate t_t**, the variant without the time gate (Eq. 6.12) exhibits a significant decrease in performance compared to **AMITA** and other versions, except for **AMITA w/o all**. This decrease is attributed to the model's inability to understand the dynamic health status of patient's illness course, as stated previously regarding the importance of using elapsed time to model the effects of interventions on the cell state. The figure, identified as Fig. 7.6, presents a parallel coordinates plot that efficiently illustrates the performance metrics of each model variation in a unified view.

Table 7.19: In-hospital & ICU mortality task on MIMIC dataset using First 24 & 48 HRS data (AMITA).

DATA	TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
			AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
24 HRS DATA	In-Hospital	**AMITA**	**0.89±0.0073**	**0.66±0.0191**	**0.60±0.0137**	**0.92±0.0067**	**0.72±0.0159**	**0.65±0.0163**
		AMITA w/o all	0.88±0.0105	0.63±0.0135	0.57±0.0245	0.91±0.0058	0.68±0.0140	0.61±0.0100
		AMITA w/o f_t^{new} (Eq. 6.11)	0.89±0.0059	0.65±0.0152	0.59±0.0142	0.91±0.0075	0.72±0.0168	0.63±0.0148
		AMITA w/o time gate t_t (Eq. 6.12)	0.88±0.0061	0.63±0.0153	0.58±0.0131	0.91±0.0055	0.70±0.0148	0.62±0.0094
		AMITA w/o last L_{x_t} (Eq. 6.7)	0.89±0.0283	0.65±0.0327	0.59±0.0374	0.92±0.0090	0.71±0.0169	0.64±0.0131
	ICU	**AMITA**	**0.92±0.0075**	**0.66±0.0215**	**0.60±0.0174**	**0.94±0.0045**	**0.74±0.0183**	**0.65±0.0164**
		AMITA w/o all	0.91±0.0092	0.63±0.0178	0.57±0.0245	0.93±0.0071	0.68±0.0205	0.62±0.0108
		AMITA w/o f_t^{new} (Eq. 6.11)	0.91±0.0066	0.64±0.0209	0.59±0.0163	0.93±0.0059	0.70±0.0183	0.63±0.0160
		AMITA w/o time gate t_t (Eq. 6.12)	0.91±0.0071	0.64±0.0191	0.58±0.0136	0.92±0.0056	0.69±0.0171	0.62±0.0152
		AMITA w/o last L_{x_t} (Eq. 6.7)	0.92±0.0115	0.65±0.0415	0.59±0.0311	0.94±0.0080	0.72±0.0251	0.65±0.0233
48 HRS DATA	In-Hospital	**AMITA**	**0.92±0.0225**	**0.73±0.0147**	**0.65±0.0014**	**0.93±0.0066**	**0.77±0.0132**	**0.67±0.0122**
		AMITA w/o all	0.90±0.0075	0.69±0.0317	0.62±0.0271	0.92±0.0128	0.72±0.0010	0.65±0.0102
		AMITA w/o f_t^{new} (Eq. 6.11)	0.91±0.0064	0.70±0.0155	0.63±0.0152	0.93±0.0051	0.75±0.0129	0.66±0.0102
		AMITA w/o time gate t_t (Eq. 6.12)	0.91±0.0049	0.70±0.0125	0.64±0.0124	0.93±0.0049	0.75±0.0111	0.67±0.0125
		AMITA w/o last L_{x_t} (Eq. 6.7)	0.92±0.0105	0.71±0.0258	0.64±0.0321	0.93±0.0045	0.73±0.0183	0.65±0.0164
	ICU	**AMITA**	**0.94±0.0155**	**0.75±0.0147**	**0.67±0.0114**	**0.95±0.0042**	**0.78±0.0175**	**0.70±0.0181**
		AMITA w/o all	0.93±0.0071	0.69±0.0165	0.64±0.0187	0.94±0.0045	0.73±0.0183	0.66±0.0164
		AMITA w/o f_t^{new} (Eq. 6.11)	0.94±0.0064	0.74±0.0141	0.66±0.0166	0.95±0.0049	0.76±0.0122	0.68±0.0107
		AMITA w/o time gate t_t (Eq. 6.12)	0.93±0.0135	0.72±0.0159	0.65±0.0173	0.94±0.0132	0.74±0.0207	0.66±0.0186
		AMITA w/o last L_{x_t} (Eq. 6.7)	0.94±0.0078	0.73±0.0425	0.66±0.0254	0.95±0.0006	0.76±0.0041	0.68±0.0078

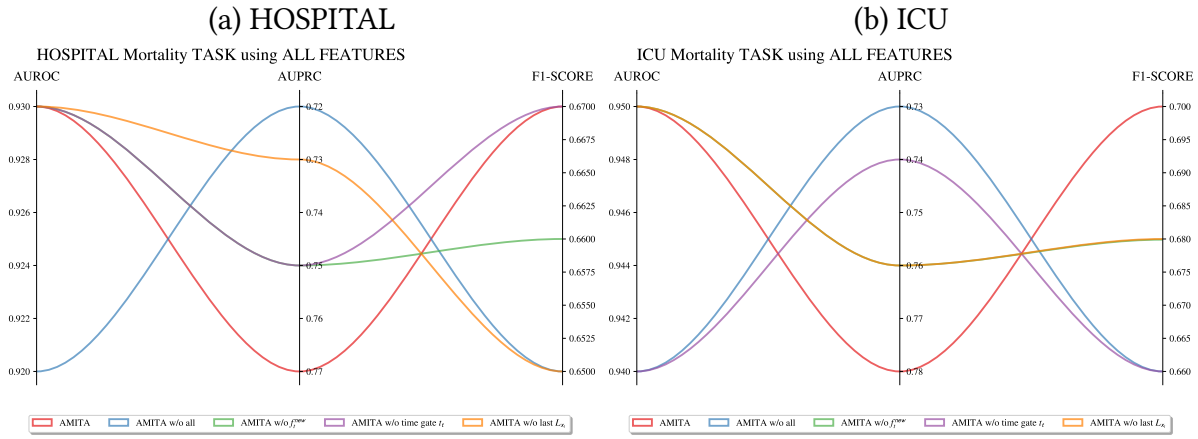


Figure 7.6: Parallel Coordinates Plot for Ablation Study results over the first 48-hours data on mortalities tasks.

7.4.3 Discussion & Conclusion

The ablation studies underscored the significance of several model components and configurations, confirming the necessity of certain features and architectural choices while identifying others as superfluous or suboptimal. These insights not only validated the model's design rationale but also guided further refinement, leading to improved efficiency and performance. Importantly, this iterative process highlighted the delicate interplay between model complexity and generalizability, reinforcing the need for careful consideration in each design decision.

The conducted ablation studies has provided a comprehensive evaluation of the proposed model, affirming the efficacy of key components and informing future improvements. This meticulous approach to model evaluation and refinement is essential for advancing machine learning research, contributing to the development of more robust, efficient, and interpretable models.

7.5 Interpretability

This section focuses on elucidating the mechanisms and decision-making processes of our deep learning model, enhancing our understanding and trust in its outputs. By implementing techniques such as feature importance analysis, saliency maps, and model-agnostic methods, we uncover the critical factors influencing predictions. This not only aids in validating the model's performance but also provides insights into the underlying patterns and relationships within the data, offering valuable clinical interpretations and steering future research and practical applications towards more informed pathways.

7.5.1 Use cases of ranking critical features using both attention values and frequency values of input features

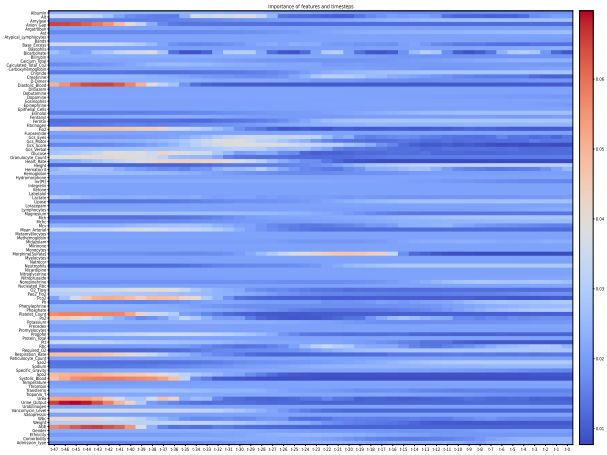
Fig. 7.7(b) showcases the top 30 critical features as identified by our predictive model for a patient with a positive prognosis and complex medical history including (Bladder carcinoma, Diabetes Type II, Hypertension etc...), displaying a bar plot that captures the average influence of each feature over time on predicting ICU mortality, each bar annotated with numerical values indicating the frequency of records made across the entire timesteps. Meanwhile, Fig. 7.7(a) presents a heatmap that delineates the variation in input feature values across various timesteps. Upon analyzing Fig. 7.7(c), we discern a combination of clinical indicators suggesting a critical and potentially life-threatening condition for this specific patient.

The presence of a heart rate(ranked 1st) less than 40, diastolic blood pressure(ranked 2) less than 20, and systolic blood pressure(ranked 5) less than 50 collectively indicates severe cardiovascular compromise. These values suggest poor cardiac output, inadequate perfusion to vital organs, and an elevated risk of shock. The high lactate level(ranked 3) of 15 is particularly concerning and is often associated with insufficient tissue oxygenation, potentially stemming from the compromised cardiovascular status, further highlighting the severity of the patient's condition. Additionally, the high pCO2(ranked 4) level above 70 suggests respiratory distress and impaired gas exchange, signifying that the respiratory system may be struggling to eliminate carbon dioxide, leading to respiratory acidosis. To assess the patient's prognosis, we have calculated the mortality rate based on SAPS-II features [85] using the initial first 24 hours data. The mortality rate surpasses 96% for SAPS-II score points of 90, underscoring the severity of the situation. The frequent use of vasopressors such as Norepinephrine also underscores the critical condition of the patient. These medications are typically administered to stabilize blood pressure in severely ill patients, and their consistent application indicates the management of serious conditions like circulatory shock or hypotension.

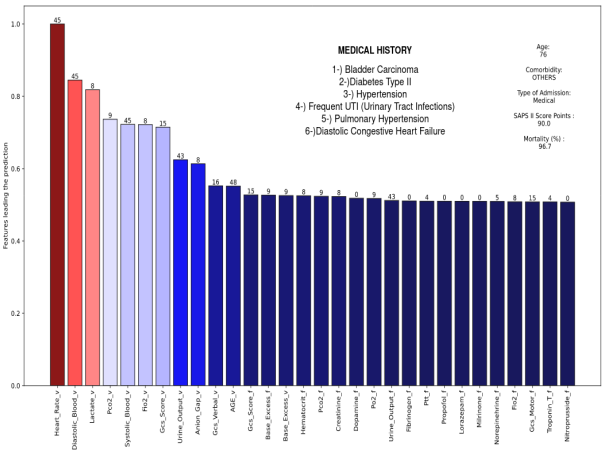
In summary, the depicted clinical features in Fig. 7.7(c) collectively paint a picture of a complex and severe medical scenario involving cardiovascular and respiratory compromise. The detailed insights provided by our model contribute to a better understanding of the patient's condition, guiding effective and timely medical interventions. Furthermore, our model tracks the importance (decay weights) of such features throughout the patient's illness course, as depicted in Fig. 7.8 and the evolution of the forget gate through the course of illness using both (time-based and frequency-weighted based) forgetting, as shown in Fig. 7.7(c) .

Fig. 7.9(b) illustrates critical features identified for a positive patient utilizing Feature Set B. Examining Fig. 7.9(c) , which delineates the patient's health trajectory during hospitalization, highlights our model's superior interpretability in discerning crucial aspects of the patient's profile. Prominently, key features such as heart rate, GCS components, Oxygen parameters (Spo2, FiO2, SaO2), (Lactate, Anion Gap, PCO2) metabolic acidosis factors, Heart Rate and Systolic Blood.

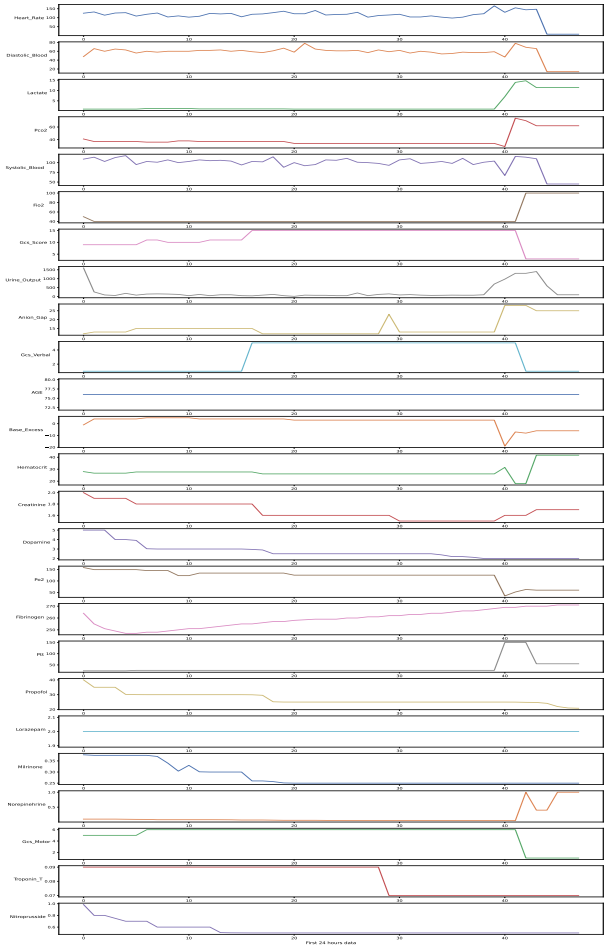
Moreover, Fig. 7.9(d) provides decay weights for these features, elucidating their respec-



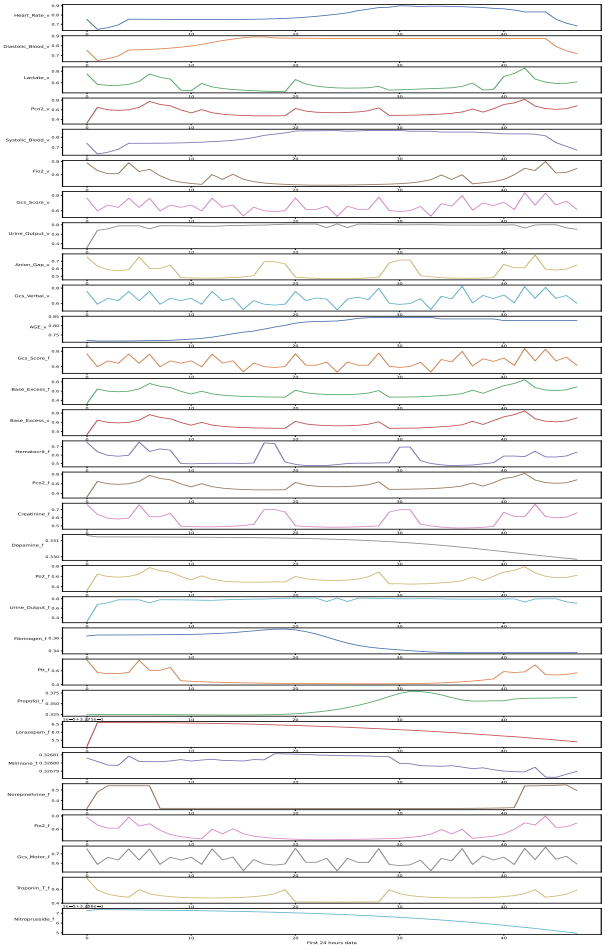
(a) The features timesteps importances.



(b) The first 30 most importants features.



(c)Patient's health trajectory during his/her hospitalization.



(d) forget gate (f_t^{new} (Eq. (6.11)) on patient's illness journey.

Figure 7.7: Positive patient on ICU Mortality task using Feature set B, decay weights for those features are shown in Fig. 7.8.

tive importance. By analyzing Fig. 7.9(c), key physiological parameters have been highlighted, These physiological changes and corresponding lab results demonstrate extensive system involvement and profound clinical deterioration. Key observations include a sharp

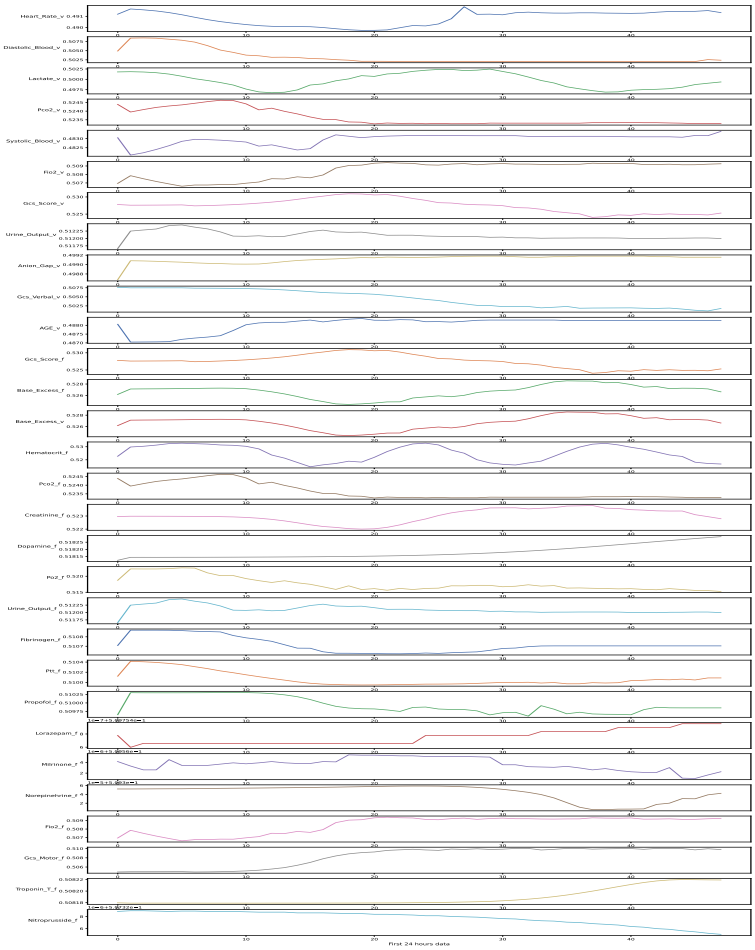
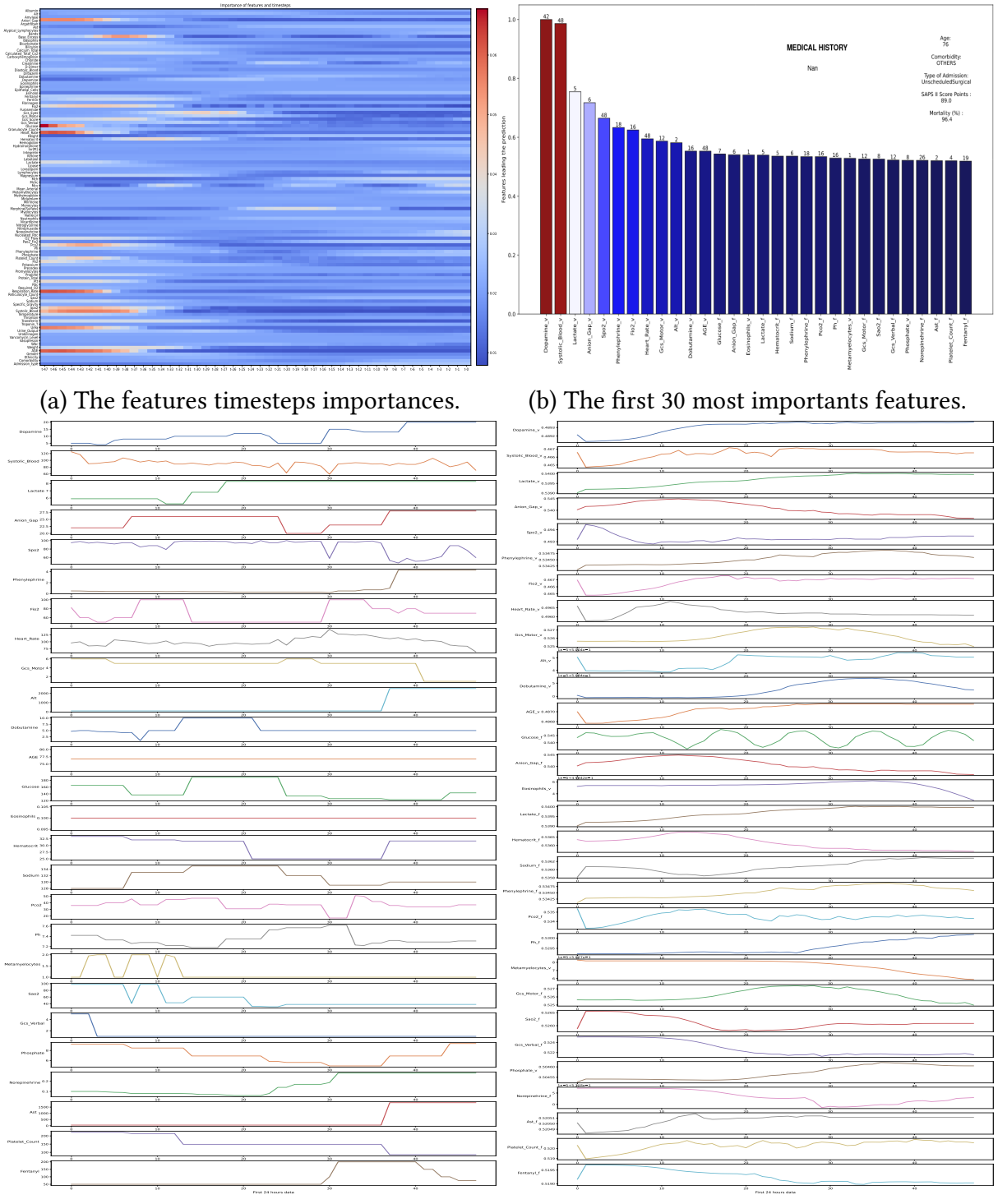


Figure 7.8: decaying weights, as expressed in (Eq. (6.2)) on patient's illness for the most 30 influential features.

decrease in heart rate(ranked 8) from 125 to below 50, and systolic blood pressure(ranked 2) dropping from 120 to below 70, both indicative of severe cardiovascular compromise. Respiratory difficulties are evident from the reduction in SpO2(ranked 5) from 98% to below 60%, signaling acute respiratory compromise requiring maximum oxygen supplementation to maintain sufficient oxygen levels as stated an increasing Fio2(ranked 7) level at 100% .

Furthermore, metabolic disturbances are indicated by a substantial rise in lactate(ranked 3) levels from 4 to 8 mmol/L and an increasing anion gap(ranked 4) from 20 to 27.5 mEq/L, suggesting worsening metabolic acidosis. These shifts, combined with a significant decrease in pCO2(ranked 20) to below 20 mmHg, point to a critical state of compensatory respiratory alkalosis, potentially triggered by hyperventilation in response to metabolic acidosis.

Other concerning changes include a severe decrease in hematocrit(ranked 17) levels from 32.5% to 25%, reflecting possible acute blood loss or dilution, and a drastic drop in platelets from 200 to below 100 x $10^3 \mu$L, heightening the risk of bleeding. Liver function is also severely compromised, as evidenced by ALT(ranked 10) levels rising from near zero to 2000 units per liter and AST(ranked 28) levels from near zero to 1500 units per liter,



(a) The features timesteps importances.

(b) The first 30 most importants features.

(c) Patient's health trajectory during his/her hospitalization.

(d) decaying weights, as expressed in (Eq. 6.2) on patient's illness.

Figure 7.9: Positive patient on ICU Mortality task.

suggesting significant hepatocellular injury.

The escalated administration of vasopressors, including dopamine(ranked 1st) increasing from 5 to 20, phenylephrine(ranked 6) from near zero to above 4, dobutamine(ranked

11) from 2 to 10, and norepinephrine(ranked 27) from near zero to 0.2, highlights the need for substantial hemodynamic support due to the patient's critically unstable condition.

For a positive patient with a comorbidity of metastatic cancer with a complex oncological history, Fig. 7.10(a) provide a detailed analysis of critical health parameters, each bar annotated with numerical values indicating the frequency of records made, utilizing insights from feature set B. The comprehensive review in Fig. 7.10(c), which charts the patient's health journey throughout his/her hospital stay, enhances our understanding by pinpointing key indicators directly associated with the unique challenges posed by metastatic cancer. In this scenario, the highlighted physiological readings signal a dire clinical situation. A heart rate(ranked 17) above 120 bpm, systolic blood pressure(ranked 1st) under 50 mmHg, and mean arterial pressure below 50 mmHg underscore a critical reduction in cardiovascular functionality, likely aggravated by the metastatic burden and associated therapies. A GCS score(ranked 27) plummeting from 15 to 3 indicates severe neurological compromise, which could stem from brain metastases or systemic effects of advanced cancer.

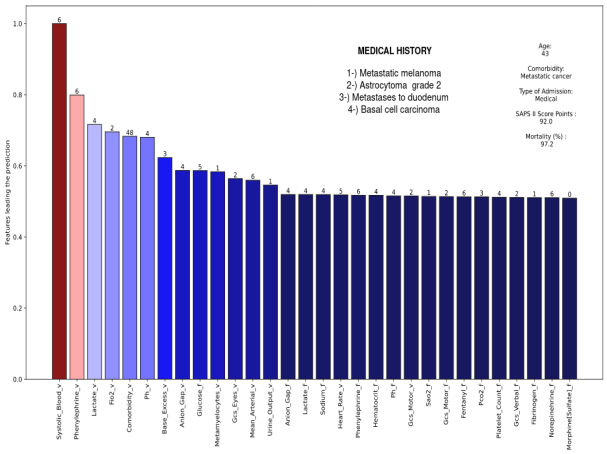
An escalating lactate level(ranked 3), rising from 3 to 6 mmol/L, alongside a FiO2(ranked 4) requirement jumping from 50% to 100%, suggests acute hypoxemia and possible lactic acidosis, conditions that are exacerbated by the cancer's progression and its impact on organ function.

