Optimized deep learning-based multimodal method for irregular medical timestamped data

Sara Rabhi

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAS003

Thèse de doctorat

Optimized Deep Learning-Based Multimodal method for Irregular Medical Timestamped data

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom SudParis

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat : Mathématiques et informatique

Thèse présentée et soutenue à Palaiseau, le 02/02/2022, par

SARA RABHI

Composition du Jury :

Sylvain LE CORFF Professeur, Télécom SudParis (TIPIC)	Président
Zahia GUESSOUM Professeure, Université de Reims Champagne-Ardenne (STIC)	Rapporteur
Adlen KSENTINI Professeur, EURECOM (COMMUNICATION SYSTEMS)	Rapporteur
Frédéric BLANCHARD Maître de conférences, Université de Reims Champagne-Ardenne (CReSTIC)	Examinateur
Djamal ZEGHLACHE Professeur, Télécom SudParis (RS2M)	Directeur de thèse
Didier SCHWAB Maître de conférences, Université Grenoble Alpes (GETALP)	Examinateur

Abstract

The wide adoption of Electronic Health Records (EHR) in hospitals' information systems has led to the definition of large databases grouping various types of data such as text clinical notes, longitudinal medical events, and tabular patient information. However, the data records are only filled during medical consultations or hospital stays that depend on the patient's health, state, and local habits. A system that can leverage the different types of data collected at different time scales is critical for reconstructing the patient's health trajectory, analyzing his history, and consequently delivering better clinical care. This thesis work addresses two main challenges of medical data processing: a) learning to represent the sequence of medical observations with irregular elapsed time between consecutive visits and b) optimizing the extraction of medical events from clinical notes. Our main goal is to design a multimodal representation of the patient's health trajectory to solve clinical prediction problems.Our first work built a generic framework for modeling irregular medical time series to evaluate the importance of considering the time gaps between medical episodes when representing a patient's health trajectory. To that end, we conducted a comparative study of sequential neural networks and irregular time representation techniques. The clinical objective was to predict retinopathy complications for type 1 diabetes patients in the french database CaRéDIAB (Champagne Ardenne Réseau Diabetes) using their history of HbA1c measurements. The study results showed that the attention-based model combined with the soft one-hot representation of time gaps led to theAUROC score of 88.65% (specificity of 85.56%, sensitivity of 83.33%), an improvement of 4.3% when compared to the LSTM-based model. Motivated by these results, we extended our framework to shorter multivariate time series and predicted in-hospital mortality for critical care patients of the publicly available MIMIC-III dataset. The proposed architecture, Hierarchical Time-aware Transformer (HiTT), improved the AUROC score by 5% over the vanilla Transformer baseline. In the second step, we were interested in extracting relevant medical information from clinical notes to enrich the patient's health trajectories. Particularly, Transformer-based architectures have shown encouraging results in medical information extraction tasks. However, these large

models often require a large annotated corpus. This requirement is hard to achieve in the medical field as it necessitates access to private patient data and high expert annotators. To reduce annotation costs, we explored active learning strategies that have been shown to be effective in many tasks, including text classification, information extraction, and speech recognition. In addition to existing methods, we defined a Hybrid Weighted Uncertainty Sampling (HWUS) active learning strategy that takes advantage of the contextual embeddings learned by the Transformer-based approach to measuring the representativeness of samples. A simulated study using the publicly available i2b22010 challenge dataset showed that our proposed metric reduces the annotation cost by 70% to achieve the same performance score as passive supervised learning.Lastly, we combined multivariate medical time series and medical concepts extracted from clinical notes of the MIMIC-III database to train a multimodal transformer-based architecture. The test results of the in-hospital mortality task showed an improvement of 5.3% when considering additional text information. This thesis contributes to patient health trajectory representation by alleviating the burden of episodic medical records and the manual annotation of free-text notes. In a nutshell, this research has three practical contributions: (1) Supporting e-Health systems such as reporting, reasoning, and efficient decision-making to benefit the overall patient management. (2) Benefiting the research in medical informatics by facilitating the development of state-of-the-art deep learning temporal models and the collection of rich annotated corpora from clinical free text resources. (3) Aiming at advancing machine learning research in the medical domain by developing an effective multimodal Transformer-based architecture for accurate health trajectories representation and an innovative domain-independent AL query strategy.

Résumé

L'adoption des dossiers médicaux électroniques (DME) dans les systèmes d'information des hôpitaux a conduit à la définition de bases de données Big Data regroupant divers types de données telles que des notes cliniques textuelles, des événements médicaux longitudinaux et des informations tabulaires sur les patients. Toutefois, les données ne sont renseignées que lors des consultations médicales ou des séjours hospitaliers. La fréquence de ces visites varie selon l'état de santé du patient et des habitudes locales. Ainsi, un système capable d'exploiter les différents types de données collectées à différentes échelles de temps est essentiel pour reconstruire la trajectoire de soin du patient, analyser son