The increase in the Anion Gap(ranked 8) to 20, possibly linked to elevated lactate levels as indicated in Fig. 7.10(c), coupled with a pH(ranked 6) decrease from 7.5 to below 7, actually reflects a dramatic shift from an initially alkalotic state towards metabolic acidosis. This condition may be precipitated by factors such as renal impairment or tumor lysis syndrome, which are common in rapidly proliferating or treated cancers. Furthermore, a significant shift in Base Excess(ranked 7) from +5 to -10 corroborates the extent of metabolic derangement, indicating severe acidosis and the body's efforts to buffer the acidotic blood.

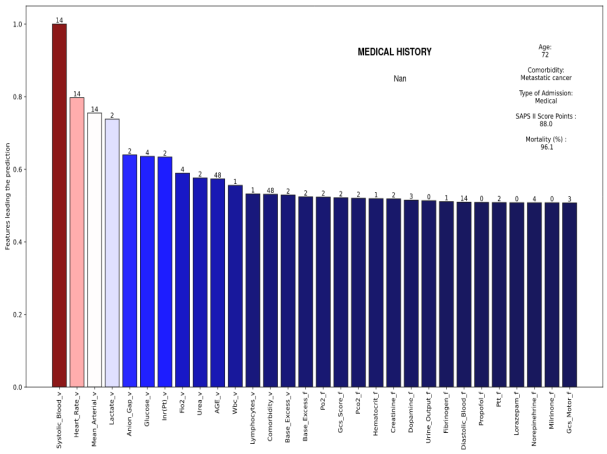
A hematocrit(ranked 19) level plummeting to 21% signals severe anemia, undermining the body's capacity for oxygen transport and exacerbating tissue hypoxia. This condition is further evidenced by an alarmingly low arterial oxygen saturation (SaO2)(ranked 22) below 60%, indicating a state of profound hypoxemia. The situation's gravity is underscored by a lactate level surging to 7 mmol/L, a clear marker of lactic acidosis and an indicator of systemic tissue hypoxia, suggesting that the body's tissues are resorting to anaerobic metabolism due to insufficient oxygen supply, a marker of severe metabolic distress.

The management of the patient's oxygenation status, with FiO2 adjusted from 50% to a full 100%, underscores the aggressive attempts to correct the hypoxemia, yet the persistently low SaO2(ranked 22) suggests these measures are failing to meet the oxygen demands of the body. Furthermore, a PCO2(ranked 25) level below 35 mmHg reflects the body's compensatory response to acidosis through hyperventilation, attempting to correct the pH balance by expelling carbon dioxide, an acid component of the blood.

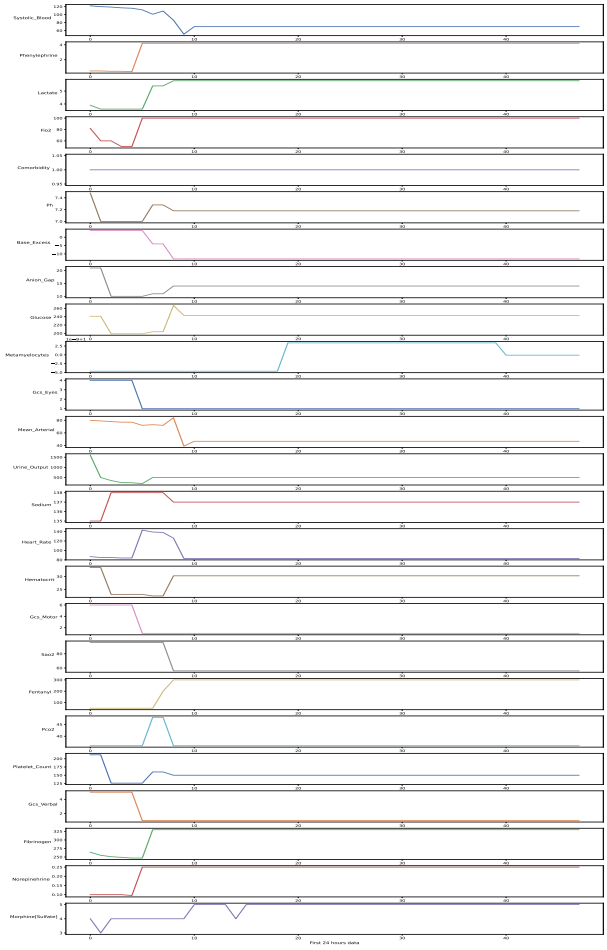
The frequent administration of Fentanyl(ranked 24), Norepinephrine(ranked 29), and Phenylephrine(ranked 2 and ranked 18) to a patient with a complex medical history that includes metastatic melanoma, Grade 2 astrocytoma, metastases to the duodenum, basal cell carcinoma, and other comorbidities associated with metastatic cancer suggests a critical care scenario. This aggressive approach to pain management and hemodynamic sta-



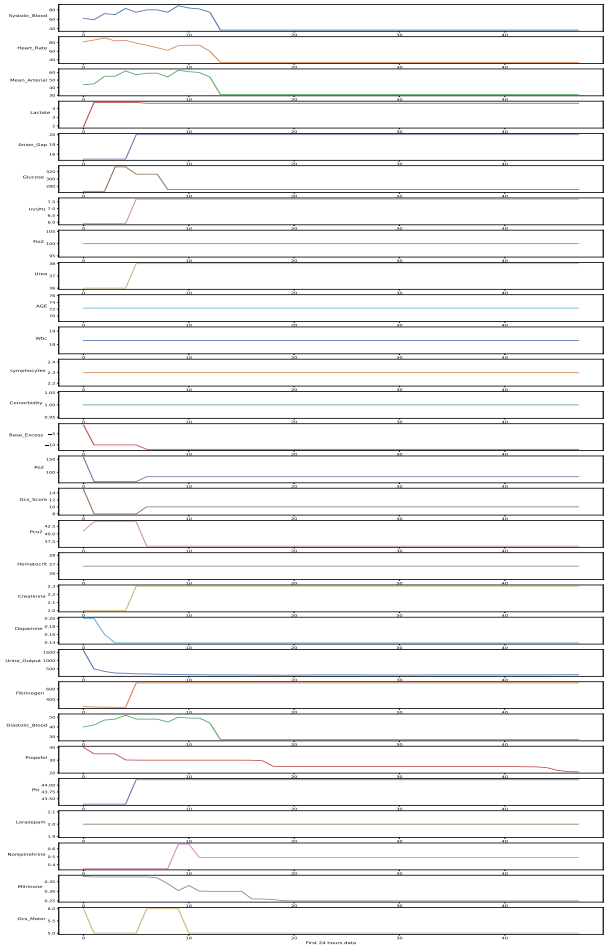
(a) The first 30 most importants features.



(b) The first 30 most importants features.



(c) Patient's health trajectory during his hospitalization.



(d) Patient's health trajectory during his hospitalization.

Figure 7.10: Two Positive patients with comorbidity (Metastatic cancer) on ICU Mortality task.

bilization reflects the challenges posed by severe pain, systemic effects of cancer, and the potential for life-threatening complications such as sepsis or shock. It underscores the need for comprehensive symptom control and cardiovascular support to maintain the patient's

comfort and vital functions amidst the complexities of advanced cancer states.

In the context of a patient with metastatic cancer without medical history, Fig. 7.10(b) offer a comprehensive depiction of crucial features identified, by utlizing features set B. The described laboratory and clinical parameters highlight a critical situation that suggests a complex interplay of acute renal failure, severe cardiovascular compromise, hypercoagulability, severe metabolic stress, and significant respiratory distress, necessitating high levels of oxygen supplementation. By examining Fig. 7.10(d), our model enhances interpretability and clearly emphasizes the most critical features uniquely associated with metastatic cancer.

For this patient with metastatic cancer, the observed physiological changes and the corresponding laboratory findings indicate multiple system involvement and severe clinical compromise. We observe a notable decline in Heart Rate(ranked 2) from 79 to below 40, Systolic Blood Pressure(ranked 1st) from 80 to under 40, Mean Arterial Pressure(ranked 3) from 60 to below 30, and Diastolic Blood Pressure(ranked 24) from 50 to under 30. These changes signal severe cardiovascular compromise, likely exacerbated by the systemic impact of metastatic cancer. Respiratory challenges are underscored by a requisite FiO2(ranked 8) of 100%, highlighting severe respiratory compromise necessitating maximal oxygen supplementation to maintain adequate oxygenation, a distress signal in the context of potential pulmonary involvement or complications. This is coupled with a pCO2(ranked 18) level dipping below 35 mmHg, suggesting a compensatory respiratory alkalosis as the body attempts to counterbalance metabolic acidosis through increased ventilation, which could be a consequence of lung metastases, pleural effusion, or other complications related to advanced cancer.

Increasing levels of urea(ranked 9) (from 36 to 38), lactate(ranked 4) (from 2 to 4), INR (PT)(ranked 7) (from 6 to 7), and anion gap(ranked 5) (from 16 to 20), coupled with a worsening base excess(ranked 14) (from 0 to -10) and a decrease in pCO2(ranked 18) (from 42 to less than 37), suggest metabolic and hemodynamic instability. These changes might reflect organ dysfunction, tissue hypoperfusion, and metabolic acidosis, often common in advanced cancer stages.

The fluctuating creatinine levels, ranging from 2 to 2.3 mg/dL, indicative of renal stress or dysfunction, possibly from cancer progression, treatment side effects, or secondary complications. Concurrently, an elevated fibrinogen level surpassing 600 mg/dL signals a heightened state of hypercoagulability and systemic inflammation, a common accompaniment in malignancies that predisposes to thrombotic events and complicates the clinical course, as evidenced by by a range of lactate levels from 2 to 6 mmol/L, denoting a shift towards anaerobic metabolism, reflective of tissue hypoxia and metabolic distress, potentially spurred by impaired perfusion or exacerbated by the cancer's metabolic demands.

The administration of critical care medications like Dopamine(ranked 21), and Norepinphrine (ranked 28) used for sedation or seizure control, cardiac function supporting, managing blood pressure and cardiac output underscore the severity of the patient's condition and the complex interplay of various systems affected by metastatic cancer. In summary, these figures and associated data highlighted by our model showcase its ability to discern keys features associated with the multifaceted and severe clinical challenges in managing

a patient with metastatic cancer.

Fig. 7.11(a) provide an in-depth visual analysis of the critical clinical features observed in patients who are positively diagnosed with a hematologic malignancy, exploiting Feature Set B. Through a detailed review presented in Fig. 7.11(c), which maps out the patient's health evolution during their hospitalization, our model not only boosts interpretability but also precisely emphasizes those features most closely associated with hematologic malignancy.

Notably, the presence of severe cardiovascular instability, indicated by a systolic blood pressure(ranked 9) reading less than 70, a diastolic(ranked 1st) reading less than 25, and a mean arterial pressure(ranked 11) under 40, alongside a heart rate(ranked 25) exceeding 120, suggests significant circulatory compromise. This situation is further complicated by hematologic malignancy's impact on the cardiovascular system. A respiration rate(ranked 6) increasing from 18 to 40 breaths per minute highlights respiratory distress, potentially exacerbated by the malignancy's effects on lung function or related complications. Additionally, a marked decrease in the Glasgow Coma Scale (GCS) score from 15 to 6 points towards significant neurological implications, including altered mental status.

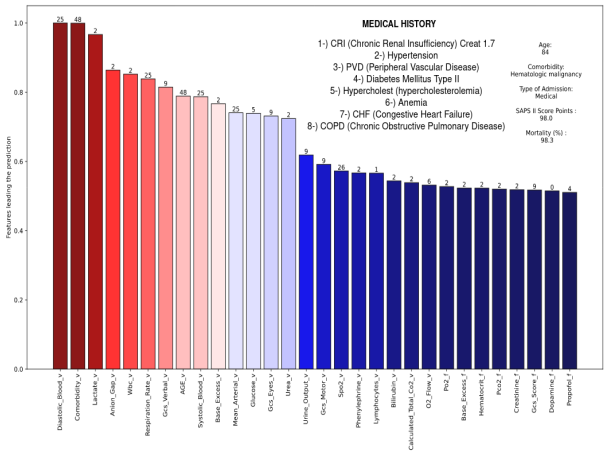
The observation of critical metabolic disturbances, such as a base excess(ranked 10) moving from -2.5 to -7.5 and an increasing lactate level(ranked 3) from 2 to 3, underscores potential metabolic acidosis. These indicators, together with an elevated anion gap(ranked 4) from 18 to 19 and a glucose level(ranked 12) of 250, reflect the complex interplay between hematologic malignancy and metabolic imbalances.

Hematologic parameters reveal the underlying bone marrow involvement characteristic of hematologic malignancies. A white blood cell (WBC)(ranked 5) count of 24, alongside the presence of promyelocytes(ranked 28) (1%), and lymphocytes(ranked 19) value of 4.4, suggests significant bone marrow disruption and often a sign of blood cancer. Furthermore, deteriorating kidney function, evidenced by rising urea(ranked 14) from 67 to 71 and bilirubin(ranked 20) from 18 to 20, along with reduced urine output(ranked 15) highlights the renal complications often seen in patients with hematologic malignancies.

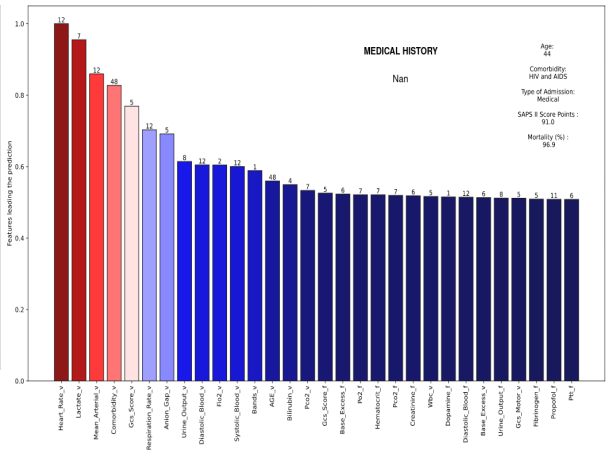
This detailed examination is critical for patients with hematologic malignancy, compounded by a comprehensive medical history of hypertension, chronic renal insufficiency (Creat 1.7), peripheral vascular disease, diabetes, superficial femoral artery (SFA) stenosis claudication, hypercholesterolemia, aortic stenosis, anemia, congestive heart failure (CHF) Class, and chronic obstructive pulmonary disease (COPD). The administration of vasopressors such as Phenylephrine and Propofol, further underscores the need for meticulous management of these complex comorbid conditions.

Fig. 7.11(b) comprehensively illustrate the key features identified in a positive patient grappling with a comorbidity of HIV-related comorbidities (AIDS), leveraging insights from features set B. By delving into the details presented in Fig. 7.11(d), illustrating the patient's health trajectory during the hospital stay, our model enhances interpretability, offering a nuanced view of critical features specifically associated with HIV-related comorbidities.

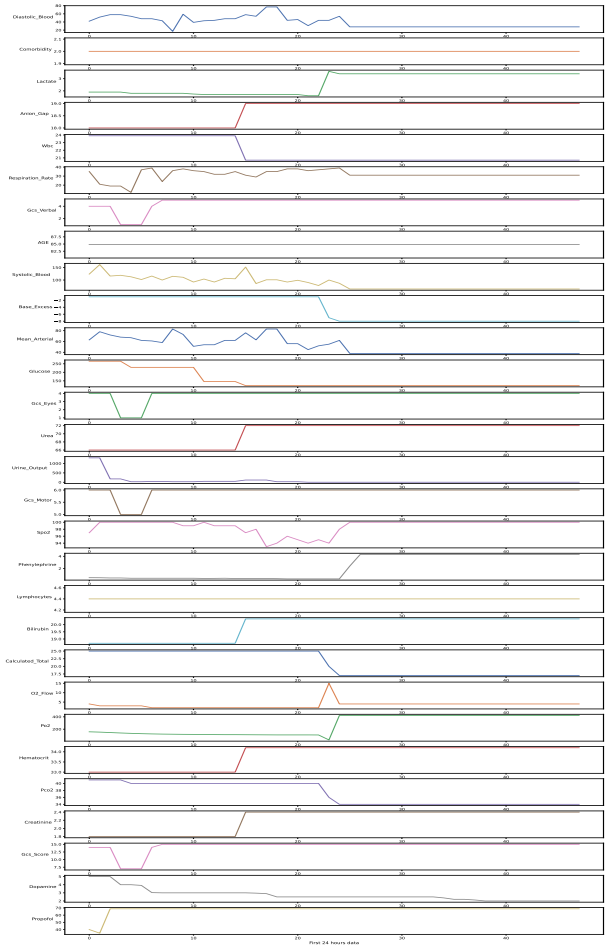
These highlighted features, including a Heart Rate(ranked 1st) below 50, an escalating lactate level(ranked 2) from 6 to 10, Mean Arterial Pressure(ranked 3) less than 45, a GCS



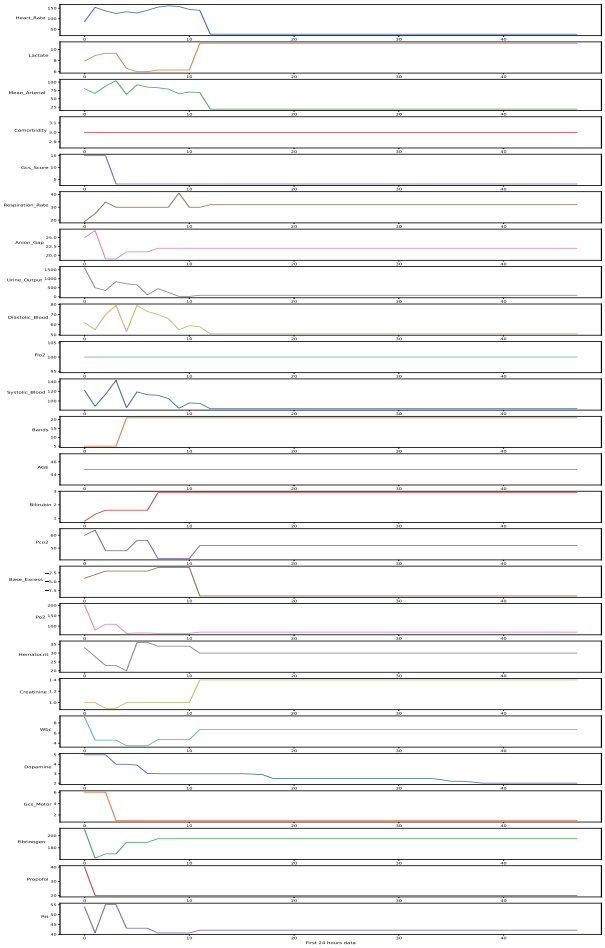
(a) The first 30 most importants features.



(b) The first 30 most importants features.



(c) Patient's health trajectory during his hospitalization.



(d) Patient's health trajectory during his hospitalization.

Figure 7.11: Two Positive patients with comorbidity (Hematologic malignancy and AIDS) on ICU Mortality task.

Score(ranked 5) of 3 indicating altered mental status, Respiration Rate(ranked 5) of 30, Diastolic Blood(ranked 9) Reading below 50, and Systolic Blood(ranked 11) Reading below 80, collectively point towards significant cardiovascular and respiratory challenges.

Additionally, an escalating bilirubin level(ranked 14) from 1 to 3 and a Urea(ranked 19) level of 12 indicating compromised kidney function, common in the intricate dynamics of HIV-related comorbidities. A WBC(ranked 14) value less than 3 signals a low white blood cell count, potentially heightening the risk of infections, which is particularly relevant for individuals with HIV-related comorbidities.

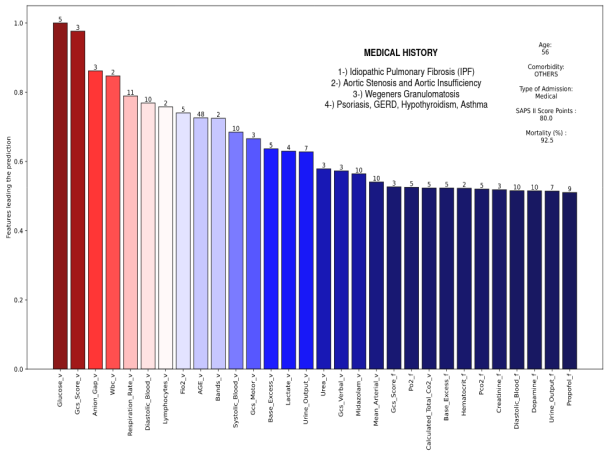
Furthermore, a Base Excess(ranked 15) level of -7.5, a calculated Total CO2(ranked 24) value greater than 25, and the presence of Promyelocytes(ranked 29) are indicative of potential metabolic imbalances and bone marrow irregularities. The elevated bands(ranked 12) value from 10 to 20 suggests an increased presence of immature neutrophils, potentially indicating an acute infection or inflammatory process. These transparent indicators contribute to a more thorough understanding of the patient's condition.

Moreover, the figures emphasize a SpO2(ranked 21) value less than 60%, highlighting severe hypoxemia, a critical consideration in managing individuals with HIV-related comorbidities. The presence of Promyelocytes(ranked 29) is also visually emphasized, offering clarity regarding bone marrow irregularities typical in this comorbidity.

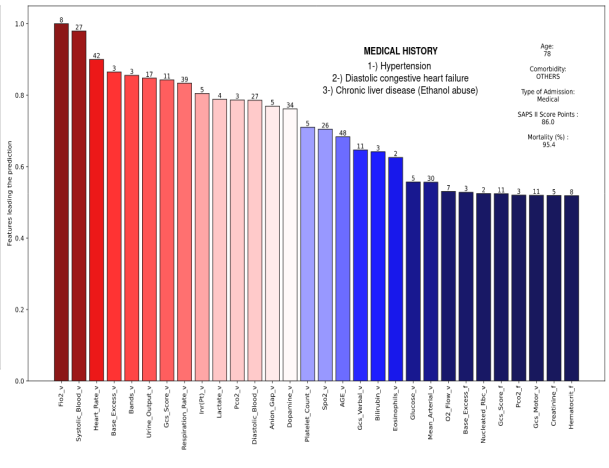
This interplay is particularly relevant in the context of administering vasopressors like Dopamine and Propofol. The intricacies underscore the dynamic relationship between HIV-related comorbidities and their impact on vital health parameters.

Upon detailed examination of Fig. 7.12(a), we unravel the complex health profile of a patient facing a range of comorbidities, including idiopathic pulmonary fibrosis, aortic stenosis, aortic insufficiency, Wegeners granulomatosis, psoriasis, GERD, hypothyroidism, and asthma. By leveraging insights from Fig. 7.12(c), which traces the patient's health trajectory, our model sheds light on key features that are intricately associated with these diverse medical conditions. This analysis vividly illustrates the critical physiological trends that demand immediate medical attention.

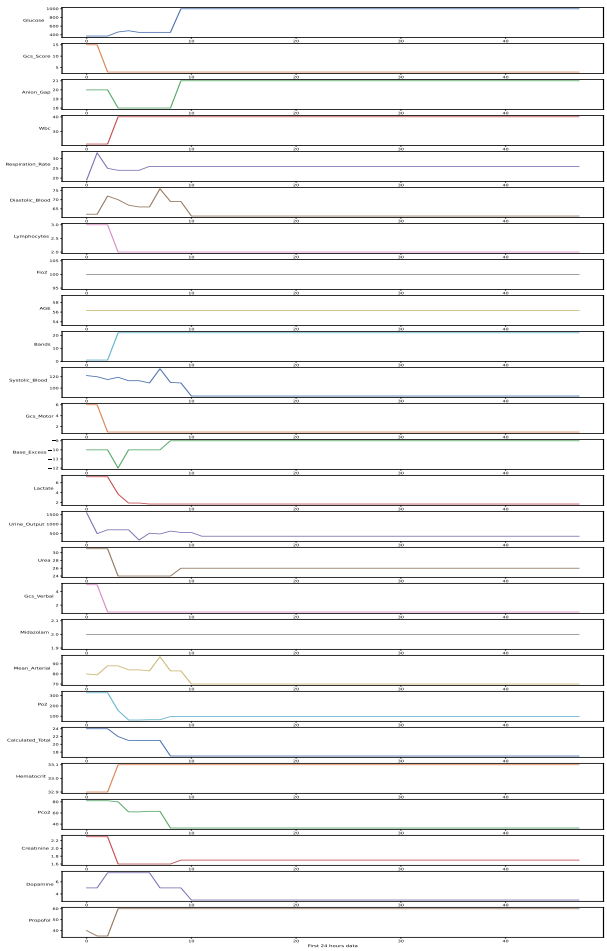
The marked escalation in glucose(ranked 1st) levels from 400 to 1000 mg/dL alongside a doubling of the WBC(ranked 4) count from 20 to 40 x $10^3\mu$L indicates severe hyperglycemia and suggests metabolic dysfunction, potentially intensified by the patient's underlying health issues. Moreover, a dramatic drop in the Glasgow Coma Scale (GCS)(ranked 2) score from 15 to 3 signals substantial neurological impairment, underscoring the gravity of the patient's state. Additionally, the anion gap(ranked 3) widening to 21, coupled with the significant increase in white blood cell count(ranked 4), suggests a profound systemic inflammatory response, likely exacerbated by inflammatory conditions such as Wegeners granulomatosis and psoriasis. Notable respiratory distress, evidenced by an increase in respiratory rate(ranked 5) from 18 to over 25 breaths per minute, and a maintained FiO2(ranked 8) at 100%, along with a critically high PCO2(ranked 25) of 80 mmHg and Lactate(ranked 14) level at 4 , points to severe respiratory acidosis and extreme hypoxemia. This condition is possibly due to hypoventilation or compromised gas exchange, further complicated by the patient's pulmonary fibrosis and potential acute respiratory distress syndrome (ARDS). Simultaneously, the observed drop in diastolic(ranked 6) , systolic blood(ranked 11) pressures to below 60 and 80 mmHg and Mean arterial(ranked 19) below 70, respectively, indicates significant cardiovascular instability, likely aggravated by the patient's aortic conditions. An elevated bands level over 20% indicates a critical inflammatory response. Concurrently, the



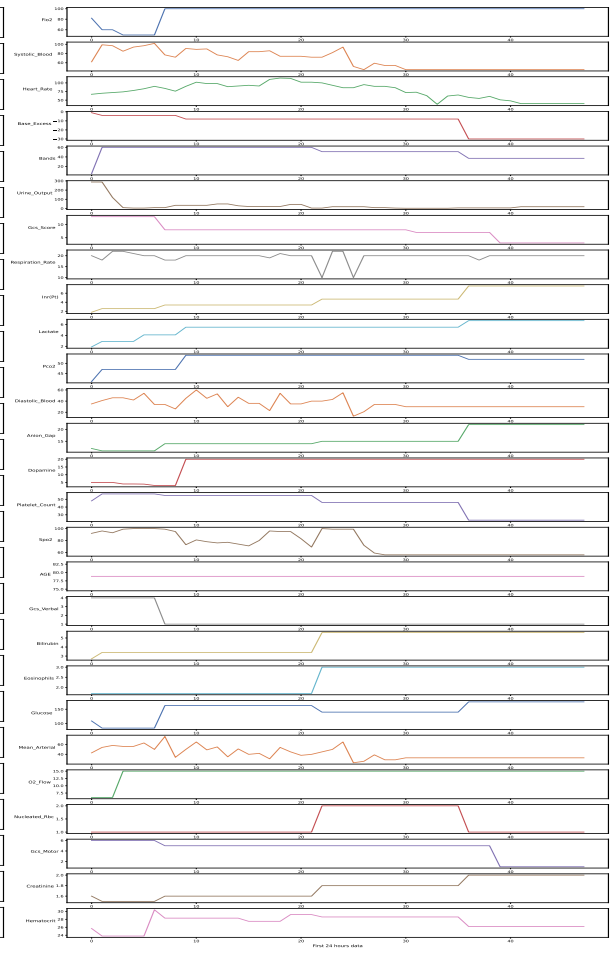
(a) Positive Patient.



(b) Positive Patient.



(c) Patient's health trajectory during his/her hospitalization.



(d) Patient's health trajectory during his/her hospitalization.

Figure 7.12: Two Positive patients on ICU Mortality task.

decline in base excess from -12 to -9 and total CO2 from 24 to 16 mmol/L signals developing metabolic acidosis, which may be compounded by renal issues related to hypothyroidism as reflected by Urea(ranked 16) level above 20 and a decreasing in Urine Output(ranked 15) from 1500 to less than 500. The noted hematocrit level of 33% and a PCO2 of 80 mmHg high-

light significant hematological and respiratory issues that require swift and comprehensive medical intervention to manage the patient's complex health dynamics effectively.

In the analysis of the medical data depicted in Fig. 7.12(b), we encounter the profound health challenges of a patient burdened with hypertension, diastolic congestive heart failure, and chronic liver disease. The progression of these conditions is meticulously chronicled in Fig. 7.12(d), which outlines the patient's health trajectory during his/her hospitalization. This comprehensive overview allows for an in-depth understanding of the complex interplay between the patient's multiple comorbidities and their overarching impact on health.

A critical observation is the escalated therapeutic oxygen requirement, marked by an FiO2(ranked 1st) increase from 50% to 100%. This dramatic adjustment is indicative of an acute effort to mitigate severe hypoxemia, underlining the gravity of the patient's respiratory distress. Concurrently, a stark reduction in both systolic(ranked 2) and diastolic blood(ranked 12) pressures to levels below 50 mmHg and 25 mmHg, respectively alongside a mean arterial pressure(ranked 22) decrease to below 40 mmHg, signals an alarming state of cardiovascular instability. For a patient with a history of hypertension and heart failure, these findings suggest the imminent threat of cardiovascular collapse. The patient's condition is further complicated by a heart rate(ranked 3) deceleration to below 40 beats per minute, indicating a severe bradycardia with potential for life-threatening arrhythmic disturbances. Moreover, a surge in Base Excess(ranked 4) to -30, an expanded Anion Gap(ranked 13) to 21, and a lactate(ranked 10) elevation to levels above 6 mmol/L collectively point to a critical state of severe metabolic acidosis. This condition reflects a significant disruption in the patient's acid-base homeostasis, exacerbated by their chronic liver disease and heart failure, demanding immediate and targeted intervention. Notably, an increase in bands(ranked 5) from 10 to 60 and a Urine Output(ranked 6) decline to nearly zero, alongside a pronounced rise in INR (Pt)(ranked 9) from 2 to above 6, herald the onset of acute renal failure and a heightened risk of disseminated intravascular coagulation (DIC). Such developments are particularly concerning, given the patient's complex medical background. The deterioration to a Glasgow Coma Scale (GCS)(ranked 7) score of 3 further emphasizes the severity of neurological compromise, reflecting the cumulative adverse effects of the patient's declining circulatory and metabolic state. Additionally, critical indicators such as a plummet in SpO2(ranked 16) to below 60%, a Platelet Count(ranked 15) decrease to less than 25, a Bilirubin(ranked 19) level ascent to above 3, and a Creatinine(ranked 29) level elevation to 2, coupled with a Hematocrit(ranked 30) level falling below 26, underscore the onset of multi-organ dysfunction syndrome (MODS). This dire condition, characterized by the concurrent failure of multiple organ systems, underscores the profound impact of the patient's heart failure, hypertension, and liver disease on their overall physiological functioning. The patient's clinical presentation reveals a constellation of critical indicators pointing towards acute cardiovascular collapse, severe metabolic and respiratory disturbances, and the looming threat of multi-organ dysfunction.

Fig. 7.13(a) and Fig. 7.13(b) reveal the critical features identified for two individual patients in the context of the hospital mortality task, utilizing the mean pooling technique. In particular, Fig. 7.13(a) highlights the significant features contributing to the prediction of mortality for the first patient, while Fig. 7.13(b) illustrates the key determinants for the

second patient. By applying mean pooling, the figures effectively summarize the influence of each feature, thus providing clear insights into the factors that most impact the prediction outcomes for these patients. This visualization underscores the model's ability to capture and interpret crucial information from patient data, enhancing the understanding of individual risk factors in hospital mortality.

We can analyze the data on an individual patient basis. It becomes evident that the attributes considered important can differ significantly from one patient to another. This raises the question: Can we identify attributes that consistently hold more importance across the entire patient cohort?

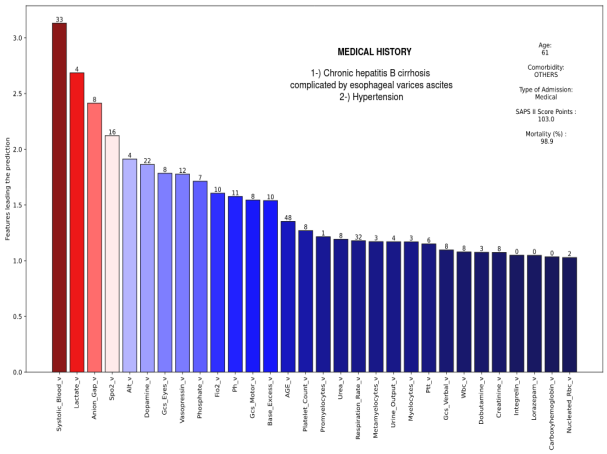
Fig. 7.14 offers a detailed look at the pivotal factors in predicting ICU mortality. It reports the average impact and standard deviation of each feature(attention weight), encompassing both vital signs and laboratory variables, on the model's predictions across all patients(10 testing folds). This analysis reveals the weight of various clinical metrics which includes both the values and their frequencies, on 108 input features, as outlined in Table 7.20a and Table 7.21a. Below, we give an unified description of some critical features as depicted in the Fig. 7.14.

Systolic Blood Pressure is highlighted as a key determinant, signaling cardiovascular stability or warning of potential issues. Heart Rate is also emphasized, serving as an indicator of cardiac health and stress levels. The Glasgow Coma Scale (GCS), focusing specifically on the Eyes, Motor, and Verbal components, is prominently featured. These metrics evaluate consciousness levels and cognitive functioning, where lower scores may indicate severe neurological impairments, marking them as key indicators of a patient's neurological health. Urine Output is noted for its relevance in evaluating renal health and fluid balance, critical in clinical settings. Blood pH, Base Excess and the Anion Gap are underscored for their roles in assessing metabolic and respiratory health, crucial in diagnosing acid-base imbalances.

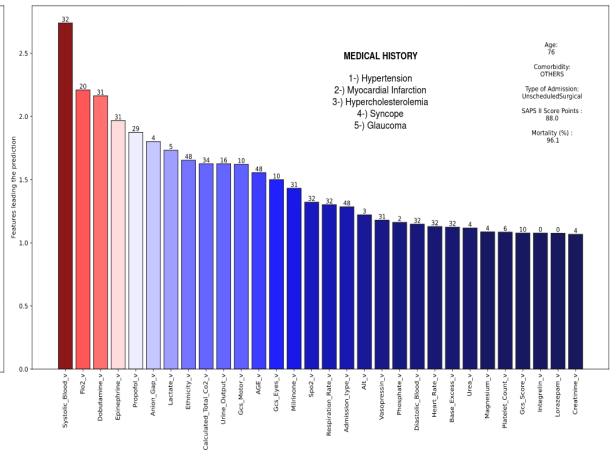
Age is presented as a factor affecting disease risk and recovery, with implications for patient care. Respiration Rate's inclusion highlights its importance in respiratory assessment, while Platelet Count is shown to be vital for understanding coagulation and bleeding risks. Lastly, SpO2 is focused on for its critical role in monitoring oxygen saturation and respiratory effectiveness. Together, these parameters provide a comprehensive framework for evaluating ICU mortality risk, guiding interventions and patient management strategies effectively. Nonetheless, it's important to note that this method does not conclusively establish that the foremost attribute is significantly more critical than the subsequent one.

7.5.2 Ranking of critical features through pairwise comparisons (P-value)

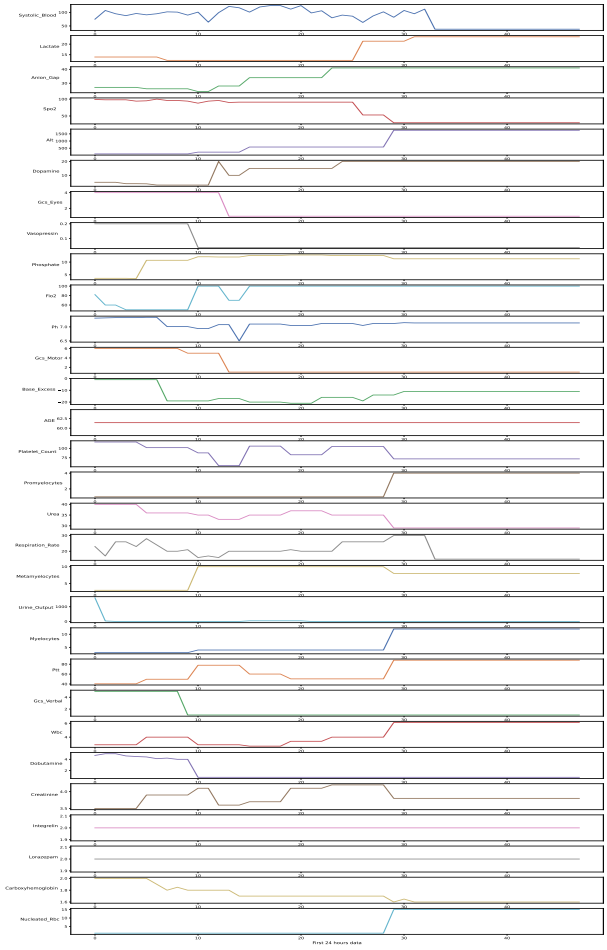
In order to ascertain the relative importance of each feature within the entire dataset spectrum, as depicted in the Fig. 7.14, we conducted pairwise comparisons for each feature against every other feature within the dataset, meticulously analyzing both the attention and frequency weights. At the heart of our methodology lies the utilization of p-value



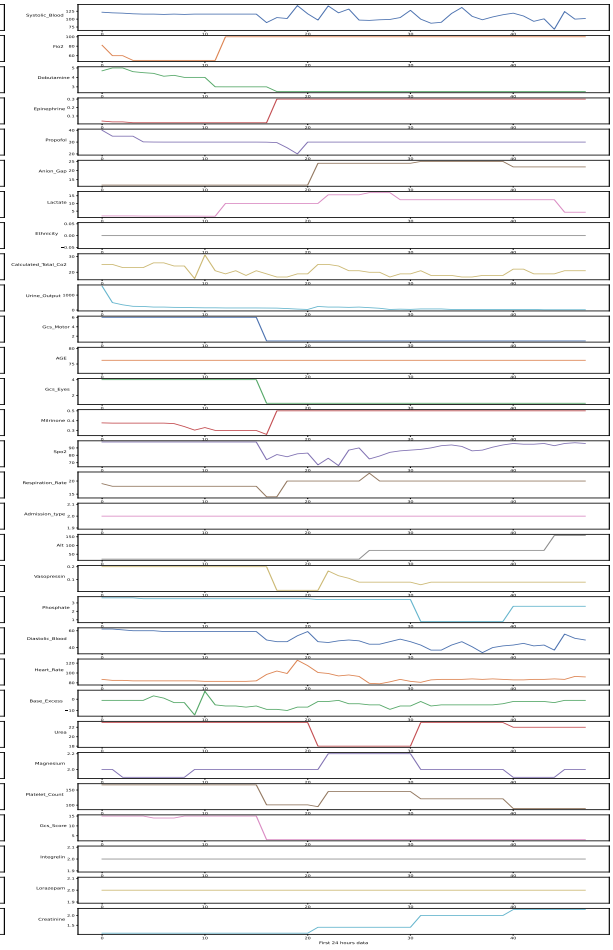
(a) Positive Patient.



(b) Positive Patient.



(c) Patient's health trajectory during his/her stay.



(d) Patient's health trajectory during his/her stay.

Figure 7.13: Two Positive patients on HOSPITAL Mortality task with Mean Pooling.

calculations to determine statistical significance. The p-value serves as a statistical metric aiding in the evaluation of whether the observed differences in weights (both attention and frequency) between two features are statistically meaningful. In our analytical process, a low p-value, typically below 0.05, signifies that the weight disparities (either attention or

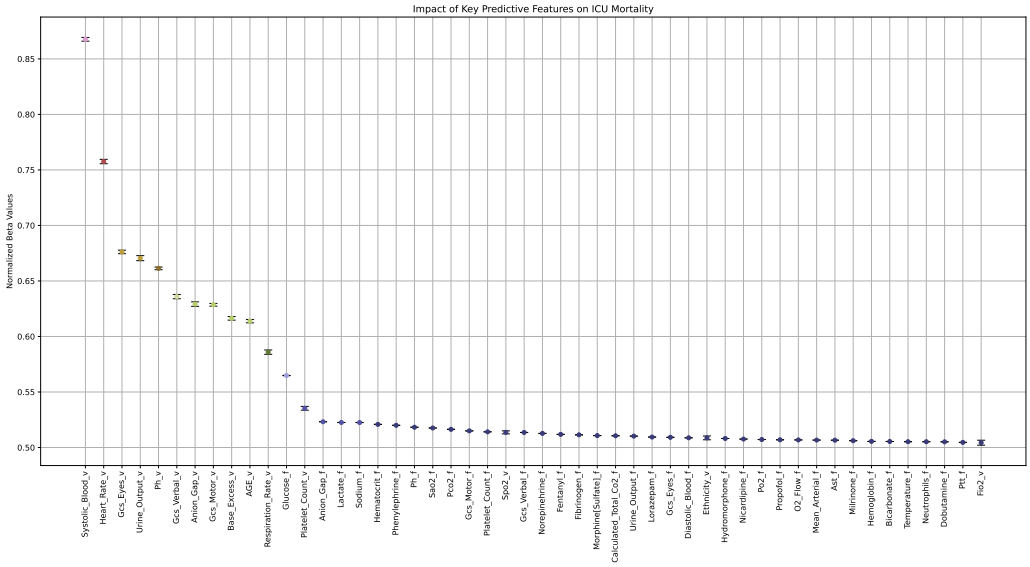


Figure 7.14: This figure displays the feature importance, ranked by their significance in predicting ICU mortality. It shows the average impact and standard deviation of each feature(attention weight), encompassing both vital signs and laboratory variables, on the model's predictions across all patients(10 testing folds).

Table 7.20: Number of measurements recorded for each feature within the first 48 hours of data collection and Aggregate function used for each feature based on experts knowledge.

(a) Lab features

Feature	No Patients	FreqsOfRecords	Agg(Min&Max&Sum)
Albumin	29246	42519	Min
Alt	32977	59015	Max
Amylase	17469	24961	Max
Anion Gap	52696	190957	Max
Ast	32954	58985	Max
Atypical Lymphocytes	3917	4524	Max
Bands	9148	12212	Min
Base Excess	34170	157063	Max
Basophils	31340	39141	Max
Bicarbonate	52867	200379	Min
Bilirubin	32643	58890	Max
Calcium Total	49690	154616	Max
Calculated Total Co2	36346	194509	Min
Carboxyhemoglobin	1509	1556	Max
Chloride	52917	219342	Max
Creatinine	52929	202324	Max
D-Dimer	1321	1518	Max
Eosinophils	32998	42678	Max
Epithelial Cells	3689	3871	Max
Ferritin	5803	6004	Max
Fibrinogen	15895	26194	Min
Fio2	30616	286696	Max
Glucose	53020	568605	Max
Granulocyte Count	1079	1562	Max
Height	22141	23928	Max
Hematocrit	52970	266941	Min
Hemoglobin	52946	217050	Min
Inr(Pt)	50336	147274	Max
Ketone	5937	6563	Max
Lactate	35388	98458	Max
Lipase	19373	26994	Max
Lymphocytes	37828	54289	Min
Magnesium	52043	178746	Min

(b) (Next)

Feature	No Patients	FreqsOfRecords	Agg(Min&Max&Sum)
Mch	52874	188748	Min
Mchc	52877	188857	Min
Mcv	52874	188745	Min
Metamyelocytes	4576	5905	Max
Methemoglobin	798	863	Max
Monocytes	37583	52980	Max
Myelocytes	2832	3463	Max
Neutrophils	37634	52717	Max
Nucleated Rbc	2051	2645	Max
O2 Flow	39262	288972	Max
Pco2	36344	194504	Min
Ph	47211	238489	Min
Phosphate	49715	155180	Max
Platelet Count	52914	198079	Min
Po2	36349	194513	Min
Potassium	52961	278704	Min
Promyelocytes	533	616	Max
Protein Total	2595	2704	Max
Ptt	50167	151074	Max
Rbc	52875	188751	Max
Required O2	8824	13646	Max
Reticulocyte Count	3704	3965	Max
Sao2	22440	78802	Min
Sodium	52941	233329	Min
Specific Gravity	35259	44289	Max
Thrombin	251	268	Max
Transferrin	6172	6376	Max
Troponin T	15796	36078	Max
Urea	52930	201647	Max
Urobilinogen	10095	11443	Max
Wbc	52892	189749	Min
Weight	47253	83340	Max

Table 7.21: Number of measurements recorded for each feature within the first 48 hours of data collection and Aggregate function used for each feature based on experts knowledge.

(a) Vital signs & Others

Feature	No Patients	FreqsOfRecords	Agg(Min&Max&Sum)
Diastolic Blood	52353	1831849	Min
Heart Rate	52357	1882472	Min
Mean Arterial	52353	1823948	Min
Respiration Rate	52327	1851377	Max
Spo2	52312	1805748	Min
Systolic Blood	52340	1820894	Min
Temperature	52352	689125	Max
Gcs Eyes	52268	634627	Min
Gcs Motor	52255	631933	Min
Gcs Verbal	52258	632431	Min
Urine Output	49717	1227457	Sum

(b) Drugs features

Feature	No Patients	FreqsOfRecords	Agg(Min&Max&Sum)
Argatroban	94	1041	Max
Diltiazem	966	9088	Max
Dobutamine	750	10515	Max
Dopamine	2921	34936	Max
Epinephrine	1687	18212	Max
Esmolol	497	5356	Max
Fentanyl	6039	87676	Max
Furosemide	947	9467	Max
Hydromorphone	95	1666	Max
Integrelin	1264	14692	Max
Labetalol	953	10125	Max
Lorazepam	563	12715	Max
Midazolam	4463	58134	Max
Milrinone	1167	20860	Max
Morphine[Sulfate]	1086	9941	Max
Natrecor	196	3697	Max
Nicardipine	976	9381	Max
Nitroglycerine	6816	65538	Max
Nitroprusside	1764	25050	Max
Norepinephrine	6210	89642	Max
Phenylephrine	9983	117947	Max
Precedex	929	4977	Max
Propofol	17438	197103	Max
Vancomycin Level	6348	9177	Max
Vasopressin	1641	19940	Max

frequency) between the compared features are statistically significant, suggesting that one feature exerts a more pronounced influence on the model's predictions than the other.

Taking Systolic Blood Pressure as a prime example, as illustrated in Table 7.22, we juxtaposed its attention weight against the attention and frequency weights of the remaining 107 predictors, including its own frequency weight. The evident statistical significance of Systolic Blood Pressure's attention weight, underscored by low p-values and its predominant statistical significance compared to the attention weight of all other features and their frequency weight with a total dominance of 215 wins over other features, highlights its critical role. This significance showcases its impact on the model's decision-making process to underscore its clinical relevance in forecasting ICU mortality.

This rigorous analysis highlights the top 50 features, as detailed in Table 7.22 identifying those particularly critical for ICU mortality prediction. It brings to light several key variables integral to ICU severity scores, like SAPS II and SOFA, and essential indicators such as Systolic Blood Pressure and Heart Rate. These indicators, along with metrics like Urine Output, Blood pH, Base Excess, Age, Respiration Rate, Platelet Count, and SpO2, noted for their high attention values, are crucial physiological metrics that can quickly change due to critical health events. Their significant attention values highlight their immediate relevance in predicting ICU mortality, serving as key indicators of a patient's current health status and necessitating swift intervention. The heightened attention values assigned to these parameters also enhance the model's proficiency in detecting subtle yet significant variations in vital signs and physiological states linked to increased mortality risks.

Conversely, frequency weights for features such as Glucose, Anion Gap, Lactate, Sodium, Hematocrit, Phenylephrine, Blood pH, Oxygen Saturation (SaO2), Carbon Dioxide Pressure (Pco2), and components of GCS Motor and Verbal, Platelet Count, and Norepinephrine use shed light on the patient's ongoing care and condition over time. Continuous monitoring of metabolic and electrolyte statuses, indicated by regular assessments of Glucose, Anion Gap, Lactate, and Sodium, is imperative for managing long-standing conditions and averting complications.

For instance, the consistent application of vasopressors like Phenylephrine and Norepinephrine highlights the critical nature of the patient's state. These medications are often deployed to stabilize blood pressure in severely ill individuals, with their frequent use signaling the management of conditions such as circulatory shock or hypotension. Additionally, frequency weights provide a panoramic view of patient care, emphasizing the necessity of perpetual monitoring and management for chronic or underlying conditions that, although not immediately life-threatening, are crucial for the patient's stabilization and survival prospects. The attentive adjustment of Glucose levels, for example, might signify the handling of diabetic complications or the strict regulation of glucose in critically ill patients to sidestep negative outcomes. Moreover, tracking parameters like Oxygen Saturation (SaO2), PaO2, hemoglobin, Carbon Dioxide Pressure (Pco2), Diastolic Blood Pressure, Oxygen Flow, Mean Arterial Pressure, Hemoglobin, Bicarbonate, Temperature, Neutrophils, and Partial Thromboplastin Time (Ptt) offers insights into the patient's respiratory, cardiovascular, and hemodynamic status, alongside their reactions to interventions like mechanical ventilation, vasopressors, sedatives, and blood transfusions.

This exhaustive process of comparison and analysis enables us to prioritize features based on their significance, illuminating the clinical and physiological factors that greatly influence patient outcomes in critical care. By spotlighting features that exhibited the highest levels of statistical significance across pairwise comparisons, we unveiled a roster of crucial determinants for predicting ICU mortality. These findings offer profound insights into the clinical and physiological factors that profoundly influence patient outcomes in critical care settings. The distinction in importance between some features' attention values over others' frequency weights also sheds light on the complex and dynamic nature of ICU care, where both immediate responses to acute alterations and sustained management of chronic conditions are essential. Meanwhile, the significance of monitoring clinical parameter illuminates care standards and ongoing management strategies, essential for a comprehensive understanding of a patient's health status, treatment requirements, and illness progression. In summary, attention values and frequency weights provide comprehensive insights into patient care within ICU settings, underscoring the complementary nature of acute interventions and long-term management in critical care. For Length of Stay (LOS) Task, the first 30 most important features are also shown in Table 7.23.

7.5.3 Causal inference explanations

To validate our findings, as illustrated in Fig. 7.14, regarding the effects of treatment variables on mortality prediction, we employed the DoWhy framework [142] to estimate causal effects from the input data. For a thorough analysis, we selected two first treatments, as

depicted in Fig. 7.14, Norepinephrine and Phenylephrine. Norepinephrine, the first-line vasopressor for shock and other critical conditions, is frequently associated with high mortality in the medical literature. Phenylephrine, on the other hand, is mainly used to raise blood pressure in the perioperative period. As it is less potent than norepinephrine, it is therefore less frequently associated with high mortality [15, 37, 117]. By analyzing the interactions of these medications, which are often associated with high mortality rates, we aimed to understand their impact on ICU mortality outcomes.

Key Components of the Causal Model are described below:

- Common Causes: These variables could confound the relationship between the treatment variables (Phenylephrine & Norepinephrine) and mortality.
- Instruments: These are variables that influence the treatment but do not directly impact the outcome, except through their effect on the treatment. They are used to address unmeasured confounding in the analysis. In this model, the instruments include variables such as Dobutamine, Fentanyl, Vancomycin level, Dopamine, Epinephrine, Vasopressin, Morphine [Sulfate], Phenylephrine, among others.
- Effect Modifiers: These variables might modify the effect of the treatment on the outcome. Variables such as Age, Diastolic Blood, Gcs Score, Heart Rate, Respiration Rate, Systolic Blood, Lactate, Comorbidity, Admission type, Bilirubin, Creatinine, Urine Output, Glucose, Sodium, Potassium, pH, Wbc, Urea, etc., are included as effect modifiers. The conditional estimates explore how Phenylephrine's effect and Norepinephrine's effect on Mortality might differ across different subgroups of patients, providing insights into potential effect modification.
- Estimand Type (NONPARAMETRIC ATE): This estimand is the Average Treatment Effect (ATE), estimated non-parametrically. This means the model estimates the causal effect of Norepinephrine on Mortality without assuming a specific functional form for the relationship.

The Causal Model provides estimated causal effects of 0.0543 for Phenylephrine and 0.3175 for Norepinephrine. This suggests that, on average, an increase in Phenylephrine is associated with an 5.43% increase in the probability of mortality, while an increase in Norepinephrine is associated with a 31.75% increase, after controlling for the included variables.

7.5.3.1 Refutation Results

Refutation methods are techniques used to validate the robustness and credibility of causal effect estimates. These methods test the assumptions and stability of the causal model by introducing perturbations to the data or the analysis. The goal is to check if the causal effect estimate holds under various scenarios designed to challenge its validity or if it fails a refutation test (typically indicated by a p-value < 0.05).

- Use of a Placebo Treatment: This refutation method checks if the estimated causal effect of the treatment (Phenylephrine & Norepinephrine) might be an artifact of the

model or data by randomly assigning a covariate as the treatment. If the assumptions are correct, this placebo treatment should have no real effect on the outcome, and the estimated causal effect should be close to zero.

This refutation results in a new estimated effect of 2e-04 with a p-value of 0.37 for Phenylephrine, and 5.94e-05 with a p-value of 0.49 for Norepinephrine. The significant difference between the original effects of both Phenylephrine (0.0543) and Norepinephrine (0.3175) compared to the placebo effects, which are close to zero, combined with high p-values of 0.37 and 0.49, indicates that the original effect estimates are likely not due to random chance. The placebo treatment's negligible effect close to zero(0), supports the conclusion that the observed effects of Phenylephrine and Norepinephrine are genuine and not artifacts of the data or model.

- Add a Random Common Cause: This method involves adding a randomly drawn covariate to the data and re-running the analysis. If the original causal effect estimate is robust, it should not change significantly when random covariates are added, as these random covariates should not have a genuine causal impact on the treatment or outcome.

The effect remains nearly unchanged for both Phenylephrine (mean value 0.0543) and Norepinephrine (mean value 0.3175), with high p-values of 0.96 and 0.98, respectively. This suggests that adding a random common cause does not significantly impact the causal estimates, reinforcing the robustness of the original findings. These results likely provide a reliable reflection of the causal relationship between Phenylephrine and mortality, as well as Norepinephrine and mortality.

7.5.3.2 Conclusion

The causal model offers a strong and reliable estimate of the effects of both Phenylephrine and Norepinephrine on ICU mortality. By accounting for a wide range of confounding variables and validating the findings through refutation tests, the model reveals a significant association between Norepinephrine use and increased mortality as unveiled in Section 2.2.4, thereby corroborating the findings of Hugerot and al [127]. Additionally, it indicates a slight increase in mortality with the use of Phenylephrine, which aligns with established clinical protocols.

Key points from this analysis include:

- Significant Covariates: The large number of covariates in the model highlights the complexity of the ICU setting and the numerous factors influencing both the treatment and the outcome.
- Robustness of Results: The refutation tests (placebo and random common cause) indicate that the causal estimate is robust and not easily swayed by arbitrary changes, adding confidence to the causal inference.

- Positive Effect of treatment variable: The positive average treatment effect suggests a potential harmful effect of both Phenylephrine and Norepinephrine on ICU mortality, necessitating careful interpretation and validation against clinical guidelines and expert opinions.

Table 7.22: Highlights the feature importance, ranked by their significance in predicting *ICU mortality*. It shows the sum impact of each feature won in rapport to the attention value and frequency weight for the comparing features, After performing p-value pairwise comparisons for all the 216 inputs features for our entire cohort (attention value and frequency weights of each feature) over the first 48-hours data using max pooling.

Feature "SS: Statistically Significant"	Sum SS/Feature(Att Value + Freq Weight)	SS/Feature Attention Value	SS/Feature Frequency Weight	No Patients(At leat one recorded)	FreqsOfRecords
Systolic_Blood_v	215	107	108	52340	1820894
Heart_Rate_v	214	106	108	52357	1882472
Gcs_Eyes_v	213	105	108	52268	634627
Urine_Output_v	212	104	108	49717	1227457
Ph_v	211	103	108	47211	238489
Gcs_Verbal_v	210	102	108	52258	632431
Anion_Gap_v	208	100	108	52696	190957
Gcs_Motor_v	208	100	108	52255	631933
Base_Excess_v	207	99	108	34170	157063
AGE_v	206	98	108	-	-
Respiration_Rate_v	205	97	108	52327	1851377
Glucose_f	204	97	107	53020	568605
Platelet_Count_v	203	96	107	52914	198079
Anion_Gap_f	202	96	106	52696	190957
Lactate_f	201	96	105	35388	98458
Sodium_f	200	96	104	52941	233329
Hematocrit_f	199	96	103	52970	266941
Phenylephrine_f	198	96	102	9983	117947
Ph_f	197	96	101	47211	238489
Sao2_f	196	96	100	22440	78802
Pco2_f	195	96	99	36344	194504
Gcs_Motor_f	194	96	98	52255	631933
Platelet_Count_f	192	95	97	52914	198079
Gcs_Verbal_f	191	95	96	52258	632431
Spo2_v	190	95	95	52312	1805748
Norepinehrine_f	190	95	95	6210	89642
Fentanyl_f	189	95	94	6039	87676
Fibrinogen_f	188	95	93	15895	26194
Morphine[Sulfate]_f	187	95	92	1086	9941
Calculated_Total_Co2_f	186	95	91	36346	194509
Urine_Output_f	185	95	90	49717	1227457
Lorazepam_f	183	94	89	563	12715
Gcs_Eyes_f	182	94	88	52268	634627
Diastolic_Blood_f	181	94	87	52353	1831849
Hydromorphone_f	180	94	86	95	1666
Ethnicity_v	179	94	85	-	-
Nicardipine_f	179	94	85	976	9381
Po2_f	178	94	84	36349	194513
Propofol_f	177	94	83	17438	197103
O2_Flow_f	176	94	82	39262	288972
Mean_Arterial_f	175	94	81	52353	1823948
Ast_f	174	94	80	32954	58985
Milrinone_f	173	94	79	1167	20860
Hemoglobin_f	171	93	78	52946	217050
Bicarbonate_f	170	93	77	52867	200379
Temperature_f	169	93	76	52352	689125
Neutrophils_f	168	93	75	37634	52717
Dobutamine_f	167	93	74	750	10515
Ptt_f	166	93	73	50167	151074
Urea_f	165	93	72	52930	201647

7.6 MWTA-LSTM ASP vs AMITA

In this section, we compare the performance of MWTA-LSTM ASP and AMITA across various clinical prediction tasks. We analyze their effectiveness in handling mortality prediction (In-hospital & ICU) on the MIMIC III and eICU datasets, and evaluate their accuracy in predicting Length of Stay (LOS). Detailed comparisons, including F1-scores, MAE, and RMSE metrics, highlight the strengths and weaknesses of each model, providing a comprehensive assessment of their capabilities.

Table 7.23: Highlights the feature importance, ranked by their significance in predicting *Length of Stay (LOS)*. It shows the sum impact of each feature won in rapport to the attention value and frequency weight for the comparing features, After performing p-value pairwise comparisons for all the 216 inputs features for our entire cohort (attention value and frequency weights of each feature) over the first 48-hours data using mean pooling.

Feature "SS: Statistically Significant"	Sum SS/Feature(Att Value + Freq Weight)	SS/Feature Attention Value	SS/Feature Frequency Weight	No Patients(At leat one recorded)	FreqsOfRecords
Admission_type_v	215	107	108	-	-
Diastolic_Blood_v	214	106	108	52353	1831849
Respiration_Rate_v	213	105	108	52327	1851377
Weight_v	212	104	108	47253	83340
Gcs_Verbal_v	211	103	108	52258	632431
Hematocrit_v	210	102	108	52970	266941
Platelet_Count_v	209	101	108	52914	198079
Glucose_v	208	100	108	53020	568605
Fibrinogen_v	207	99	108	15895	26194
Temperature_v	206	98	108	52352	689125
Pao2/Fio2_v	205	97	108	25681	87154
Heart_Rate_v	204	96	108	52357	1882472
Lactate_v	203	95	108	35388	98458
Vasopressin_v	202	94	108	1641	19940
Sao2_v	201	93	108	22440	78802
Gcs_Motor_v	200	92	108	52255	631933
Gcs_Score_v	199	91	108	52242	628743
Mchc_v	198	90	108	52877	188857
Bicarbonate_v	194	86	108	52867	200379
Hemoglobin_v	194	86	108	52946	217050
Phenylephrine_v	194	86	108	9983	117947
Morphine[Sulfate]_v	193	85	108	1086	9941
Mch_v	193	85	108	52874	188748
Chloride_v	192	84	108	52312	1805748
Mcv_v	191	83	108	52874	188745
Rbc_v	190	82	108	52875	188751
Systolic_Blood_v	188	80	108	52340	1820894
Reticulocyte_Count_v	188	80	108	3704	3965
Calculated_Total_Co2_v	187	79	108	36346	194509

7.6.1 Mortalities tasks & Length of stay

Both of our approaches achieve remarkable results in mortality prediction tasks (in-hospital and ICU) across both datasets (MIMIC III and eICU). However, when comparing the two models, we observed that AMITA outperforms MWTA-LSTM on MIMIC for all mortality prediction metrics, with an F1-score exceeding 2%, as shown in Table 7.24. This is noteworthy given that MWTA-LSTM incorporates additional components such as Gating and ASP layers.

AMITA's superiority can be attributed to its ability to effectively handle timing irregularities in the patient's medical history. By distinguishing between short-term and long-term memory contributions of clinical features, AMITA integrates contextual information, measurement frequency, and elapsed times into the cell state. This nuanced approach allows AMITA to accurately capture short-term impacts, where frequent measurements are crucial, and long-term influences, where less frequent measurements still hold significant clinical relevance. In contrast, MWTA-LSTM relies solely on elapsed times to manage timing irregularities, limiting its capacity to retain important earlier data, especially in cases where infrequent measurements remain clinically significant, such as in renoprotective strategies. On the eICU dataset, we observed similar trends, although MWTA-LSTM performed slightly better than AMITA, as detailed in Table 7.24.

In the context of the Length of Stay (LOS) prediction task, detailed error metrics, including MAE and RMSE, are summarized in Table 7.25. Remarkably, AMITA yielded better RMSE and MAE values of less than one day on both datasets, specifically 20.88 and 13.2 hours for MIMIC and 19.44 and 12.24 hours for eICU, respectively, compared to the ground

truth data, with an Adjusted R^2 of 74% when using the first 48 hours of data. In contrast, MWTA-LSTM achieved a MAE of 1.21 days and an RMSE of 3.27 days on MIMIC, and a MAE of 2.69 days and an RMSE of 5.84 days on eICU.

Fig. 7.15, Fig. 7.16 and Fig. 7.17 present the results of both approaches on both datasets using all features within the first 48 hours data as inputs.

Table 7.24: MWTA-LSTM vs AMITA on In-hospital & ICU using MIMIC III & eICU.

DATA	TASKS	MODELS	SAPS II FEATURES			ALL FEATURES		
			AUROC	AUPRC	F1-SCORE	AUROC	AUPRC	F1-SCORE
24 HRS DATA	In-Hospital	**AMITA on MIMIC**	0.89±0.0073	0.66±0.0191	**0.60±0.0137**	0.92±0.0067	0.72±0.0159	**0.65±0.0163**
		MWTA-LSTM on MIMIC	0.89±0.0017	0.68±0.0075	0.60±0.0084	0.91±0.0097	0.70±0.0120	0.62±0.0100
		AMITA on eICU	0.91±0.0030	0.63±0.0134	**0.60±0.0127**	0.93±0.0055	0.68±0.0090	0.63±0.0098
		MWTA-LSTM on eICU	0.91±0.0008	0.62±0.0032	0.57±0.0087	0.91±0.0023	0.68±0.0011	**0.63±0.0030**
	ICU	**AMITA on MIMIC**	0.92±0.0075	0.66±0.0215	**0.60±0.0174**	0.94±0.0045	0.74±0.0183	**0.65±0.0164**
		MWTA-LSTM on MIMIC	0.91±0.0092	0.63±0.0178	0.57±0.0245	0.93±0.0071	0.68±0.0205	0.62±0.0108
		AMITA on eICU	0.93±0.0211	0.68±0.0217	**0.63±0.0102**	0.95±0.0055	0.71±0.0081	0.67±0.0154
		MWTA-LSTM on eICU	0.93±0.0053	0.65±0.0024	0.61±0.0030	0.93±0.0008	0.71±0.0052	**0.67±0.0050**
48 HRS DATA	In-Hospital	**AMITA on MIMIC**	0.92±0.0225	0.73±0.0147	**0.65±0.0014**	0.93±0.0066	0.77±0.0132	**0.67±0.0122**
		MWTA-LSTM on MIMIC	0.90±0.0075	0.69±0.0317	0.62±0.0271	0.92±0.0128	0.72±0.0010	0.65±0.0102
		AMITA on eICU	0.93±0.0144	0.68±0.0094	**0.65±0.0177**	0.93±0.0252	0.72±0.0098	0.67±0.0168
		MWTA-LSTM on eICU	0.92±0.0008	0.67±0.0132	0.62±0.0237	0.94±0.0073	0.75±0.0131	**0.68±0.0070**
	ICU	**AMITA on MIMIC**	0.94±0.0155	0.75±0.0147	**0.67±0.0114**	0.95±0.0042	0.78±0.0175	**0.70±0.0181**
		MWTA-LSTM on MIMIC	0.93±0.0071	0.69±0.0165	0.64±0.0187	0.94±0.0045	0.73±0.0183	0.66±0.0164
		AMITA on eICU	0.94±0.0071	0.71±0.0017	**0.68±0.0102**	0.95±0.0079	0.77±0.0086	0.71±0.0153
		MWTA-LSTM on eICU	0.94±0.0088	0.74±0.0240	0.68±0.0251	0.96±0.0068	0.80±0.0040	**0.73±0.0423**

Table 7.25: MWTA-LSTM vs AMITA on LOS using MIMIC III & eICU.

DATA USED	MODELS	SAPS II FEATURES		ALL FEATURES	
		RMSE	MAE	RMSE	MAE
24 HRS DATA	**AMITA on MIMIC**	**1.23±0.0134**	**0.90±0.0112**	**1.17±0.0117**	**0.86±0.0075**
	MWTA-LSTM on MIMIC	4.57±0.1308	2.08±0.0387	3.68±0.1712	1.57±0.0128
	AMITA on eICU	**1.18±0.0261**	**0.87±0.0131**	**1.13±0.0159**	**0.85±0.0137**
	MWTA-LSTM on eICU	6.01±0.1536	3.16±0.0632	5.89±0.1926	3.06±0.0342
48 HRS DATA	**AMITA on MIMIC**	**0.89±0.0207**	**0.57±0.0168**	**0.87±0.0192**	**0.55±0.0112**
	MWTA-LSTM on MIMIC	4.03±0.1722	1.37±0.0216	3.27±0.1038	1.21±0.0186
	AMITA on eICU	**0.86±0.0266**	**0.58±0.0138**	**0.81±0.0163**	**0.51±0.0137**
	MWTA-LSTM on eICU	5.88±0.1764	2.92±0.0345	5.84±0.1732	2.69±0.0126

7.6.2 Runtime comparison

Both models (MWTA & AMITA) were evaluated on an NVIDIA GeForce GTX 1080 Ti GPU with 10 GB of total memory and a system RAM of 30 GB. During training, MWTA-LSTM takes 164 seconds per epoch while AMITA takes 170 seconds per epoch. Both achieves an inference speed of 5.52 milliseconds per prediction, demonstrating their suitability for real-time applications. Their performance are comparable to other state-of-the-art methods, indicating that MWTA & AMITA are efficient in both training and inference phases. All models were implemented using PyTorch 1.13.0.

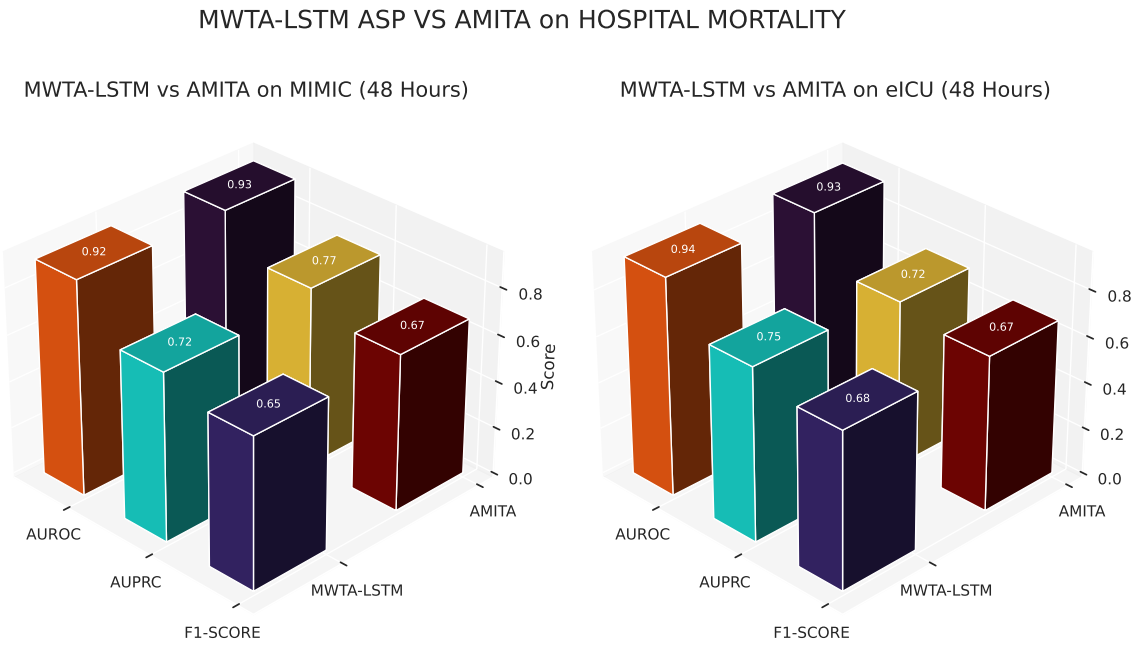


Figure 7.15: MWTA-LSTM ASP VS AMITA on HOSPITAL using ALL FEATURES (MIMIC III & eICU).

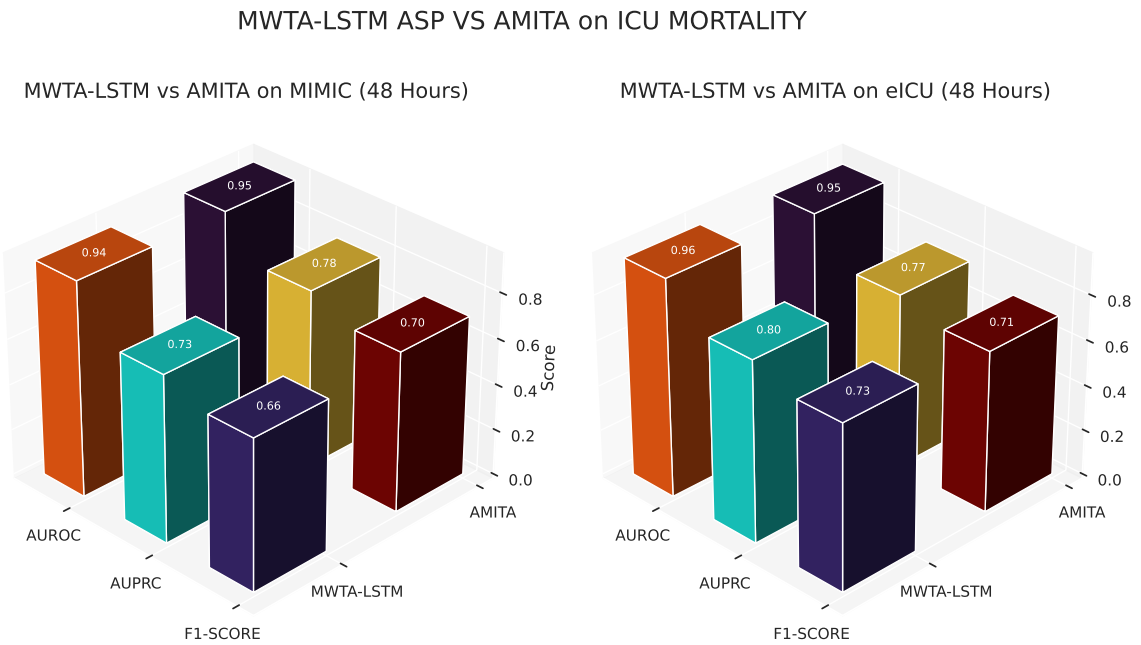


Figure 7.16: MWTA-LSTM ASP VS AMITA on ICU using ALL FEATURES (MIMIC III & eICU).

7.6.3 Closing remarks

In summary, both AMITA and MWTA-LSTM demonstrate strong performance in mortality prediction tasks (In-hospital & ICU) across the MIMIC III and eICU datasets. However, a

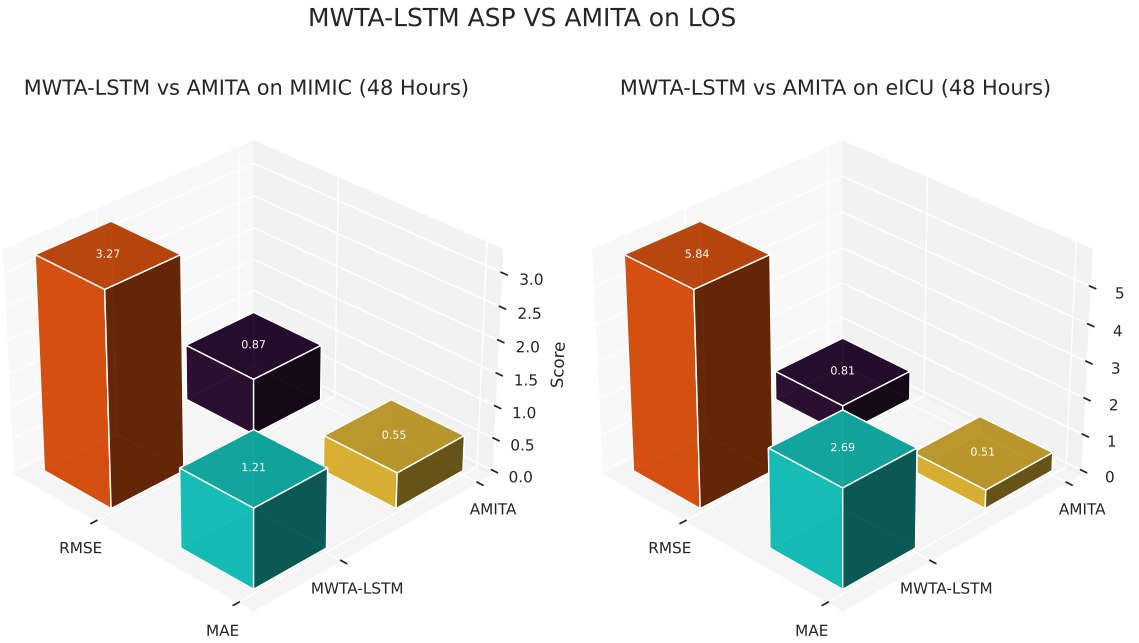


Figure 7.17: MWTA-LSTM ASP VS AMITA on LOS using ALL FEATURES (MIMIC III & eICU).

closer examination reveals distinct strengths and weaknesses for each model.

MWTA-LSTM is designed with additional components such as Gating and ASP layers, which contribute to its robust handling of complex medical data. However, it relies primarily on elapsed times to manage timing irregularities, limiting its ability to retain significant earlier data. This makes it less effective in scenarios where the infrequent measurement of some clinical parameters is crucial. In contrast, AMITA effectively handles timing irregularities by incorporating contextual information, measurement frequency, and elapsed times into its model. This leads to superior performance in both short-term and long-term memory retention.

Overall, AMITA's sophisticated approach to managing timing irregularities and its superior performance in most tasks highlight its potential as a more advanced model for clinical predictions. Nevertheless, MWTA-LSTM's robust components and slight edge in specific datasets suggest it remains valuable in particular contexts. Future work could explore combining the strengths of both approaches to develop an even more powerful predictive model for clinical applications.

Chapter 8

Conclusion & Perspectives

"Stories don't end", he says. "They just turn into new beginnings."

– Lindsay Eagar, Hour of the Bees

In this section, we summarize the key findings of our study and explore potential directions for future research. Our approach has shown significant promise, addressing critical challenges and opening up new possibilities for enhancing predictive performance through the integration of clinical notes and physiological variables. We also discuss the potential for real-time predictions and advanced techniques for handling missing data, highlighting the future scope of our work.

8.1 Contribution

In this thesis, we present three significant contributions as listed below:

MWTA-LSTM and AMITA mark significant advancements in personalized healthcare, distinguishing themselves from existing models such as [5, 24, 124, 171]. Unlike its predecessors MWTA-LSTM and AMITA handle irregular sampling more effectively using elapsed time, frequency of measurement, and the last observation, thereby addressing the issue of timing irregularity.

Firstly, MWTA-LSTM operates as an end-to-end deep dynamic memory neural network. It autonomously extracts relevant features from medical records by integrating time decay mechanisms and parameterization, effectively managing the temporal dynamics of healthcare data. This adaptability allows MWTA-LSTM to handle irregular timing of events and observations more efficiently. Furthermore, we introduced an innovative adaptive pooling mechanism into MWTA-LSTM to manage outliers in patient records, ensuring robust inference on future outcomes even with noisy or irregular data points, distinguishing it from existing models [5, 124, 171].

Secondly, we introduce AMITA, a novel deep dynamic memory neural network that extends LSTM's capabilities by incorporating time intervals between events, measurement frequency, and patient context. This refinement enables nuanced memory cell adjustments, reducing the influence of distant memories while retaining significant past events. Dual gate mechanisms further enhance AMITA's ability to capture the impact of clinical interventions on current illness states, with a refined forget gate that considers event timing, frequency, and broader historical context.

Lastly, an innovative approach to enhance interpretability through a comprehensive analysis incorporating attention values and frequency weights for each feature. This method identifies critical features relevant to prediction tasks such as ICU mortality and Length of Stay (LOS), ensuring transparency and insight into model predictions.

MWTA-LSTM and AMITA have demonstrated superior performance in predicting healthcare outcomes such as mortality and length of stay across diverse datasets. Comparative evaluations underscore their effectiveness compared to state-of-the-art models [130], as detailed in Table 7.9.

In summary, MWTA-LSTM and AMITA represent significant advancements in personalized healthcare, leveraging deep learning to extract relevant features autonomously, manage temporal dynamics, handle outliers, and enhance interpretability. These innovations

promise to advance predictive personalized medicine substantially.

8.2 Perspectives

Currently, predictions are made exactly 24 or 48 hours after ICU admission, but there is significant potential in transitioning towards real-time predictions, ideally updated hourly or after each medical intervention. Real-time predictions could dramatically enhance the responsiveness and precision of clinical interventions, enabling healthcare providers to adjust treatments dynamically as a patient's condition evolves. While the current approach provides a structured and operationally feasible framework for monitoring patients, it comes with notable limitations.

One major challenge in implementing real-time predictions is the variability in patient data across ICU stays. A single patient may experience multiple ICU admissions during a single hospital stay, each characterized by different health conditions and varying levels of monitoring. This results in a highly non-uniform distribution of clinical parameters across ICU stays, as some admissions require more intensive monitoring than others. Sequential models, such as LSTM, GRU even MWTA-LSTM or AMITA, are typically designed to handle data with uniform distribution, expecting input at each time step, as they operate under the assumption of homogeneity in data collection. However, ICU data is inherently sporadic, as measurements may be taken at unpredictable intervals, driven by a patient's rapidly changing condition or the availability of healthcare personnel. This lack of regularity presents a fundamental challenge, as the current models struggle to process or learn from such non-uniform data patterns, which can lead to degraded performance or missed clinical insights. The complexity of patient trajectories within the ICU exacerbates this issue further. Different stages of illness, responses to treatment, or complications can all influence the frequency and type of data collected, meaning that even within a single patient's ICU stay, the data distribution can shift dramatically. This variability calls for more sophisticated modeling approaches that can not only accommodate the non-linear and uneven nature of ICU data but also leverage it to make more accurate, context-aware predictions.

This inherent limitation in sequential models highlights the need to explore alternative approaches. First, how can we develop real-time prediction models that account for varying distributions of clinical data, especially when a patient experiences multiple ICU stays within the same hospital admission? Splitting the data into training and testing sets under these conditions presents a challenge, as the distribution of parameters across stays can differ greatly. Second, how can we ensure that real-time predictions—whether updated hourly or following each intervention—do not lead to alarm fatigue? This phenomenon, where the sheer volume of alerts overwhelms clinicians, risks reducing the overall effectiveness of real-time monitoring.

Additionally, another primary challenge in working with Electronic Health Record (EHR) data is the irregularity in the recording of clinical parameters. Unlike standard time series datasets that feature uniform intervals between observations, EHR data is characterized by uneven time stamps. Different clinical variables such as heart rate, blood pressure, and lab test results are measured at varying intervals based on patient needs or resource availabil-

ity. This variability creates gaps in the data, introducing uncertainty into model predictions. Additionally, EHR data often suffers from sparsity, where many clinical parameters are infrequently recorded or entirely absent for certain patients. This inconsistency complicates the task for machine learning models, which struggle to build accurate representations of patient health. Furthermore, the inherent heterogeneity of healthcare processes means that different clinical measurements decay at varying rates, with some being more time-sensitive than others.

Although our current approach addresses timing irregularities and data sparsity in EHR datasets, it does not specifically aim to replace traditional missing value imputation methods. Missing values in clinical data often stem from logistical constraints, such as limited resources or specific clinical decisions. Our model primarily focuses on making robust predictions based on the available data rather than reconstructing the missing portions of the dataset. To handle missing values, we employ straightforward imputation techniques, including mean and median imputation, forward and backward filling. While these methods offer a quick fix, they can introduce inaccuracies by failing to consider the underlying clinical context or variability inherent in the data. Consequently, our predictions may be compromised by incomplete or inaccurately imputed data, hindering the model's ability to accurately capture the true trajectory of a patient's condition.

To address these limitations, our methodology could be significantly enhanced by leveraging advanced deep learning architectures capable of simultaneously managing missing value imputation and outcome prediction. Techniques such as multi-task learning could enable the model to learn how to impute missing values while predicting clinical outcomes concurrently. This approach involves training the model on two interrelated tasks: one focused on estimating the missing values and the other on predicting outcomes. By integrating these tasks, the model can improve its overall accuracy; as it learns to predict clinical outcomes such as ICU mortality and also becomes adept at filling in missing data in a clinically relevant manner. This ensures that the imputed values are not arbitrary but rather learned representations from the data that directly contribute to the accuracy of the primary task. This is particularly advantageous for EHR data, where the absence of critical variables can significantly skew results.

In contrast to simple imputation methods that rely on averages or heuristic approaches, advanced architectures such as variational autoencoders (VAEs) and recurrent neural networks (LSTMs) offer significant advantages when trained on multi-task objectives. These models can infer missing values by identifying underlying patterns in the data, taking into account both temporal dependencies and clinical correlations among variables. As a result, they generate more accurate imputed values that closely align with real-world patient trajectories.

By utilizing imputed values derived from deep learning models, we can also mitigate the bias and noise commonly associated with traditional imputation methods. Concurrently learning these tasks allows us to leverage the interdependence between accurate imputation and precise predictions, resulting in more reliable and robust outcomes—especially for patients with highly irregular or sparse records. This approach is particularly suited to a variety of sequence modeling tasks.

The evaluation of our proposed models has so far been limited to retrospective observational studies, primarily due to the constraints posed by publicly available datasets. However, the practical implications of our methods go far beyond these limitations. By implementing our approaches in prospective observational studies, we could further validate the findings from retrospective analyses and ensure the models perform reliably in real-world clinical environments.

Nevertheless, prospective validation introduces unique challenges, particularly when applying A/B testing in ICU settings. One key issue is the heterogeneity of ICU patient populations, as patients present with a wide range of conditions that make it difficult to generalize model predictions across all cases. To address this, we would need to randomly assign patients to either Group A or Group B to minimize selection bias. Moreover, the trial should be conducted in a double-blinded manner, ensuring that clinical staff remain unaware of group assignments to prevent any unintentional influence on patient management and outcomes.

Another challenge is the non-uniformity of data distribution in ICU settings. The variability in ICU stays and clinical interventions leads to inconsistencies in the data, especially in real-time prediction scenarios. For example, vital signs may be measured at different frequencies based on a patient's condition. Some patients may have continuous monitoring, while others are assessed less frequently. This variability complicates real-time predictions and highlights the need to account for causality, as the relationship between the timing and frequency of data collection and a patient's health condition can introduce confounding factors that are not easily scalable across different cases. These potential confounding factors, where changes in patient status could influence both the treatment and the frequency of data collection, making it difficult to distinguish correlation from causality. Addressing these complexities is essential for improving the real-world applicability and generalizability of our models in prospective clinical trials and ICU decision-making.

In summary, while our models effectively tackle key challenges in EHR data by addressing timing irregularities and sparse records, there remains substantial potential for further development. Future directions could include real-time predictions, advanced imputation techniques, and validations through prospective studies, all of which would enhance the utility and impact of our approaches. Our research establishes a strong foundation for future innovations aimed at improving patient outcomes through more accurate and timely predictions.

Bibliography

[1] Julia Adler-Milstein and Ashish K Jha. "HITECH Act drove large gains in hospital electronic health record adoption." In: *Health Affairs, 36(8):1416–1422* (2017).

[2] Denis Agniel, Isaac S Kohane, and Griffin M Weber. "Biases in electronic health record data due to processes within the healthcare system: retrospective observational study". In: *Bmj* 361 (2018).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[4] Iyad Batal et al. "A temporal pattern mining approach for classifying electronic health record data". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 4.4 (2013), pp. 1–22.

[5] Inci M Baytas et al. "Patient subtyping via time-aware LSTM networks". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 65–74.

[6] Anat Reiner Benaim et al. "Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies". In: *JMIR medical informatics* 8.2 (2020), e16492.

[7] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2 (1994), pp. 157–166.

[8] Shishira Bharadwaj et al. "Malnutrition: laboratory markers vs nutritional assessment". In: *Gastroenterology report* (2016), gow013.

[9] Filippo Maria Bianchi, Lorenzo Livi, and Cesare Alippi. "Investigating echo-state networks dynamics by means of recurrence analysis". In: *IEEE transactions on neural networks and learning systems* 29.2 (2016), pp. 427–439.

[10] Filippo Maria Bianchi et al. "Multiplex visibility graphs to investigate recurrent neural network dynamics". In: *Scientific reports* 7.1 (2017), p. 44037.

[11] Christopher M Bishop. "Mixture density networks". In: (1994).

[12] Y-Lan Boureau, Jean Ponce, and Yann LeCun. "A theoretical analysis of feature pooling in visual recognition". In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 111–118.

[13] Kathleen M Brelsford, Susan E Spratt, and Laura M Beskow. "Research use of electronic health records: patients' perspectives on contact by researchers". In: *Journal of the American Medical Informatics Association* 25.9 (2018), pp. 1122–1129.

[14] Ao Buke et al. "Healthcare algorithms by wearable inertial sensors: a survey". In: *China Communications* 12.4 (2015), pp. 1–12.

[15] Maurizio Cecconi et al. "Consensus on circulatory shock and hemodynamic monitoring. Task force of the European Society of Intensive Care Medicine". In: *Intensive care medicine* 40 (2014), pp. 1795–1815.

[16] Basit Chaudhry et al. "Systematic review: impact of health information technology on quality, efficiency, and costs of medical care". In: *Annals of internal medicine* 144.10 (2006), pp. 742–752.

[17] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. "Special issue on learning from imbalanced data sets". In: *ACM SIGKDD explorations newsletter* 6.1 (2004), pp. 1–6.

[18] Zhengping Che et al. "Recurrent neural networks for multivariate time series with missing values". In: *Scientific reports* 8.1 (2018), p. 6085.

[19] Yu Cheng et al. "Risk prediction with electronic health records: A deep learning approach". In: *Proceedings of the 2016 SIAM international conference on data mining*. SIAM. 2016, pp. 432–440.

[20] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[21] Changkyu Choi. "Time series forecasting with recurrent neural networks in presence of missing data". MA thesis. UiT Norges arktiske universitet, 2018.

[22] Edward Choi et al. "Doctor ai: Predicting clinical events via recurrent neural networks". In: *Machine learning for healthcare conference*. PMLR. 2016, pp. 301–318.

[23] Edward Choi et al. "GRAM: graph-based attention model for healthcare representation learning". In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 787–795.

[24] Edward Choi et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in neural information processing systems* 29 (2016).

[25] Edward Choi et al. "Using recurrent neural network models for early detection of heart failure onset". In: *Journal of the American Medical Informatics Association* 24.2 (2017), pp. 361–370.

[26] Anna Choromanska et al. "The loss surfaces of multilayer networks". In: *Artificial intelligence and statistics*. PMLR. 2015, pp. 192–204.

[27] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[28] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)". In: *arXiv:1511.07289* (2015).

[29] European Commission. "European Commission. Overview of the national laws on electronic health records in the eu member states." In: *link*. (2017).

[30] Sajad Darabi et al. “Taper: Time-aware patient ehr representation”. In: *IEEE journal of biomedical and health informatics* 24.11 (2020), pp. 3268–3275.

[31] Yann N Dauphin et al. “Language modeling with gated convolutional networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 933–941.

[32] Misha Denil et al. “Learning where to attend with deep architectures for image tracking”. In: *Neural computation* 24.8 (2012), pp. 2151–2184.

[33] Rahul Dey and Fathi M Salem. “Gate-variants of gated recurrent unit (GRU) neural networks”. In: *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*. IEEE. 2017, pp. 1597–1600.

[34] Centers for Disease Control, Prevention, et al. “HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services”. In: *MMWR: Morbidity and mortality weekly report* 52.Suppl 1 (2003), pp. 1–17.

[35] Cristóbal Esteban et al. “Predicting clinical events by combining static and dynamic information using recurrent neural networks”. In: *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. Ieee. 2016, pp. 93–101.

[36] Cristóbal Esteban et al. “Predicting sequences of clinical events by using a personalized temporal latent embedding model”. In: *2015 International conference on healthcare informatics*. IEEE. 2015, pp. 130–139.

[37] Laura Evans et al. “Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021”. In: *Critical care medicine* 49.11 (2021), e1063–e1143.

[38] Yuchen Fan et al. “TTS synthesis with bidirectional LSTM based recurrent neural networks”. In: *Fifteenth annual conference of the international speech communication association*. 2014.

[39] Charles K Fisher, Aaron M Smith, and Jonathan R Walsh. “Machine learning for comprehensive forecasting of Alzheimer’s Disease progression”. In: *Scientific reports* 9.1 (2019), p. 13622.

[40] Milena A Gianfrancesco et al. “Potential biases in machine learning algorithms using electronic health record data”. In: *JAMA internal medicine* 178.11 (2018), pp. 1544–1547.

[41] P Girardet et al. “Scores de gravité en réanimation”. In: *Conférences d’actualisation du 41e Congrès national d’anesthésie et de réanimation. Paris: Elsevier-SFAR* (1999), pp. 659–678.

[42] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 315–323.

[43] Ary L Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23 (2000), e215–e220.

[44] Ian Goodfellow et al. "Maxout networks". In: *International conference on machine learning*. PMLR. 2013, pp. 1319–1327.

[45] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[46] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee. 2013, pp. 6645–6649.

[47] Alex Graves and Jürgen Schmidhuber. "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks* 18.5-6 (2005), pp. 602–610.

[48] Alex Graves et al. "A novel connectionist system for unconstrained handwriting recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 31.5 (2008), pp. 855–868.

[49] WJ Guan et al. "Unsupervised learning technique identifies bronchiectasis phenotypes with distinct clinical characteristics". In: *The International Journal of Tuberculosis and Lung Disease* 20.3 (2016), pp. 402–410.

[50] Tian Guo, Tao Lin, and Nino Antulov-Fantulin. "Exploring interpretable LSTM neural networks over multi-variable data". In: *International conference on machine learning*. PMLR. 2019, pp. 2494–2504.

[51] Robert M Hamer and Pippa M Simpson. *Last observation carried forward versus mixed models in the analysis of psychiatric clinical trials*. 2009.

[52] Hrayr Harutyunyan et al. "Multitask learning and benchmarking with clinical time series data". In: *Scientific data* 6.1 (2019), pp. 1–18.

[53] Moeen Hassanalieragh et al. "Health monitoring and management using Internet-of-Things (IoT) sensing with cloud-based processing: Opportunities and challenges". In: *2015 IEEE international conference on services computing*. IEEE. 2015, pp. 285–292.

[54] Milos Hauskrecht et al. "Outlier-based detection of unusual patient-management actions: an ICU study". In: *Journal of biomedical informatics* 64 (2016), pp. 211–221.

[55] Haibo He and Edwardo A Garcia. "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9 (2009), pp. 1263–1284.

[56] Kaiming He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[57] Kaiming He et al. "Identity mappings in deep residual networks". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 630–645.

[58] Tsipi Heart, Ofir Ben-Assuli, and Itamar Shabtai. "A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy". In: *Health Policy and Technology* 6.1 (2017), pp. 20–25.

[59] Georges Hebrail and Alice Berard. *Individual household electric power consumption*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58K54. 2012.

[60] JaWanna Henry et al. "Adoption of electronic health record systems among US non-federal acute care hospitals: 2008–2015". In: *ONC data brief* 35.35 (2016), pp. 2008–15.

[61] William R Hersh. "Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance". In: *Clin Pharmacol Ther* 81.126-128 (2007), p. 42.

[62] Sepp Hochreiter. "The vanishing gradient problem during learning recurrent neural nets and problem solutions". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6.02 (1998), pp. 107–116.

[63] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[64] Sepp Hochreiter et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

[65] Oliver Ibe. *Markov processes for stochastic modeling*. Newnes, 2013.

[66] Hassan Ismail Fawaz et al. "Deep learning for time series classification: a review". In: *Data mining and knowledge discovery* 33.4 (2019), pp. 917–963.

[67] Martin Jacobsen and Joseph Gani. "Point process theory and applications: marked point and piecewise deterministic processes". In: (2006).

[68] Abhyuday N Jagannatha and Hong Yu. "Bidirectional RNN for medical event detection in electronic health records". In: *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. Vol. 2016. NIH Public Access. 2016, p. 473.

[69] Ashish K Jha et al. "Use of electronic health records in US hospitals". In: *New England Journal of Medicine* 360.16 (2009), pp. 1628–1638.

[70] Alistair EW Johnson et al. "MIMIC-III, a freely accessible critical care database". In: *Scientific data* 3.1 (2016), pp. 1–9.

[71] Robert L Kane et al. "The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis". In: *Medical care* (2007), pp. 1195–1204.

[72] Arif Khwaja. "KDIGO clinical practice guidelines for acute kidney injury". In: *Nephron Clinical Practice* 120.4 (2012), pp. c179–c184.

[73] Mehmet Kılıç et al. "Cost analysis on Intensive Care Unit costs based on the length of stay". In: *Turkish journal of anaesthesiology and reanimation* 47.2 (2019), p. 142.

[74] John Frank Charles Kingman. *Poisson processes*. Vol. 3. Clarendon Press, 1992.

[75] William A Knaus et al. "APACHE II: a severity of disease classification system." In: *Critical care medicine* 13.10 (1985), pp. 818–829.

[76] Roopa Kohli-Seth and John M Oropello. "The future of bedside monitoring". In: *Critical care clinics* 16.4 (2000), pp. 557–578.

[77] Sunil Kripalani et al. "Reducing hospital readmission rates: current strategies and future directions". In: *Annual review of medicine* 65 (2014), pp. 471–485.

[78] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[79] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[80] Hugo Larochelle and Geoffrey E Hinton. "Learning to combine foveal glimpses with a third-order Boltzmann machine". In: *Advances in neural information processing systems* 23 (2010).

[81] Hugo Larochelle et al. "Exploring strategies for training deep neural networks." In: *Journal of machine learning research* 10.1 (2009).

[82] Thomas A Lasko, Joshua C Denny, and Mia A Levy. "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data". In: *PloS one* 8.6 (2013), e66341.

[83] Günter Last and Andreas Brandt. *Marked Point Processes on the real line: the dynamical approach*. Springer Science & Business Media, 1995.

[84] Günter Last and Mathew Penrose. *Lectures on the Poisson process*. Vol. 7. Cambridge University Press, 2017.

[85] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. "A new simplified acute physiology score (SAPS II) based on a European/North American multicenter study". In: *Jama* 270.24 (1993), pp. 2957–2963.

[86] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[87] Garam Lee et al. "Predicting Alzheimer's disease progression using multi-modal deep learning approach". In: *Scientific reports* 9.1 (2019), p. 1952.

[88] Jae Won Lee. "Stock price prediction using reinforcement learning". In: *ISIE 2001. 2001 IEEE International Symposium on Industrial Electronics Proceedings (Cat. No. 01TH8570)*. Vol. 1. IEEE. 2001, pp. 690–695.

[89] Jeong Min Lee and Milos Hauskrecht. "Modeling multivariate clinical event time-series with recurrent temporal mechanisms". In: *Artificial intelligence in medicine* 112 (2021), p. 102021.

[90] Jeong Min Lee and Milos Hauskrecht. "Recent context-aware lstm for clinical event time-series prediction". In: *Conference on Artificial Intelligence in Medicine in Europe*. Springer. 2019, pp. 13–23.

[91] Christophe Leys et al. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median". In: *Journal of experimental social psychology* 49.4 (2013), pp. 764–766.

[92] Pengpeng Li et al. "Bidirectional gated recurrent unit neural network for Chinese address element segmentation". In: *ISPRS International Journal of Geo-Information* 9.11 (2020), p. 635.

[93] Yang Li, Nan Du, and Samy Bengio. "Time-dependent representation for neural event sequence prediction". In: *arXiv preprint arXiv:1708.00065* (2017).

[94] Yikuan Li et al. "BEHRT: transformer for electronic health records". In: *Scientific reports* 10.1 (2020), p. 7155.

[95] Bryan Lim et al. "Temporal fusion transformers for interpretable multi-horizon time series forecasting". In: *International Journal of Forecasting* 37.4 (2021), pp. 1748–1764.

[96] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[97] Luchen Liu et al. "Learning hierarchical representations of electronic health records for clinical outcome prediction". In: *AMIA Annual Symposium Proceedings*. Vol. 2019. American Medical Informatics Association. 2019, p. 597.

[98] Siqi Liu and Milos Hauskrecht. "Nonparametric regressive point processes based on conditional gaussian processes". In: *Advances in Neural Information Processing Systems* 32 (2019).

[99] Zitao Liu and Milos Hauskrecht. "A regularized linear dynamical system framework for multivariate time series analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[100] Zitao Liu, Lei Wu, and Milos Hauskrecht. "Modeling clinical time series using gaussian process sequences". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM. 2013, pp. 623–631.

[101] Lorenzo Livi, Filippo Maria Bianchi, and Cesare Alippi. "Determination of the edge of criticality in echo state networks through Fisher information maximization". In: *IEEE transactions on neural networks and learning systems* 29.3 (2017), pp. 706–717.

[102] Junyu Luo et al. "Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 647–656.

[103] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. "Effective approaches to attention-based neural machine translation". In: *arXiv preprint arXiv:1508.04025* (2015).

[104] Liantao Ma et al. "Concare: Personalized clinical feature embedding via capturing the healthcare context". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 833–840.

[105] Tengfei Ma, Cao Xiao, and Fei Wang. "Health-atm: A deep architecture for multifaceted patient health record representation and risk prediction". In: *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM. 2018, pp. 261–269.

[106] Iain L MacDonald and Walter Zucchini. *Hidden Markov and other models for discrete-valued time series*. Vol. 110. CRC Press, 1997.

[107] Emmanuel Maggiori et al. "Recurrent neural networks to correct satellite image classification maps". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.9 (2017), pp. 4962–4971.

[108] Salim Malakouti and Milos Hauskrecht. "Hierarchical adaptive multi-task learning framework for patient diagnoses and diagnostic category classification". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 701–706.

[109] Benjamin M Marlin et al. "Unsupervised pattern discovery in electronic health care data using probabilistic clustering models". In: *Proceedings of the 2nd ACM SIGHIT international health informatics symposium*. 2012, pp. 389–398.

[110] CJ McKay et al. "High early mortality rate from acute pancreatitis in Scotland, 1984–1995". In: *British journal of surgery* 86.10 (1999), pp. 1302–1305.

[111] Eddie McKenzie. "Ch. 16. discrete variate time series". In: *Handbook of statistics* 21 (2003), pp. 573–606.

[112] Hongyuan Mei and Jason M Eisner. "The neural hawkes process: A neurally self-modulating multivariate point process". In: *Advances in neural information processing systems* 30 (2017).

[113] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013).

[114] Riccardo Miotto et al. "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records". In: *Scientific reports* 6.1 (2016), pp. 1–10.

[115] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. "Recurrent models of visual attention". In: *Advances in neural information processing systems* 27 (2014).

[116] Frank J Molnar, Brian Hutton, and Dean Fergusson. "Does analysis using "last observation carried forward" introduce bias in dementia research?" In: *Cmaj* 179.8 (2008), pp. 751–753.

[117] Andrea Morelli et al. "Phenylephrine versus norepinephrine for initial hemodynamic support of patients with septic shock: a randomized, controlled trial". In: *Critical care* 12 (2008), pp. 1–11.

[118] Ishna Neamatullah et al. “Automated de-identification of free-text medical records”. In: *BMC medical informatics and decision making* 8.1 (2008), pp. 1–17.

[119] Shamim Nemati et al. “An interpretable machine learning model for accurate prediction of sepsis in the ICU”. In: *Critical care medicine* 46.4 (2018), p. 547.

[120] Rajendran Nirthika et al. “Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study”. In: *Neural Computing and Applications* 34.7 (2022), pp. 5321–5347.

[121] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. “On the difficulty of training recurrent neural networks”. In: *International conference on machine learning*. PMLR. 2013, pp. 1310–1318.

[122] Chris Paxton, Alexandru Niculescu-Mizil, and Suchi Saria. “Developing predictive models using electronic medical records: challenges and pitfalls”. In: *AMIA Annual Symposium Proceedings*. Vol. 2013. American Medical Informatics Association. 2013, p. 1109.

[123] Trang Pham et al. “Faster training of very deep networks via p-norm gates”. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE. 2016, pp. 3542–3547.

[124] Trang Pham et al. “Predicting healthcare trajectories from medical records: A deep learning approach”. In: *Journal of biomedical informatics* 69 (2017), pp. 218–229.

[125] Tom J Pollard et al. “The eICU Collaborative Research Database, a freely available multi-center database for critical care research”. In: *Scientific data* 5.1 (2018), pp. 1–13.

[126] Michael I Posner and Steven E Petersen. “The attention system of the human brain”. In: *Annual review of neuroscience* 13.1 (1990), pp. 25–42.

[127] “Posologie maximale de noradrénaline des 24 premières heures d’hospitalisation en réanimation : étude de l’impact pronostique et détermination des facteurs prédictifs de succès du traitement”. In: (2021).

[128] Peter J Pronovost et al. “Physician staffing patterns and clinical outcomes in critically ill patients: a systematic review”. In: *Jama* 288.17 (2002), pp. 2151–2162.

[129] Sanjay Purushotham et al. “Benchmarking deep learning models on large healthcare datasets”. In: *Journal of biomedical informatics* 83 (2018), pp. 112–134.

[130] Yao Qin et al. “A dual-stage attention-based recurrent neural network for time series prediction”. In: *arXiv preprint arXiv:1704.02971* (2017).

[131] Lawrence R Rabiner. “A tutorial on hidden Markov models and selected applications in speech recognition”. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[132] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.

[133] Alvin Rajkomar et al. "Scalable and accurate deep learning with electronic health records". In: *NPJ digital medicine* 1.1 (2018), p. 18.

[134] Daniel Ramage. "Hidden Markov models fundamentals". In: *CS229 Section Notes* 1 (2007).

[135] Marian-Andrei Rizoiu et al. "Hawkes processes for events in social media". In: *Frontiers of multimedia research*. 2017, pp. 191–218.

[136] T Ruf. "The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series". In: *Biological Rhythm Research* 30.2 (1999), pp. 178–201.

[137] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[138] DE Rumelhart. "David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams Learning representations by back-propagating errors Nature 323: 533-536". In: *nature* 323 (1986), pp. 533–536.

[139] Jeffrey D Scargle. "Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data". In: *Astrophysical Journal, Part 1, vol. 263, Dec. 15, 1982, p. 835-853.* 263 (1982), pp. 835–853.

[140] Peter Schulam, Fredrick Wigley, and Suchi Saria. "Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1. 2015.

[141] Mike Schuster and Kuldip K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[142] Amit Sharma and Emre Kiciman. "DoWhy: An end-to-end library for causal inference". In: *arXiv preprint arXiv:2011.04216* (2020).

[143] Benjamin Shickel et al. "DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning". In: *Scientific reports* 9.1 (2019), p. 1879.

[144] Satya Narayan Shukla and Benjamin Marlin. "Multi-Time Attention Networks for Irregularly Sampled Time Series". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=4c0J6lwQ4_.

[145] Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos". In: *Advances in neural information processing systems* 27 (2014).

[146] Mervyn Singer et al. "The third international consensus definitions for sepsis and septic shock (Sepsis-3)". In: *Jama* 315.8 (2016), pp. 801–810.

[147] Huan Song et al. "Attend and diagnose: Clinical time series analysis using attention models". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[148] Gabriel de Souza Pereira Moreira et al. "Transformers4rec: Bridging the gap between nlp and sequential/session-based recommendation". In: *Proceedings of the 15th ACM Conference on Recommender Systems*. 2021, pp. 143–153.

[149] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. "Highway networks". In: *arXiv preprint arXiv:1505.00387* (2015).

[150] Ruslan Leont'evich Stratonovich. "Conditional markov processes". In: *Non-linear transformations of stochastic processes*. Elsevier, 1965, pp. 427–453.

[151] Chenxi Sun et al. "A review of deep learning methods for irregularly sampled medical time series data". In: *arXiv preprint arXiv:2010.12493* (2020).

[152] Ilya Sutskever. *Training recurrent neural networks*. University of Toronto Toronto, ON, Canada, 2013.

[153] Qingxiong Tan et al. "Data-gru: Dual-attention time-aware gated recurrent unit for irregular multivariate time series". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 930–937.

[154] Kristian Thygesen et al. "Fourth universal definition of myocardial infarction (2018)". In: *Circulation* 138.20 (2018), e618–e651.

[155] Sindhu Tipirneni and Chandan K Reddy. "Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 16.6 (2022), pp. 1–17.

[156] Sandeep Kumar Vashist. "Non-invasive glucose monitoring technology in diabetes management: A review". In: *Analytica chimica acta* 750 (2012), pp. 16–27.

[157] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[158] Petar Velickovic et al. "Deep graph infomax." In: *ICLR (Poster)* 2.3 (2019), p. 4.

[159] Marzia Verhaeghe and Saïd Hachimi-Idrissi. "Blood lactate and lactate kinetics as treatment and prognosis markers for tissue hypoperfusion". In: *Acta Clinica Belgica* 75.1 (2020), pp. 1–8.

[160] Ronald J Williams and Jing Peng. "An efficient gradient-based algorithm for online training of recurrent network trajectories". In: *Neural computation* 2.4 (1990), pp. 490–501.

[161] Ronald J Williams and David Zipser. "Gradient-based learning algorithms for recurrent". In: *Backpropagation: Theory, architectures, and applications* 433 (1995), p. 17.

[162] Cao Xiao, Edward Choi, and Jimeng Sun. "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review". In: *Journal of the American Medical Informatics Association* 25.10 (2018), pp. 1419–1428.

[163] Shuai Xiao et al. "Modeling the intensity function of point process via recurrent neural networks". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[164] Kelvin Xu et al. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057.

[165] Andrew Yale et al. "Generation and evaluation of privacy preserving synthetic health data". In: *Neurocomputing* 416 (2020), pp. 244–255.

[166] Dingjun Yu et al. "Mixed pooling for convolutional neural networks". In: *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*. Springer. 2014, pp. 364–375.

[167] Ke Yu et al. "Monitoring ICU mortality risk with a long short-term memory recurrent neural network". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*. World Scientific. 2019, pp. 103–114.

[168] Matthew D Zeiler and Rob Fergus. "Stochastic pooling for regularization of deep convolutional neural networks". In: *arXiv preprint arXiv:1301.3557* (2013).

[169] Jinghe Zhang et al. "Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record". In: *IEEE Access* 6 (2018), pp. 65333–65346.

[170] Xi Zhang et al. "Data-driven subtyping of Parkinson's disease using longitudinal clinical records: a cohort study". In: *Scientific reports* 9.1 (2019), p. 797.

[171] Yuan Zhang. "ATTAIN: Attention-based time-aware LSTM networks for disease progression modeling." In: *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-2019), pp. 4369-4375, Macao, China.* 2019.

[172] Yutao Zhang et al. "LEAP: learning to prescribe effective and safe treatment combinations for multimorbidity". In: *proceedings of the 23rd ACM SIGKDD international conference on knowledge Discovery and data Mining*. 2017, pp. 1315–1324.

[173] Jing Zhao et al. "Learning from heterogeneous temporal data in electronic health records". In: *Journal of biomedical informatics* 65 (2017), pp. 105–119.

[174] Rui Zhao et al. "Machine health monitoring using local feature-based gated recurrent unit networks". In: *IEEE Transactions on Industrial Electronics* 65.2 (2017), pp. 1539–1548.

[175] Xiaokang Zhou, Yue Li, and Wei Liang. "CNN-RNN based intelligent recommendation for online medical pre-diagnosis support". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18.3 (2020), pp. 912–921.

[176] Yu Zhu et al. "What to Do Next: Modeling User Behaviors by Time-LSTM." In: *IJCAI*. Vol. 17. 2017, pp. 3602–3608.

[177] Bo Zong et al. “Deep autoencoding gaussian mixture model for unsupervised anomaly detection”. In: *International conference on learning representations*. 2018.

[178] Quan Zou et al. “Predicting diabetes mellitus with machine learning techniques”. In: *Frontiers in genetics* 9 (2018), p. 515.

Part IV

Résumé de Thèse en Français

Chapter 9

Résumé de Thèse en Français

"Change your language and you change your thoughts."

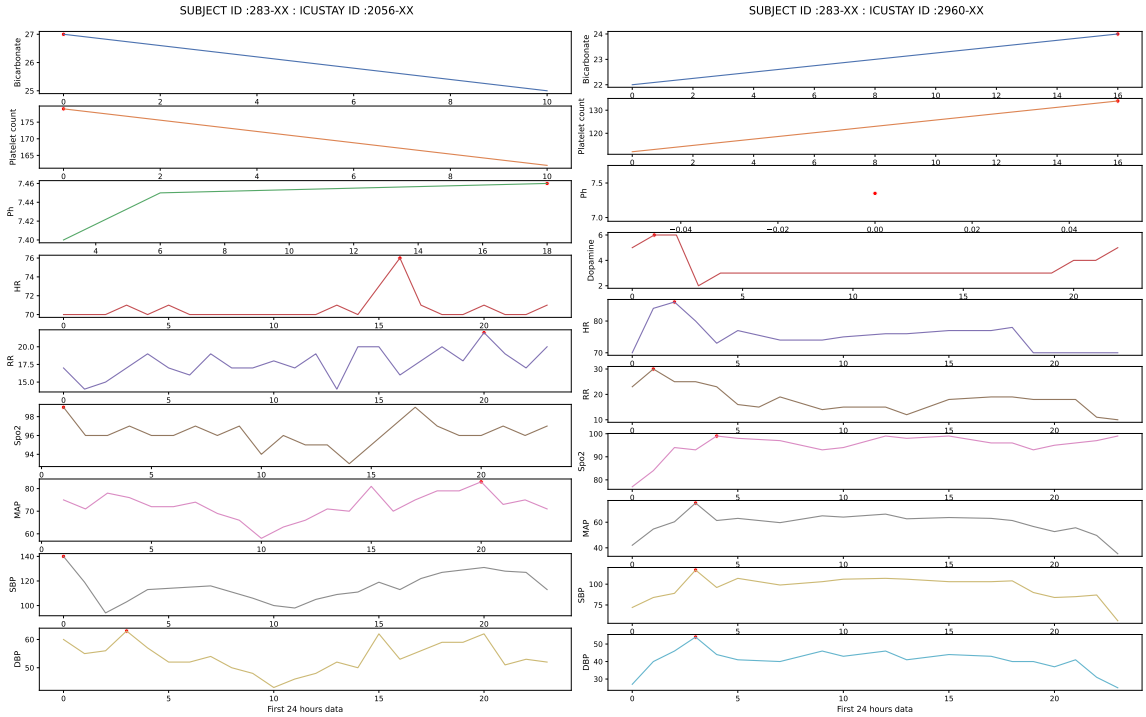
– Karl Albrecht

9.1 Introduction

Les systèmes de dossiers de santé électroniques (DSE) sont essentiels pour améliorer les soins aux patients en regroupant des informations médicales complètes telles que les diagnostics, les traitements et les résultats collectés à partir de divers capteurs hospitaliers. Ces données variées stockées sans structure spatiale ou séquentielle stricte incluent des symptômes, des enregistrements de médicaments, des tests de laboratoire, des procédures, des signaux physiologiques et des notes cliniques rédigées par les médecins. De plus, l'adoption généralisée des DSE dans le domaine de la santé a suscité un intérêt considérable de la part de la communauté de l'apprentissage automatique. Les chercheurs s'efforcent de développer des outils de prise de décision clinique basés sur des preuves en utilisant divers algorithmes d'apprentissage automatique [162]. De nombreuses études dans le domaine médical ont démontré l'efficacité de l'utilisation des réseaux neuronaux séquentiels pour représenter les trajectoires de soins des patients et réaliser des analyses cliniques prédictives. Des modèles basés sur les réseaux de neurones récurrents (RNN), l'attention et les transformers ont été appliqués avec succès à diverses tâches de prédiction d'événements cliniques.

Dans le contexte des séries temporelles d'événements au sein des DSE, chaque entrée représente un événement clinique significatif (par exemple, une ordonnance de médicament) avec des attributs détaillés tels que l'horodatage, le type, l'élément et la valeur. Ces événements forment collectivement une série temporelle complexe multivariée, cruciale pour comprendre la dynamique des soins aux patients, comme illustré par la Figure. 9.1, qui montre l'historique des soins cliniques du patient à partir des dossiers DSE.

Les modèles basés sur les RNN, en particulier les réseaux Long Short-Term Memory [63](LSTM) et Gated Recurrent Unit [27](GRU) [5, 18, 124, 129, 153, 171] sont largement utilisés pour modéliser les séries temporelles médicales et ont montré des résultats prometteurs par rapport aux algorithmes de machine learning traditionnels. Par exemple, Choi et al [25] ont développé un processus en deux étapes où ils ont d'abord défini des vecteurs d'embedding pour les codes d'événements médicaux en utilisant le modèle skip-gram [113]. Ces embeddings ont ensuite été intégrés dans une couche GRU pour générer des représentations cachées du patient afin de prédire le risque d'insuffisance cardiaque. Zhang et al [171] ont proposé ATTAIN, un modèle de progression des maladies basé sur l'attention et prenant en compte le temps pour la prédiction précoce du choc septique, qui intègre le mécanisme d'attention et modélise l'irrégularité temporelle entre les événements. Ils ont spécifiquement ajusté la cellule mémoire précédente de Long Short-Term Memory (LSTM)[63] $\boldsymbol{c}^{t-1}$ pour accumuler les informations antérieures. Au lieu d'ajouter des informations provenant d'un seul événement antérieur, ils ont rétrospectivement évalué les mémoires de tous ou de plusieurs événements antérieurs et les ont pondérées par des poids générés à partir du mécanisme d'attention et des intervalles de temps entre ces événements et l'événement actuel. Les poids globaux représentent l'importance de chaque événement précédent pour l'événement actuel afin d'identifier l'évolution de la condition. Trois mécanismes d'attention sont explorés : global ($\boldsymbol{g}$), local ($\boldsymbol{l}$) et flexible ($\boldsymbol{f}$), pour générer les poids d'attention. D'autre part, les intervalles de temps sont transformés en poids de décroissance par une fonction de décroissance, de sorte que les événements obsolètes jouent un rôle moins important que les événements récents dans la prédiction de l'issue de l'événement actuel comme suit



(a) Première admission en soins intensifs (b) Deuxième admission en soins intensifs

Figure 9.1: L'illustration montre les trajectoires de santé d'un patient au cours de son hospitalisation, où il a été admis deux fois en soins intensifs, avec à chaque fois un schéma distinct. L'axe X indique le moment de la mesure de chaque paramètre clinique (horaire), et l'axe Y représente la valeur mesurée.

: $\boldsymbol{C}^{t-1} = \sum_{i=t-m}^{\boldsymbol{t-1}} \alpha_{ti} \cdot \boldsymbol{c}^{i} \cdot \mathrm{g}(\Delta_{t_{ti}})$ où α_{ti} est le poids d'attention de l'événement i^{me} vers l'événement actuel t, $\Delta_{t_{ti}}$ est l'intervalle de temps entre l'événement i^{me} et l'événement actuel, $\mathrm{g}(\cdot)$ est une fonction de décroissance, et m représente le nombre d'événements à regarder en arrière.

Plus récemment, les chercheurs ont investigué l'efficacité des modèles basés sur les transformers pour la modélisation des séries temporelles médicales, démontrant ainsi des performances supérieures à celles des modèles GRU/LSTM [102, 104, 147]. Song et al [147] ont utilisé un vecteur d'embedding multidimensionnel pour représenter chaque épisode, encapsulant toutes les informations sur les événements médicaux. Ils ont traité ces embeddings à travers un encodeur basé sur le transformer avec attention causale (où une observation t ne peut tenir compte que des informations passées $j < t$) et une couche d'interpolation dense pour tenir compte des écarts temporels entre les observations. Tipirneni et al [155] ont proposé un modèle basé sur un transformateur avec une nouvelle composante d'embedding d'entrée en triplet, qui représente le temps de l'observation t, les caractéristiques $(f_j)_{1\leq j\leq K}$ observées à t et leurs valeurs $(v_j)_{1\leq j\leq K}$.

Cependant, la construction de modèles de machine learning à partir des séries temporelles des DSE présente plusieurs défis en raison de leurs caractéristiques distinctives. Modéliser efficacement les prédictions médicales nécessite de prendre en compte les défis liés au traitement des données d'observation hautement diversifiées présentes dans les bases de données cliniques du monde réel. Ces défis peuvent être classés en six (6) domaines clés :

(1) **Haute dimensionnalité:** Les données des DSE sont intrinsèquement multidimensionnelles en raison des nombreux types d'événements cliniques et de diagnostics qui peuvent survenir au cours de l'hospitalisation d'un patient. Le grand nombre de types d'événements et de concepts cliniques augmente la complexité de la modélisation, car cela oblige le modèle à apprendre et à maintenir des connaissances sur chaque type d'événement. Ce problème de haute dimensionnalité est souvent qualifié de "malédiction de la dimensionnalité". De plus, la prédiction des événements futurs implique d'apprendre des dépendances complexes entre un événement futur et une série d'événements antérieurs, ce qui peut être difficile à énumérer et à apprendre [173].

(2) **Observations manquantes et irrégulières:** Les événements cliniques dans les données des DSE ne sont souvent pas observés de manière universelle chez tous les patients en raison de leur association avec des maladies ou conditions spécifiques. Par exemple, des événements comme l'administration d'insuline sont généralement enregistrés uniquement pour les patients diabétiques, ce qui entraîne des données clairsemées. Cette rareté pose des défis pour la formation de modèles robustes capables de bien se généraliser à tous les types d'événements et à des patients non observés. De plus, les intervalles irréguliers entre les observations compliquent davantage l'analyse des données des DSE, qui servent principalement à stocker des informations sur les patients plutôt qu'à des fins de recherche clinique. Cette collecte de données de routine introduit une variabilité dans l'intégrité des données au fil du temps, ce qui entraîne des défis en matière d'informations manquantes [2, 40]. En

outre, les données des DSE contiennent souvent des valeurs aberrantes [13] qui compromettent la cohérence et l'exactitude des donnée entraînant souvent des résultats non fiables. Par conséquent, un prétraitement rigoureux des données est essentiel pour garantir la fiabilité des outils de machine learning.

(3) **Variabilité des patients:** Les données des DSE reflètent l'hétérogénéité des séquences des patients à travers différentes populations de patients. Chaque patient peut avoir une combinaison unique de complications cliniques, de régimes médicamenteux et de dynamiques de séquence observées. Bien que les comportements moyens puissent être capturés par un modèle unique, ce dernier peut avoir du mal à représenter les dynamiques détaillées des séquences de chaque patient.

(4) **Temporalité:** Les données des DSE sont intrinsèquement longitudinales, avec un ensemble de visites de patients créant plusieurs séries temporelles. L'ordre et le temps entre ces visites sont essentiels et contiennent des informations importantes pour comprendre l'évolution de l'état de santé et la trajectoire du patient au cours de la période de soins, et pour extraire des connaissances cliniques appropriées. Les événements cliniques sont échantillonnés de manière irrégulière. L'ordre des événements cliniques et la différence de temps entre les événements sont des informations précieuses pour apprendre des modèles prédictifs. Cependant, l'encodage des informations temporelles dans la modélisation prédictive est un défi relativement nouveau, qui manque de méthodologies établies, et freine l'application des méthodes classiques de l'apprentissage automatique (Machine Learning).

(5) **Tailles des groupes cibles (Biais de données):** Dans les applications de machine learning clinique, il est crucial de disposer de suffisamment de données patients pour former des modèles prédictifs précis et obtenir des informations significatives. Cependant, la rareté des dossiers, en particulier pour certains événements ou conditions cliniques spécifiques, conduit fréquemment à des problèmes de déséquilibre de classes CIP (Class Imbalance Problems) dans l'analyse des données de santé. Le CIP (Class Imbalance Problems) survient lorsque certaines classes, telles que les événements cliniques rares, sont significativement sous-représentées par rapport aux autres [55]. Les algorithmes de prédiction traditionnels supposent une distribution équilibrée des classes dans les ensembles d'entraînement. Dans les ensembles de données déséquilibrés, cette supposition entraîne un biais en faveur de la classe majoritaire, négligeant un apprentissage adéquat de la distribution de la classe minoritaire [17]. Ce problème est particulièrement critique en santé, où la précision de la prédiction des événements de classe minoritaire est souvent plus importante.

(6) **Sécurité et confidentialité:** Les données cliniques sont hautement sensibles et leur utilisation est limitée par des restrictions de confidentialité, des réglementations et des directives organisationnelles [6]. Les DSE sont régulés par des lois protégeant la confidentialité des patients, telles que la Health Insurance Portability and Accountability Act aux États-Unis et le Règlement général sur la protection des données dans l'Union européenne [165]. Le partage et l'utilisation des données cliniques nécessitent l'approbation d'un comité ou d'une organisation de santé, ce qui est restrictif et chronophage, limitant l'accès et freinant la recherche et la mise en place d'études innovantes [6].

En somme, les observations médicales sont enregistrées de manière irrégulière lors des visites hospitalières des patients, ce qui aboutit à des historiques de soins variés influencés par l'état de santé et les pratiques locales. Cela conduit à des trajectoires de santé diverses de longueurs et d'intervalles différents entre les observations. Chaque point de ces trajectoires reflète un épisode de soins, générant divers types de données (comme des rapports textuels, des prescriptions, des tests de laboratoire, des résultats, des paramètres médicaux, des diagnostics et des codes administratifs) provenant de multiples prestataires. L'intégration de ces divers types de données dans un modèle unifié pose un défi important, nécessitant un système multimodal pour extraire les informations essentielles de chaque entrée tout en évitant la redondance.

À titre d'exemple, après avoir effectué un échantillonnage horaire des données pour un ensemble de 103 variables utilisées dans notre étude, nous avons découvert que l'absence d'enregistrement à la fois des données de laboratoire et des signes vitaux dépassait 93 % pour tous les patients. Ce phénomène s'explique par le fait que chaque patient peut ne subir que quelques tests médicaux en fonction de ses besoins spécifiques, et que ces tests sont généralement planifiés de manière sporadique sur une longue durée, comme le montrent la Fig. 9.2 pour les variables biologiques, où nous constatons que certaines variables, comme la glycémie et la FiO2, sont enregistrées plus fréquemment que d'autres, telles que l'albumine and Fig. 9.3 pour les signes vitaux, médicaments et les variables neurologiques.

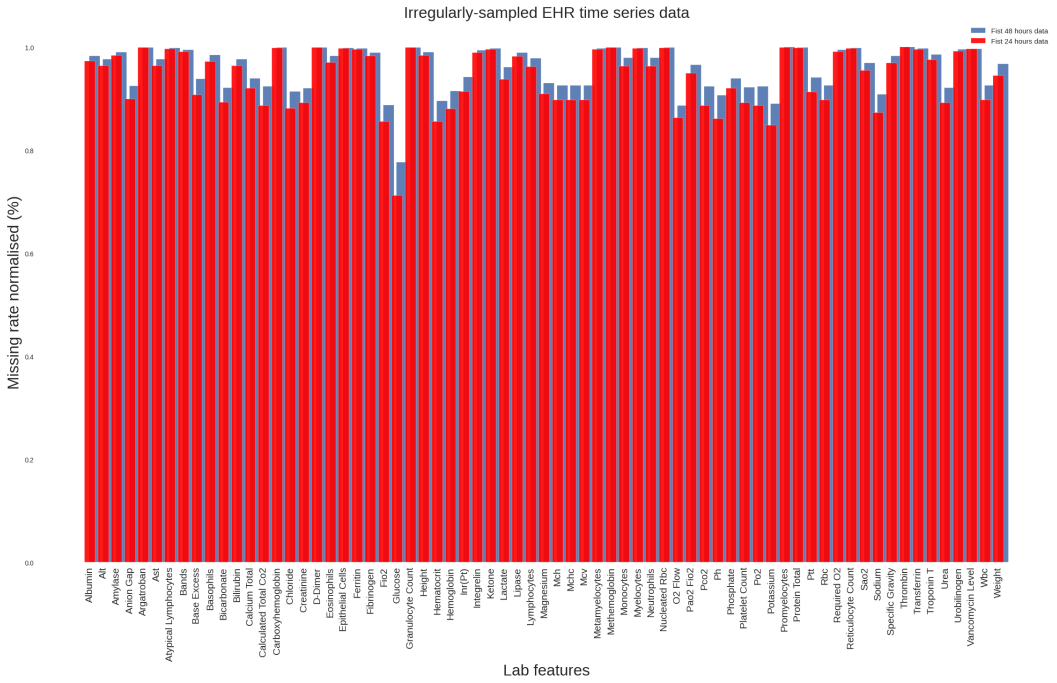


Figure 9.2: Ratio de valeurs manquantes normalisé (%) pour les caractéristiques de laboratoire au cours des premières 24 et 48 heures de collecte des données (MIMIC III).

Pour exploiter pleinement les données des DSE, il est crucial de relever ces défis. Cette thèse se concentre donc sur le développement de méthodologies efficaces et évolutives pour surmonter ces obstacles. Des discussions détaillées sur ces défis et nos méthodologies sont présentées dans les sections suivantes.

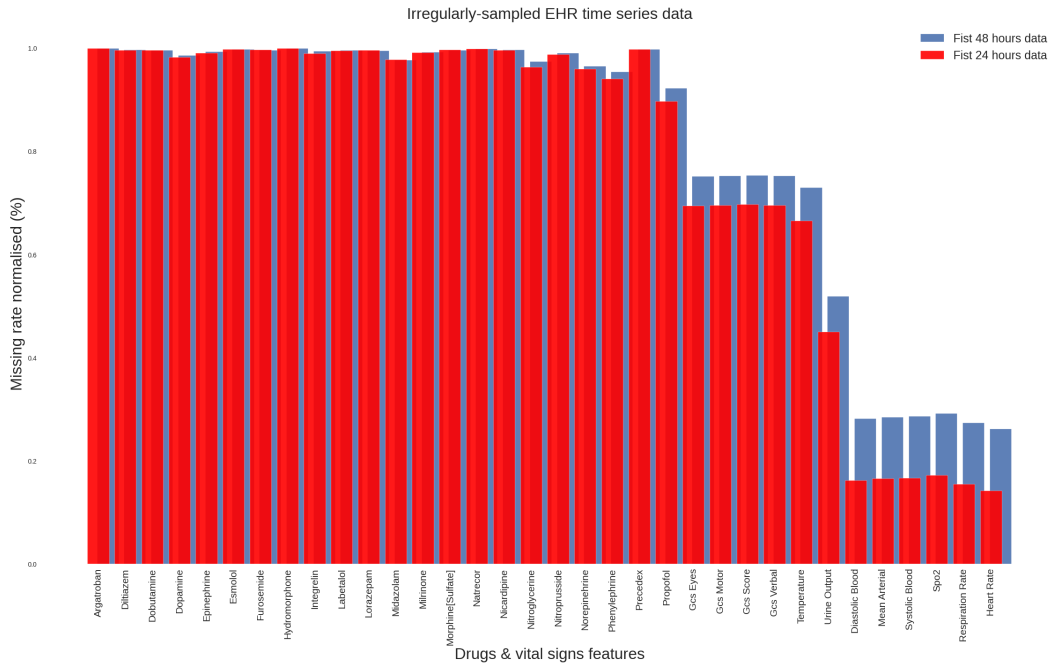


Figure 9.3: Ratio de données manquantes normalisé (%) pour les signes vitaux et les caractéristiques des médicaments au cours des premières 24 et 48 heures de collecte de données (MIMIC III).

9.2 Contexte général et objectifs scientifiques

Comme discuté dans les chapitres précédents, les dossiers de santé électroniques (DSE) sont des versions numériques des dossiers de santé des patients qui stockent un large éventail de données hétérogènes liées à leurs visites à l'hôpital. Ces données incluent les diagnostics, les résultats de laboratoire, les médicaments prescrits, les procédures, les images médicales, ainsi que les informations démographiques enregistrées par les professionnels de santé pour rendre compte de l'état du patient et des événements médicaux survenus au cours de son séjour [122]. Ces dossiers sont également caractérisés par une haute dimensionnalité et une grande hétérogénéité, allant de variables continues à des données d'événements tels que les interventions et l'administration de médicaments.

En raison de la complexité des données issues des DSE, les méthodes actuelles [5, 124, 171] et les architectures standard de LSTM [63], conçues pour des éléments de séquence uniformément espacés comme dans la modélisation linguistique, rencontrent des difficultés majeures à les gérer. Ces méthodes peinent à capturer les irrégularités temporelles et à conserver l'historique médical pertinent, risquant ainsi de négliger des épisodes de santé critiques parmi des données non pertinentes. Dans le domaine médical, le temps entre les points de données du patient varie considérablement, ce qui est essentiel pour comprendre l'état de santé et l'évolution de la maladie, comme le montre la Fig. 9.1, illustrant les trajectoires de santé uniques d'un patient lors de multiples admissions en soins intensifs. Ces schémas soulignent la nécessité d'interventions personnalisées basées sur l'état de santé

dynamique des patients.

Par exemple, des intervalles courts entre les mesures de lactate peuvent indiquer des changements rapides et une détérioration de l'état nécessitant une intervention immédiate. À l'inverse, des intervalles plus longs peuvent indiquer une stabilité. Une surveillance continue de variables telles que les niveaux de lactate, la posologie de norépinéphrine et la pression artérielle à des intervalles variables est essentielle pour évaluer la santé d'un patient et l'efficacité du traitement. Pour répondre aux complexités des données issues des DSE, il est nécessaire de développer de nouveaux modèles basés sur les LSTM, capables d'accommoder les irrégularités temporelles et les interactions entre les paramètres cliniques, offrant ainsi une représentation plus précise de la trajectoire de santé d'un patient.

Ainsi, lors du développement de modèles pour apprendre à partir de la chronologie temporelle d'un patient, il est crucial de prendre en compte l'hétérogénéité et l'irrégularité des données, ainsi que les relations complexes entre les événements cliniques. Cela permet de s'assurer que le modèle reflète avec précision l'état de santé du patient et peut fournir des informations précieuses pour orienter la prise de décision clinique.

Nous proposons que traiter les irrégularités temporelles et ajuster chaque caractéristique en fonction de son importance a le potentiel d'améliorer considérablement la précision des prédictions des résultats des patients. Cette affirmation repose sur deux facteurs clés qui soulignent les avantages de cette approche.

Premièrement, l'analyse des irrégularités temporelles pour chaque caractéristique de manière individuelle offre une compréhension plus nuancée et précise de l'information temporelle. Cela permet de mieux comprendre l'évolution de chaque caractéristique spécifique au fil du temps et son influence sur les résultats du patient.

Deuxièmement, il est crucial de reconnaître que différentes caractéristiques des données des DSE peuvent présenter des schémas de décroissance distincts, certaines se dégradant plus rapidement que d'autres dans le même intervalle de temps. Il devient donc essentiel de traiter ces caractéristiques de manière personnalisée et spécifique, en tenant compte de leurs caractéristiques temporelles uniques. Cette approche permet de capturer avec précision les dynamiques variées inhérentes aux différentes caractéristiques, facilitant une analyse complète de leur comportement temporel. Ainsi, mesurer la fréquence de chaque paramètre clinique offre une indication plus précise de l'état de gestion d'un patient pour plusieurs raisons importantes, notamment en considérant la manière dont chaque caractéristique se dégrade dans l'état de santé en cours du patient. Voici les trois principales raisons :

1. **Surveillance longitudinale:** En suivant la fréquence des paramètres cliniques au fil du temps, les professionnels de santé peuvent identifier les tendances et les changements dans l'état du patient. Cela est crucial pour détecter une détérioration ou une amélioration progressive qui peut ne pas être évidente à partir d'une seule mesure.

2. **Détection précoce des complications:** Une surveillance régulière peut détecter les premiers signes de complications. Par exemple, des contrôles fréquents du niveau de glucose chez les patients diabétiques peuvent permettre de détecter précocement une hyperglycémie ou une hypoglycémie, permettant ainsi des interventions rapi-

des [156]. De même, chez les patients atteints d'insuffisance cardiaque chronique, une surveillance fréquente de la pression artérielle et de la fréquence cardiaque peut aider à détecter l'aggravation de l'insuffisance cardiaque ou les effets des ajustements médicamenteux.

(3) **Médecine personnalisée:** La surveillance fréquente soutient un traitement personnalisé en fournissant des informations détaillées sur la manière dont un patient réagit à différentes thérapies au fil du temps. Cette approche est bénéfique pour la gestion des maladies chroniques comme l'insuffisance cardiaque, car elle permet aux cliniciens de prendre des décisions éclairées concernant l'ajustement des doses ou le changement de médicaments pour obtenir de meilleurs résultats.

En résumé, l'objectif principal de cette thèse est de construire une architecture à réseaux neuronaux à mémoire dynamique qui exploite les différents types d'informations contenues dans les données des DSE et apprend à représenter la chronologie du patient. Ainsi, il s'agit de concevoir un cadre permettant de représenter l'irrégularité temporelle présente dans le dossier medical lors de la modelisation de la trajectoire de santé du patient.

9.2.1 Contributions

L'importance de cette thèse s'articule autour de deux axes principaux : (i) l'introduction de nouvelles architectures RNN pour modéliser différents types de données structurées et (ii) l'application de ces modèles à un large éventail de problèmes pratiques dans le domaine de la santé, notamment la prévision de séries temporelles. Plus précisément, nos contributions sont les suivantes :

(1) Dans la phase initiale de cette thèse, nous menons une étude approfondie centrée sur des variables clés. Notre objectif est d'évaluer leur importance prognostique et d'identifier les facteurs prédictifs cruciaux pour un traitement efficace dans les scénarios de mortalité en soins intensifs. Cette investigation vise à valider les conclusions exposées dans les travaux de Hugerot et al. [127].

(2) La deuxième partie de ce travail de thèse aborde la modélisation des horodatages irréguliers dans les séries temporelles d'événements cliniques. La contribution principale réside dans la mise en œuvre d'un cadre générique pour les séries temporelles d'événements irréguliers et l'évaluation de l'effet de différentes techniques d'imputation sur les performances du modèle. À l'aide de ce cadre, nous avons réalisé une étude empirique de la mortalité et de la durée de séjour basée sur deux ensembles de données réels, en la comparant avec les approches neuronales temporelles actuelles.

(3) Troisièmement, nous proposons un nouveau modèle de deep learning appelé MWTA, opérant sur les dossiers de santé électroniques (DSE), dans lequel l'historique médical d'un patient est représenté comme une séquence d'événements pour chaque caractéristique clinique. Nous étendons l'unité de mémoire à long terme (LSTM), une variante des RNN, pour gérer les événements à chronologie irrégulière en modélisant le temps écoulé entre deux enregistrements consécutifs dans l'oubli de la mémoire.

MWTA modélise également explicitement l'interaction entre la progression de la maladie et les interventions (par exemple, les traitements médicaux), où les interventions modifient le cours de la maladie et influencent les risques médicaux futurs. Nous démontrons que MWTA est efficace pour différentes tâches d'apprentissage sur divers ensembles de données.

(4) Quatrièmement, nous proposons une nouvelle stratégie de pooling adaptatif appelée ASP (Adaptive Pooling Strategy) qui cible spécifiquement et résout les problèmes d'outliers pouvant potentiellement survenir lors de l'analyse des données DSE.

(5) Cinquièmement, nous introduisons un nouveau cadre appelé AMITA (Adaptive Multi-Way Interpretable Time-Aware LSTM), qui gère efficacement les irrégularités temporelles et est capable de capturer les interactions complexes entre les différentes caractéristiques cliniques à différents stades. AMITA étend les $portes$ standard de LSTM (oubli, entrée et sortie) en utilisant l'intervalle de temps entre les événements, la fréquence des mesures et les informations contextuelles issues de l'historique du patient. Cette extension permet d'ajuster la cellule de mémoire, en réduisant l'impact de la mémoire précédente à mesure que le temps écoulé entre les événements augmente, tout en conservant l'effet durable des événements antérieurs. De plus, nous incorporons deux mécanismes de porte pour refléter efficacement les effets des interventions sur l'état de la maladie actuelle.

- Dans le domaine de la santé, les données des patients arrivent souvent de manière irrégulière en raison de protocoles cliniques variés et de la surveillance des conditions. Nous avons amélioré le mécanisme de porte d'oubli d'AMITA pour tenir compte du timing et de la fréquence des événements ainsi que du contexte historique du patient. Cette adaptation permet au modèle de s'ajuster aux schémas temporels uniques de chaque patient, en mettant l'accent sur les données critiques tout en minimisant l'impact des mesures moins pertinentes ou sporadiques.

(6) Enfin, nous avons proposé une nouvelle méthodologie d'interprétabilité en expliquant comment chaque caractéristique clinique se comporte dans le parcours de santé du patient, en utilisant des fonctions agrégées sur les données historiques médicales du patient.

9.3 Cadre méthodologique

Cette section est consacrée à une brève présentation des bases des données et de l'ensemble des outils de méthodologie utilisés pour les analyses qui seront illustrées dans les sections suivantes.

9.3.1 Bases de données MIMIC & eICU

Cette étude repose sur l'analyse de bases de données publiques et anonymisées, déjà approuvées par des comités d'éthique (Institutional Review Board, IRB) [34]. Les comités

d'éthique du Beth Israel Deaconess Medical Center à Boston et du Massachusetts Institute of Technology à Cambridge ont approuvé la collecte, le traitement et la diffusion des données pour la base de données MIMIC-III. Le comité de recherche de l'eICU a également approuvé la collecte, le traitement et la diffusion des données pour la base de données eICU, qui ne nécessite pas d'approbation de l'IRB, conformément aux lois européennes sur la protection des données.

1. **MIMIC III [70]:** MIMIC-III est une base de données publique issue de données de dossiers médicaux électroniques réels et anonymisés, contenant les dossiers médicaux d'environ 46 000 patients admis en soins intensifs au Beth Israel Deaconess Medical Center entre 2001 et 2012. La base de données offre des informations détaillées sur divers événements médicaux survenus pendant le séjour des patients. Les caractéristiques incluent les signes vitaux, les médicaments, les mesures de laboratoire, les observations et notes des soignants, l'équilibre des fluides, les codes de procédures, les codes diagnostiques, les rapports d'imagerie, la durée du séjour hospitalier et les données de survie. Elle couvre les informations de santé de 38 597 patients adultes et de 49 785 admissions hospitalières. Ce jeu de données présente les défis typiques de toute grande base de données clinique, notamment des séquences de longueurs variables, des distributions biaisées, des valeurs manquantes, des visites épisodiques et des intervalles de temps très variables entre les visites successives.

2. **eICU [125]:** La base de données eICU Collaborative Research est une base de données multi-centres de soins intensifs, avec des données de haute granularité pour plus de 200 000 admissions dans des unités de soins intensifs suivies par les programmes eICU à travers les États-Unis. La base de données eICU comprend 200 859 rencontres unitaires de patients pour 139 367 patients uniques admis entre 2014 et 2015 dans 208 hôpitaux situés à travers les États-Unis.

   La base de données contient des dossiers complets, incluant les signes vitaux, les médicaments, les mesures de laboratoire, les observations et notes des soignants, des informations démographiques, des diagnostics (codes de la Classification internationale des maladies, neuvième révision) et d'autres données cliniques recueillies lors des soins médicaux de routine.

9.3.2 Notations

Soit $n \in N$ l'ensemble des séjours en réanimation considérés dans une étude médicale. Pour constituer notre cohorte de sujets, nous représentons chaque échantillon comme suit :

$$X = \left\{ \left(X^{(n)}, X_{last}^{(n)}, F_X^{(n)}, \Delta_t^{(n)} \right) \right\}_{n=1}^{N}, \tag{9.1}$$

où chaque échantillon est associé à un indice unique n.

$X^{(n)}$ représente les données temporelles du patient, $\Delta_t^{(n)}$ désigne les intervalles de temps entre les enregistrements consécutifs, $X_{last}^{(n)}$ fait référence à la dernière observation (la plus récente), et $F_X^{(n)}$ capture la fréquence de mesure de chaque variable physiologique.

Pour chaque séjour au sein de notre cohorte de N séjours, nous considérons une série temporelle multivariée englobant D variables physiologiques, telles que les examens de laboratoire et les signes vitaux, observées au fil du temps T (taille de la fenêtre) :

$$X_t^d = \{x_1, x_2, \ldots, x_T\} \in \mathbb{R}^{T \times D} \tag{9.2}$$

où chaque x_t^d représente l'observation de la d^{me} variable (par exemple, la pression artérielle diastolique ou la fréquence cardiaque, etc.) mesurée à l'instant $t \in \{1, 2, \ldots, T\}$.

Pour gérer les valeurs manquantes dans X_t^d, nous introduisons un vecteur de masquage $M_t^d = \{m_1, m_2, \ldots, m_T\} \in \mathbb{R}^{T \times D}$ de la même taille que X. Ce vecteur indique quelles variables sont observées ou manquantes à chaque instant. Nous initialisons M comme suit :

$$m_t^d = \begin{cases} 1, & \text{if } x_t^d \text{ is observed in the input data} \\ 0, & \text{otherwise} \end{cases} \tag{9.3}$$

Cette initialisation garantit que m_t^d vaut 1 si la variable x_t^d est observée dans les données d'entrée, et 0 sinon, fournissant une indication claire de la disponibilité ou de l'absence de données à chaque instant.

Nous calculons la fréquence des mesures de f_x en fonction du nombre de fois où n_x, x a été mesuré sur une fenêtre temporelle donnée T.

$$f_x = n_{x^d}/T \tag{9.4}$$

Cette formule exprime le ratio du nombre de mesures de la variable x^d par rapport à la taille de la fenêtre T, fournissant une mesure de la fréquence des observations dans le cadre temporel spécifié.

Nous extrayons également la dernière observation x_{last}^d de chaque variable physiologique x^d dans une fenêtre de taille T à l'aide de la représentation suivante :

$$x_{\text{last}}^d = \{x_l^d : i \in \{t - T + 1, \ldots, t\}\} \tag{9.5}$$

Ici, x^d désigne une caractéristique clinique dans le jeu de données D, x_l^d représente la dernière observation de la caractéristique x^d au point temporel t, T représente la taille de la fenêtre ou la longueur de la fenêtre historique des observations. *t* indique l'instant actuel ou le plus récent à partir duquel la dernière observation est effectuée. *t-T+1* à t définit la plage des instants dans la fenêtre de taille T.

Nous introduisons un intervalle de temps, $\Delta_t^d \in \mathbb{R}^D$, pour chaque variable x_t^d, en tenant compte des informations cruciales que portent les écarts de temps entre les collectes de données observées. Nous supposons que l'observation initiale est effectuée à l'instant $t = 1$, et $\Delta_1 = 0$. L'intervalle de temps est défini comme suit :

$$\Delta_t^d = \begin{cases} 0, & t = 1 \\ \Delta_{t-1}^d + 1, & t > 1, m_{t-1}^d = 0 \\ 0, & t > 1, m_{t-1}^d = 1 \end{cases} \tag{9.6}$$

Si $t = 1$, alors Δ_t^d est zéro, représentant la valeur de la variable au premier point temporel pris comme référence ou point de départ.

Si $t > 1$ et $m_{t-1}^d = 0$, alors $\Delta_t^d = \Delta_{t-1}^d + 1$. Cela signifie une tendance linéaire de la variable au fil du temps, où chaque point temporel est d'une unité plus grand que le point temporel précédent, car le sujet n'a pas subi d'événement (représenté par $m_{t-1}^d = 0$) à l'instant $t - 1$.

Si $t > 1$ et $m_{t-1}^d = 1$, alors Δ_t^d est zéro. Cela indique un "redémarrage" de la variable après qu'un événement soit survenu (représenté par $m_{t-1}^d = 1$) à l'instant $t - 1$.

- **Interventions :**

Dans ce contexte, nous considérons également que le vecteur de temps écoulé $\boldsymbol{\Delta}_t^{(d)}$ peut être considéré comme une intervention, car l'enregistrement des mesures des caractéristiques cliniques d'un patient au fil du temps est crucial pour évaluer son état de santé. La fréquence de ces mesures est directement corrélée à l'état de santé du patient, avec des intervalles plus courts entre les mesures pouvant indiquer une détérioration rapide de l'état, tandis que des intervalles plus longs peuvent suggérer un état stable. Globalement, le suivi des changements de ces mesures au fil du temps peut permettre d'ajuster le plan de traitement de manière appropriée, assurant ainsi une gestion efficace de l'état du patient.

9.4 Contribution

9.4.1 Corrélation entre la Norépinéphrine et le Lactate

L'insuffisance circulatoire aiguë, communément appelée choc, est l'une des principales raisons d'admission en unité de soins intensifs (réanimation). Le choc se caractérise par un déséquilibre entre le transport de l'oxygène dans les artères et les besoins en oxygène des tissus, ce qui entraîne l'activation du métabolisme anaérobie au niveau cellulaire et tissulaire. Ce changement métabolique conduit à une dysfonction organique, qui peut aggraver l'état de choc [159] et contribuer à l'apparition de défaillances d'organes supplémentaires. La reconnaissance et le traitement précoces du choc sont essentiels pour prévenir la défaillance multiviscérale et, finalement, le décès.

L'objectif de cette étude est d'évaluer l'impact pronostique sur la mortalité en réanimation de la dose maximale de norépinéphrine administrée et des niveaux de lactate au cours des premières 24 heures après l'admission en réanimation, indépendamment de l'indication chez les patients du $4^{\text{ème}}$ quartile. Pour ce faire, nous avons construit des courbes de survie pour ces patients, classés en quatre groupes : le lactate et la norépinéphrine diminuent au cours des 24 heures, le lactate diminue et la norépinéphrine augmente, le lactate augmente et la norépinéphrine diminue, et le lactate et la norépinéphrine augmentent.

Afin d'évaluer l'impact pronostique de la dose maximale de norépinéphrine reçue au cours des premières 24 heures suivant l'admission, nous avons mené une étude rétrospective incluant plus de 40 000 patients ayant séjourné dans les unités de soins intensifs du Beth Israel Deaconess Medical Center entre 2001 et 2012 [70]. Deux critères d'inclusion ont

été utilisés pour sélectionner les patients formant la cohorte de cette étude. Tout d'abord, nous avons identifié tous les patients adultes en utilisant l'âge enregistré au moment de l'admission en réanimation. Conformément aux études précédentes [70], tous les patients âgés de plus de 15 ans au moment de l'admission en réanimation sont considérés comme adultes dans notre étude. Le deuxième critère était d'inclure tous les patients pour lesquels un soutien vasopresseur par norépinéphrine avait été prescrit et validé par une infirmière au cours des premières 24 heures suivant l'admission. Les patients de moins de 15 ans, même s'ils avaient reçu de la norépinéphrine dans les 24 heures suivant leur admission, ont été exclus de l'étude.

Après application de ces critères, un total de 5230 patients, correspondant à 5735 séjours distincts dans les differentes unités de soins intensifs ont satisfait aux critères d'inclusion pour la prescription de norépinéphrine dans les 24 heures suivant l'admission en réanimation et ont été répartis en quatre quartiles. Les différents quartiles sont détaillés dans Table 9.1.

Table 9.1: Description of the different quartiles

		Missing	Overall	QUARTILE 1	QUARTILE 2	QUARTILE 3	QUARTILE 4	P-Value
Number of Patients			5735	1429	1430	1442	1434	
Survival ICU, mean (SD)		2010	248.1 (537.4)	370.5 (610.2)	300.6 (568.4)	229.3 (531.7)	133.3 (424.1)	<0.001
LOS, mean (SD)		0	7.6 (9.2)	6.6 (7.6)	7.6 (8.8)	8.8 (10.5)	7.5 (9.5)	<0.001
AGE, mean (SD)		0	67.3 (15.6)	67.8 (15.3)	68.1 (15.2)	67.2 (15.5)	66.2 (16.1)	0.006
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.2 [0.1,0.3]	0.1 [0.0,0.1]	0.1 [0.1,0.1]	0.2 [0.2,0.3]	0.5 [0.4,0.6]	<0.001
Epinephrine, mean (SD)		5156	0.2 (0.9)	0.0 (0.0)	0.1 (0.3)	0.1 (0.2)	0.3 (1.4)	0.012
Dopamine, mean (SD)		4700	16.0 (87.4)	9.6 (6.0)	11.4 (9.4)	13.1 (7.1)	22.5 (137.7)	0.268
Dobutamine, mean (SD)		5409	6.7 (5.0)	4.2 (2.2)	6.2 (4.9)	6.9 (4.3)	7.5 (5.7)	0.001
Lactate(mmol/L), mean (SD)		0	4.8 (4.0)	3.2 (2.4)	3.8 (3.0)	4.9 (3.9)	7.0 (5.1)	<0.001
Heart Rate, mean (SD)		1	118.4 (24.1)	111.4 (21.8)	115.7 (23.2)	121.0 (24.6)	125.7 (24.4)	<0.001
Systemic Blood Press(mmHg), mean (SD)		6	150.4 (25.9)	151.6 (24.3)	151.2 (25.4)	149.6 (25.4)	149.0 (28.1)	0.019
Mean Blood Press, mean (SD)		0	44.6 (14.2)	48.6 (12.7)	46.9 (12.8)	44.2 (13.4)	38.8 (15.8)	<0.001
Temperature(°C), mean (SD)		151	37.8 (1.1)	37.8 (0.9)	37.8 (0.9)	37.9 (1.1)	37.8 (1.3)	<0.001
Urine Output, mean (SD)		404	2981.7 (2979.7)	3324.1 (2144.1)	3165.8 (2216.2)	3010.5 (4167.5)	2410.4 (2829.8)	<0.001
Urea(mmol/l), mean (SD)		2	15.3 (10.0)	14.2 (9.6)	14.6 (9.4)	16.3 (10.6)	16.3 (10.2)	<0.001
Bilirubin(mg/dL), mean (SD)		857	2.5 (5.2)	1.9 (4.7)	1.9 (3.8)	2.7 (5.9)	3.3 (5.7)	<0.001
Bicarbonate(mEq/l), mean (SD)		16	18.3 (5.2)	20.3 (4.8)	19.3 (4.7)	17.8 (5.1)	15.7 (5.0)	<0.001
PaO2/FiO2, mean (SD)		990	173.0 (187.1)	192.2 (164.2)	173.8 (157.4)	175.1 (226.9)	155.1 (184.6)	<0.001
GCS Score, mean (SD)		243	7.3 (4.5)	8.4 (4.7)	7.9 (4.7)	7.0 (4.3)	6.0 (4.0)	<0.001
ICU Death, n (%)	0	0	3901 (68.0)	1210 (84.7)	1118 (78.2)	947 (65.7)	626 (43.7)	<0.001
	1		1834 (32.0)	219 (15.3)	312 (21.8)	495 (34.3)	808 (56.3)	
GENDER, n (%)	F	0	2466 (43.0)	613 (42.9)	590 (41.3)	656 (45.5)	607 (42.3)	0.128
	M		3269 (57.0)	816 (57.1)	840 (58.7)	786 (54.5)	827 (57.7)	
IGS II, mean (SD)		0	54.3 (16.7)	47.3 (15.7)	51.0 (16.2)	56.5 (15.7)	62.2 (15.3)	<0.001
SOFA, mean (SD)		0	10.7 (3.7)	8.7 (3.3)	10.2 (3.4)	11.3 (3.5)	12.5 (3.6)	<0.001

- **Analyse cinétique entre la noradrénaline et le lactate chez les patients du 4^{e} quartile**

Le 4^{e} quartile a été très significativement associé à un excès de mortalité à J1 (16%), J7 (45%), en réanimation (62%) et en milieu hospitalier (69%), avec un risque de décès en réanimation multiplié par 6 pour les patients de ce quartile (p<0,001). L'association entre la dose de noradrénaline supérieure à 0,31 µg/kg/min et la mortalité a été confirmée par une analyse multivariée utilisant un modèle de régression logistique multiple. Ce modèle prenait en compte les principaux facteurs de gravité associés à la mortalité en réanimation et/ou identifiés dans notre étude, à savoir le ratio PaO2/FiO2, l'utilisation du lactate, le score SOFA,

la survenue d'un arrêt cardiopulmonaire et l'utilisation d'adrénaline, avec une p-value < 0,001. Nous avons également constaté que, hormis le ratio PaO2/FiO2, tous les autres éléments étaient indépendamment associés à un excès de mortalité des patients, comme détaillé dans la Table 9.1.

Table 9.2: Description of the different kinetics

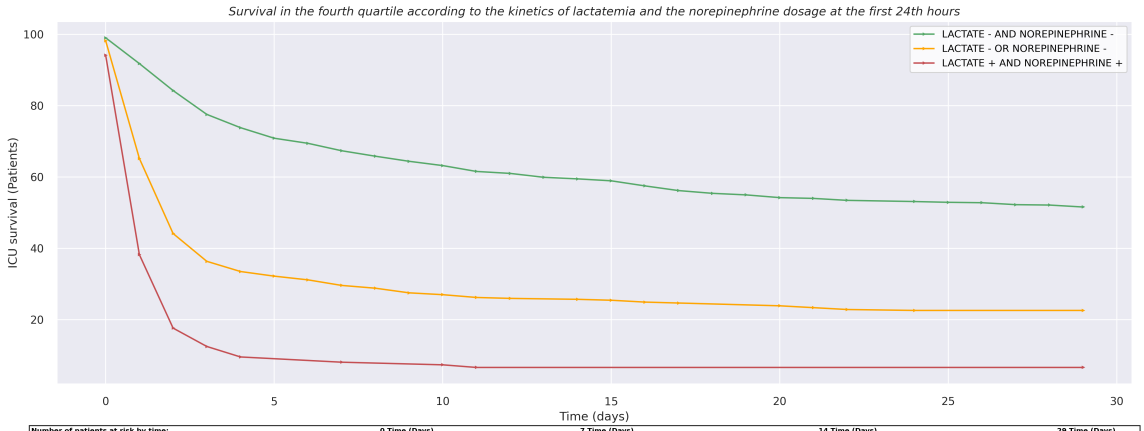
		Missing	Overall	NOP+ AND LACTATE+	NOP- AND LACTATE-	NOP- OR LACTATE-	P-Value
Number of Patients			1434	136	913	385	
Survival ICU, mean (SD)		331	133.3 (424.1)	25.7 (199.8)	181.2 (480.9)	83.9 (354.4)	<0.001
AGE, mean (SD)		0	66.2 (16.1)	68.4 (16.1)	65.9 (16.3)	65.9 (15.6)	0.251
Norepinephrine max (mcg/kg/min), median [Q1,Q3]		0	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.001
Epinephrine, mean (SD)		1242	0.3 (1.4)	0.2 (0.5)	0.3 (1.8)	0.4 (0.7)	0.940
Dopamine, mean (SD)		1020	22.5 (137.7)	18.6 (7.7)	13.6 (8.0)	38.0 (238.9)	0.264
Dobutamine, mean (SD)		1291	7.5 (5.7)	10.5 (7.1)	7.0 (5.4)	7.1 (5.5)	0.048
Lactate(mmol/L), mean (SD)		0	7.0 (5.1)	9.7 (6.0)	6.2 (4.4)	7.9 (6.0)	<0.001
Systemic Blood Press(mmHg), mean (SD)		4	149.0 (28.1)	141.9 (29.8)	151.9 (26.8)	144.8 (29.6)	<0.001
Mean Blood Press(mmHg), mean (SD)		0	38.8 (15.8)	32.9 (14.5)	41.2 (15.2)	35.1 (16.5)	<0.001
Temperature(°C), mean (SD)		39	37.8 (1.3)	37.3 (1.8)	38.0 (1.1)	37.5 (1.4)	<0.001
Urea(mmol/l), mean (SD)		2	16.3 (10.2)	16.4 (10.5)	16.3 (10.4)	16.1 (9.5)	0.942
Bilirubin(mg/dL)„ mean (SD)		168	3.3 (5.7)	2.9 (5.1)	3.1 (5.3)	3.8 (6.9)	0.118
PaO2/FiO2 ratio (mmHg), mean (SD)		109	155.1 (184.6)	154.6 (261.8)	153.2 (145.6)	159.7 (231.8)	0.857
GCS Score, mean (SD)		94	6.0 (4.0)	6.4 (4.5)	5.8 (3.8)	6.4 (4.3)	0.050
ICU Death, n (%)	0	0	626 (43.7)	9 (6.6)	519 (56.8)	98 (25.5)	<0.001
	1		808 (56.3)	127 (93.4)	394 (43.2)	287 (74.5)	
GENDER, n (%)	F	0	607 (42.3)	60 (44.1)	390 (42.7)	157 (40.8)	0.736
	M		827 (57.7)	76 (55.9)	523 (57.3)	228 (59.2)	
IGS II, mean (SD)		0	62.2 (15.3)	63.2 (16.6)	61.9 (15.1)	62.6 (15.4)	0.564
SOFA, mean (SD)		0	12.5 (3.6)	12.4 (3.4)	12.5 (3.5)	12.5 (3.8)	0.956

En plus de la dose maximale de noradrénaline et de sa durée d'administration, nous avons évalué l'impact pronostique de la cinétique des doses de noradrénaline à 24 heures. Parmi les 1434 patients recevant plus de 0,5 µg/kg/min, 896 (62%) avaient une dose réduite à 24 heures, tandis que 108 (7%) avaient une dose augmentée à 24 heures ou au moment du décès. Une diminution de la dose de noradrénaline à 24 heures était fortement associée à la survie des patients, les patients survivants ayant montré une diminution de 46,0% contre 85,2% chez les non-survivants (p<0,001).

Nous avons également identifié une association significative (p<0,001) entre la diminution des niveaux de noradrénaline et de lactate et la survie. Sur la base de ces résultats, nous avons établi des courbes de survie pour les patients du quatrième ($4^{ème}$) quartile, répartis en trois(3) groupes distincts, comme détaillé dans la Table 9.2 : ceux présentant une diminution du lactate et de la noradrénaline à 24 heures, ceux ayant une diminution du lactate ou de la noradrénaline, et enfin, ceux avec une augmentation du lactate et de la noradrénaline. L'analyse de survie a révélé que les patients présentant une augmentation des niveaux de lactate et de noradrénaline étaient associés à un excès de mortalité, tandis que ceux ayant une diminution simultanée de ces deux marqueurs avaient un pronostic moins critique que les patients présentant une diminution de l'un ou l'autre.

En résumé, notre étude dérivée d'une analyse rétrospective de patients ayant reçu de la noradrénaline dans les 24 premières heures suivant l'admission en réanimation, visait à évaluer les implications pronostiques de la dose maximale administrée au cours de cette période et à identifier les facteurs prédictifs influençant les résultats du traitement, corroborant ainsi les conclusions de Hugerot et al [127].

En conclusion, nos résultats suggèrent un lien entre la dose maximale de noradrénaline administrée dans les 24 premières heures et une augmentation du taux de mortalité au-delà



Number of patients at risk by time:		0 Time (Days)	7 Time (Days)	14 Time (Days)	29 Time (Days)
LACTATE - AND NOREPINEPHRINE -	(913 PATIENTS) :	904	615	543	471
LACTATE - OR NOREPINEPHRINE -	(385 PATIENTS) :	378	114	99	87
LACTATE + AND NOREPINEPHRINE +	(136 PATIENTS) :	128	11	9	9

Fourth Quartile		Overall	NOREPINEPHRINE + AND LACTATE +	NOREPINEPHRINE - AND LACTATE -	NOREPINEPHRINE - OR LACTATE -	P-Value
(n/N)		1434	136	913	385	
NAD max (mcg/kg/min), median [Q1,Q3]		0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	0.5 [0.4,0.6]	<0.001
AGE, mean (SD)		66.2 (16.1)	68.4 (16.1)	65.9 (16.3)	65.9 (15.6)	0.251
LACTATE(mmol/L), mean (SD)		7.0 (5.1)	9.7 (6.0)	6.2 (4.4)	7.9 (6.0)	<0.001
GENDER, n (%)	F M	607 (42.3) 827 (57.7)	60 (44.1) 76 (55.9)	390 (42.7) 523 (57.3)	157 (40.8) 228 (59.2)	0.736
ICU_DEATH, n (%)	0 1	626 (43.7) 808 (56.3)	9 (6.6) 127 (93.4)	519 (56.8) 394 (43.2)	98 (25.5) 287 (74.5)	<0.001

Figure 9.4: Mortalité en réanimation pour différentes cinétiques de noradrénaline et de lactate.

d'un seuil de 0,4 µg/kg/min. En particulier, les patients recevant des doses supérieures à 0,85 µg/kg/min ont montré une augmentation notable de la mortalité, ce qui incite à réévaluer l'efficacité du traitement dans de telles conditions. Cette observation est particulièrement cruciale lorsque la cinétique des doses de noradrénaline et les niveaux de lactate à 24 heures sont défavorables, ou lorsque des facteurs associés à un risque de mortalité plus élevé, tels que l'utilisation de l'adrénaline, la nécessité de la purification extra-rénale, ou la survenue d'arrêts cardiaques sont observés.

9.4.2 Multi-Way adaptive Time Aware LSTM

Dans le domaine de la médecine prédictive personnalisée, il est essentiel de modéliser avec précision l'évolution de la maladie d'un patient et les processus de soins, compte tenu des dépendances temporelles à long terme inhérentes à ces derniers. Cependant, les dossiers de santé électroniques (DSE) contiennent des données épisodiques et irrégulièrement horodatées, en raison des visites des patients à l'hôpital basées sur leurs besoins en soins, ce qui entraîne des schémas uniques pour chaque séjour hospitalier. Par conséquent, lors de la construction d'un modèle prédictif personnalisé, il est crucial de prendre en compte ces facteurs afin de capturer avec précision le parcours de santé du patient.

Pour relever ce défi, nous présentons un nouveau réseau de neurones à mémoire dynamique appelé MWTA-LSTM (Multi-Way adaptive Time Aware LSTM). L'objectif principal de MWTA-LSTM est de tirer parti des dossiers médicaux pour mémoriser les trajectoires de maladie et les processus de soins, estimer les états actuels de la maladie et prédire

les risques futurs, offrant ainsi un haut niveau de précision et de pouvoir prédictif. Pour améliorer ses capacités, MWTA-LSTM étend le modèle Long Short-Term Memory (LSTM) conventionnel de trois manières clés. Premièrement, il intègre la mesure de la fréquence et la plus récente observation pour améliorer la modélisation prédictive personnalisée des maladies des patients. Cette inclusion permet de mieux comprendre l'état du patient. Deuxièmement, MWTA-LSTM paramètre le temps pour gérer efficacement les irrégularités temporelles, permettant ainsi de modéliser les interventions et leur impact sur le cours de la maladie grâce à deux mécanismes de décroissance. Enfin, nous introduisons une nouvelle stratégie de regroupement adaptatif qui cible spécifiquement et résout les problèmes de valeurs aberrantes susceptibles de se produire lors de l'analyse des données des DSE. En intégrant ces fonctionnalités, MWTA-LSTM améliore considérablement sa capacité à capturer les dynamiques temporelles des données de santé, en s'adaptant aux variations et irrégularités dans le timing des événements et des observations.

- MWTA-LSTM gère efficacement les irrégularités temporelles et est capable de capturer les interactions complexes entre différentes caractéristiques cliniques à différents stades. MWTA-LSTM étend les $portes$ standards de LSTM [63] (*oublie*, *entrée*, et *sortie*) en ajoutant une $porte\ temporelle$ supplémentaire pour ajuster la cellule de mémoire de manière à diminuer l'influence de la mémoire précédente à mesure que le temps écoulé augmente. Cet ajustement est réalisé grâce à deux mécanismes de porte temporelle pour refléter les effets des interventions sur la sortie actuelle, dont l'un a été proposé par Baytas et al [5], tandis que l'autre est notre approche.
- MWTA-LSTM adopte une approche en traitant chaque caractéristique de manière différenciée, en apprenant leurs paramètres de décroissance en fonction des mesures de fréquence et des intervalles de temps entre les événements. Cette méthodologie permet au modèle de répondre efficacement aux irrégularités temporelles et de surmonter le défi des enregistrements clairsemés couramment rencontrés dans les données des DSE.
- Un mécanisme appelé Adaptive Pooling Strategy (ASP) est proposé pour aborder les problèmes de valeurs aberrantes qui peuvent surgir lors de l'analyse des données des DSE, renforçant ainsi la robustesse et la fiabilité de MWTA-LSTM.

Malgré leurs succès considérables, les modèles existants [5, 124, 171] ainsi que MWTA-LSTM rencontrent des difficultés substantielles à modéliser précisément les trajectoires de santé des patients, en particulier lorsqu'ils sont confrontés à des fréquences de mesure variables. En effet, ils négligent souvent un facteur crucial : la fréquence des mesures des caractéristiques cliniques, lors de l'ajustement de l'état de la cellule précédente. Ils se basent principalement sur le temps écoulé pour prendre en compte les irrégularités temporelles dans les dossiers médicaux électroniques des patients. Bien que cette approche soit valide pour certaines variables dont l'impact sur l'évolution des patients est retardé, elle n'est pas universelle. Par exemple, un niveau élevé de variables telles que le lactate ou la troponine T, bien que rarement mesuré, peut exercer une influence prolongée et significative sur l'évolution de la maladie du patient. Leur mesure peu fréquente n'en réduit pas l'importance clinique, surtout dans des situations critiques comme l'hypoxie tissulaire ou d'autres conditions médicales graves.

De plus, certaines interventions cliniques, telles que les stratégies de protection rénale [72], peuvent ne pas produire d'effets immédiatement observables sur des biomarqueurs comme la fonction rénale. Cependant, elles peuvent avoir un impact significatif sur les résultats des patients à long terme. Reconnaître et suivre l'effet à long terme de ces variables est essentiel pour une gestion précise des patients. De nombreuses études [72, 146, 154] ont souligné l'importance cruciale de cette considération. Singer et al. [146] ont notamment mis en évidence l'importance des niveaux élevés de lactate chez les patients gravement malades, les liant à des taux de mortalité accrus et à des séjours hospitaliers prolongés. Leur étude souligne l'importance de prendre en compte les mesures antérieures de lactate sur une longue période. Par conséquent, se fier uniquement au temps écoulé pour ajuster l'état de la cellule pourrait conduire à une sous-estimation de l'impact des variables qui ne sont pas mesurées fréquemment, risquant ainsi de passer à côté de la complexité complète du contexte clinique.

Il est donc impératif, lors de l'adaptation de l'état de la cellule pour tenir compte des irrégularités temporelles, d'intégrer les informations contextuelles issues de l'historique du patient, englobant à la fois la fréquence des mesures et les temps écoulés. Cette approche aide à distinguer les contributions de la mémoire à court et à long terme au sein de l'état de la cellule. En dépit de leurs réussites, ces modèles rencontrent également des obstacles significatifs lorsqu'ils sont appliqués aux données des dossiers médicaux électroniques. En premier lieu, beaucoup d'entre eux manquent de la capacité à fournir des résultats interprétables, un aspect crucial dans le domaine de la santé.

9.4.3 Adaptive Multi-Way Interpretable Time-Aware LSTM

Comme discuté précédemment, les modèles existants rencontrent d'importantes difficultés pour modéliser efficacement les trajectoires de santé des patients. Pour surmonter ces limitations, nous introduisons un nouveau réseau de neurones à mémoire dynamique profonde, nommé AMITA (Adaptive Multi-Way Interpretable Time-Aware LSTM) pour les données séquentielles collectées de manière irrégulière. L'objectif principal d'AMITA est d'exploiter les dossiers médicaux pour mémoriser les trajectoires de maladies et les processus de soins, estimer les états de santé actuels et prédire les risques futurs avec un haut degré de précision et de pouvoir prédictif.

Pour améliorer ses capacités, AMITA étend le modèle LSTM standard de deux manières clés. Premièrement, il intègre la mesure de fréquence et l'observation la plus récente pour affiner la modélisation prédictive personnalisée des maladies des patients, permettant ainsi une compréhension plus précise de l'état du patient. Deuxièmement, il paramètre l'état de la cellule pour gérer efficacement les temps irréguliers, en utilisant à la fois les temps écoulés et un facteur de décroissance basé sur la fréquence, qui prend en compte à la fois la fréquence des mesures et les informations contextuelles. De plus, le modèle utilise ces éléments pour comprendre l'impact des interventions sur l'évolution de la maladie dans l'état de la cellule, facilitant ainsi la mémorisation des trajectoires de maladie et améliorant sa capacité à capturer la dynamique temporelle des données de soins, en s'adaptant aux variations et irrégularités des moments et observations.

Conformément au concept fondamental introduit dans [5], nous avons systématiquement décomposé la mémoire antérieure $\boldsymbol{c}_{t-1}$ en deux composantes distinctes : l'une est le composant à long terme, conçu pour encapsuler l'impact durable des événements antérieurs, et l'autre concerne les mémoires à court terme ajustées, adaptées pour saisir les développements récents et les influences immédiates.

Après cette décomposition, la mémoire à court terme ajustée subit un mécanisme de réduction, intégrant un facteur de décroissance temporelle. Ce facteur est soigneusement dérivé d'une fusion du temps écoulé ($\boldsymbol{\Delta}_t$) et d'un facteur de décroissance basé sur la fréquence, qui prend en compte à la fois la fréquence des mesures et les informations contextuelles. Ainsi, la signification des données antérieures est pondérée de manière réfléchie, garantissant que les variables avec des mesures peu fréquentes, potentiellement cruciales pour la santé globale du patient, conservent leur pertinence même avec des intervalles prolongés entre les mesures.

Finalement, les mémoires à long terme et les mémoires à court terme ajustées se rejoignent, aboutissant à une mémoire mise à jour qui intègre de manière fluide l'impact prolongé des événements historiques avec les influences immédiates et les développements récents pour une compréhension complète de l'évolution de la santé du patient.

Cet ajustement permet de personnaliser le contenu de la mémoire dans la nouvelle unité d'état de la cellule, favorisant une compréhension plus complète de la trajectoire de santé du patient à partir de ses données. La fusion de ces composants de mémoire permet au modèle de tirer parti des forces de l'information à court terme et à long terme, offrant une réflexion équilibrée et précise de l'état évolutif du patient au fil du temps. Grâce à cette combinaison, le modèle acquiert la capacité de faire des prédictions plus fiables sur les événements futurs basées sur les données historiques du patient, offrant des prédictions à la fois précises et interprétables.

Dans le domaine de la santé, les données des patients arrivent souvent à des intervalles irréguliers en raison de divers facteurs, tels que les différences dans les protocoles cliniques ou la nature de la condition surveillée. Pour y faire face, nous avons étendu le mécanisme de porte d'oubli d'AMITA. Cette extension prend en compte à la fois l'écart de temps entre les événements successifs, la fréquence des mesures et les informations contextuelles provenant de l'historique du patient. Cela permet au modèle de s'adapter aux motifs temporels uniques des données de chaque patient. Le mécanisme favorise également la parcimonie dans le temps, en se concentrant sur les points de données critiques sans être excessivement influencé par les mesures moins informatives ou peu fréquentes. Objectivement, cette recherche apporte les contributions suivantes :

- Cadre innovant pour gérer les irrégularités temporelles : AMITA fait progresser les mécanismes standard des LSTM en intégrant l'intervalle de temps entre les événements, la fréquence des mesures et les données contextuelles de l'historique médical du patient. Cela permet des ajustements nuancés des cellules de mémoire, réduisant l'influence des mémoires précédentes avec les délais croissants tout en préservant les impacts des événements passés significatifs. De plus, nous intégrons des mécanismes à double porte pour mieux capturer les effets des interventions cliniques sur les états actuels de la maladie.

- Amélioration de la porte d'oubli : Reconnaissant l'irrégularité de l'acquisition des données en santé, nous avons affiné la porte d'oubli d'AMITA pour prendre en compte non seulement le timing, mais aussi la fréquence des événements et le contexte historique plus large de chaque patient. Cette amélioration aide le modèle à s'adapter aux chronologies uniques des patients, en veillant à se concentrer sur les données critiques tout en minimisant l'impact des points de données moins pertinents ou peu fréquents.

- Approche innovante pour améliorer l'interprétabilité : Nous avons proposé une approche novatrice pour améliorer l'interprétabilité en réalisant une analyse approfondie qui intègre à la fois les valeurs d'attention et les poids de fréquence pour chaque caractéristique. Cette méthode nous a permis d'identifier et de mettre en évidence les caractéristiques les plus critiques pour chaque tâche de prédiction, y compris la mortalité en unité de soins intensifs (USI) et la durée de séjour (LOS).

9.5 Résultats

Nous avons validé l'efficacité de nos modèles proposés à travers des expériences empiriques menées sur les deux ensembles de données cliniques réels et trois ensembles de séries temporelles réelles. Les résultats ont montré que nos modèles MWTA-LSTM et AMITA surpassent les modèles actuels de pointe ainsi que d'autres bases robustes.

Dans le domaine de la prédiction de la mortalité à l'hôpital et en unité de soins intensifs (réanimation), nos modèles obtiennent régulièrement des scores F1 supérieurs, dépassant de 10 % les modèles de pointe. De plus, l'intégration de la stratégie ASP améliore les performances de MWTA-LSTM de 3,2 %.

Pour la tâche de prédiction de la durée de séjour (LOS), nous évaluons l'exactitude prédictive en utilisant l'erreur absolue moyenne (MAE) et l'erreur quadratique moyenne (RMSE). Nos modèles montrent des améliorations significatives par rapport aux approches de base, notre meilleur modèle atteignant un MAE de 1,21 jours comparé à 2,11 jours pour les modèles de base. Avec le modèle AMITA, nous obtenons des résultats encore meilleurs, avec des valeurs de RMSE et de MAE inférieures à 1 jour, soit spécifiquement 20,88 heures et 13,2 heures respectivement, et un coefficient de détermination ajusté (Adjusted R^2) de 74 % par rapport aux données réelles (étiquettes réelles).

9.6 Conclusion & Perspectives

Dans cette thèse, nous présentons trois contributions majeures, comme indiqué ci-dessous, MWTA-LSTM et AMITA représentent des avancées significatives dans le domaine de la santé personnalisée, se distinguant des modèles existants tels que [5, 24, 124, 171]. Contrairement à leurs prédécesseurs, MWTA-LSTM et AMITA gèrent plus efficacement les échantillonnages irréguliers en utilisant le temps écoulé, la fréquence des mesures et la dernière observation, ce qui permet de résoudre le problème des irrégularités temporelles.

Ils extraient de manière autonome les caractéristiques pertinentes des dossiers médicaux en intégrant des mécanismes de décroissance temporelle et de paramétrisation, gérant efficacement les dynamiques temporelles des données de santé. Cette adaptabilité permet à MWTA-LSTM et AMITA de mieux gérer les moments et les observations irréguliers. Ils étendent les capacités des LSTM en incorporant les intervalles de temps entre les événements, la fréquence des mesures et le contexte du patient. Cette amélioration permet des ajustements nuancés des cellules de mémoire, réduisant l'influence des événements éloignés tout en conservant les événements passés significatifs. Les mécanismes à double porte renforcent la capacité d'AMITA à capturer l'impact des interventions cliniques sur les états pathologiques actuels, avec une porte d'oubli affinée prenant en compte le timing des événements, la fréquence et le contexte historique plus large.

Enfin, nous proposons une approche innovante pour améliorer l'interprétabilité d'AMITA à travers une analyse complète incorporant les valeurs d'attention et les poids de fréquence pour chaque caractéristique. Cette méthode identifie les caractéristiques critiques pertinentes pour les tâches de prédiction telles que la mortalité en réanimation et la durée de séjour (LOS), assurant transparence et compréhension des prédictions du modèle. En ce qui concerne MWTA-LSTM, nous avons introduit un mécanisme de pooling adaptatif innovant pour gérer les valeurs aberrantes dans les dossiers des patients, garantissant des inférences robustes sur les résultats, distinguant ainsi mos modèles des approches existantes [5, 124, 171].

MWTA-LSTM et AMITA ont montré une performance supérieure dans la prédiction des résultats de santé tels que la mortalité et la durée de séjour à travers divers ensembles de données. Les évaluations comparatives soulignent leur efficacité par rapport aux modèles les plus récents [130], comme détaillé dans Table 7.9.

Comme toute approche basée sur l'apprentissage automatique (ML), nos méthodes présentent certaines limites. Cependant, nos approches montrent des capacités prometteuses et ouvrent plusieurs nouvelles voies pour la recherche future.

Actuellement, les prédictions sont effectuées exactement 24 et 48 premières heures après l'admission. Cependant, permettre des prédictions en temps réel, idéalement toutes les heures ou après chaque intervention du personnel médical, serait fortement souhaitable. Ce passage aux prédictions en temps réel pourrait améliorer considérablement la réactivité et l'efficacité des interventions cliniques. De plus, nos approches ne sont pas explicitement conçues pour remplacer les valeurs manquantes habituelles. Cet travail pourrait être élargi en intégrant des architectures avancées d'apprentissage profond pour prédire les valeurs manquantes et faire des prédictions simultanément par le biais de l'apprentissage multi-tâche, plutôt que de s'appuyer sur des méthodes d'imputation simples.

Université de Strasbourg

Mamadou Ben Hamidou CISSOKO

Adaptive Time-Aware LSTM for Predicting and Interpreting ICU Patient Trajectories from Irregular Data

École doctorale Mathématiques, sciences de l'information et de l'ingénieur | ED 269
Université de Strasbourg

Résumé

En médecine prédictive personnalisée, modéliser avec précision la maladie et les processus de soins d'un patient est crucial en raison des dépendances temporelles à long terme inhérentes. Cependant, les dossiers de santé électroniques (DSE) se composent souvent de données épisodiques et irrégulières, issues des admissions hospitalières sporadiques, créant des schémas uniques pour chaque séjour hospitalier. Par conséquent, la construction d'un modèle prédictif personnalisé nécessite une considération attentive de ces facteurs pour capturer avec précision le parcours de santé du patient et aider à la prise de décision clinique.

LSTM sont efficaces pour traiter les données séquentielles comme les DSE, mais ils présentent deux limitations majeures : l'incapacité à interpréter les résultats des prédictions et à prendre en compte des intervalles de temps irréguliers entre les événements consécutifs. Pour surmonter ces limitations, nous_introduisons de nouveaux réseaux neuronaux à mémoire dynamique profonde appelés Multi-Way Adaptive et Adaptive Multi-Way Interpretable Time-Aware LSTM (MWTA-LSTM et AMITA), conçus pour les données séquentielles collectées de manière irrégulière. L'objectif principal des deux modèles est de tirer parti des dossiers médicaux pour mémoriser les trajectoires de maladie et les processus de soins, estimer les états de maladie actuels et prédire les risques futurs, offrant ainsi un haut niveau de précision et de pouvoir prédictif.

Keywords : EHR, LSTM, Irrégularité temporelle, Modélisation de la trajectoire de santé du patient, Médecine prédictive personnalisée, Prédiction de la mortalité et de la durée de séjour, Prévision des séries temporelles.

Résumé en anglais

In personalized predictive medicine, accurately modeling a patient's illness and care processes is crucial due to the inherent long-term temporal dependencies. However, Electronic Health Records (EHRs) often consist of episodic and irregularly timed data, stemming from sporadic hospital admissions, which create unique patterns for each hospital stay. Consequently, constructing a personalized predictive model necessitates careful consideration of these factors to accurately capture the patient's health journey and assist in clinical decision-making.

LSTM networks are effective for handling sequential data like EHRs, but they face two significant limitations: the inability to interpret prediction results and to take into account irregular time intervals between consecutive events. To address these limitations, we introduce novel deep dynamic memory neural networks called Multi-Way Adaptive and Adaptive Multi-Way Interpretable Time-Aware LSTM (MWTA-LSTM and AMITA) designed for irregularly collected sequential data. The primary objective of both models is to leverage medical records to memorize illness trajectories and care processes, estimate current illness states, and predict future risks, thereby providing a high level of precision and predictive power.

Keywords : EHR, LSTM, Timing Irregularity, Patient's health trajectory modeling, Personalized predictive medicine, Mortality and Length of stay Prediction, Time series forecasting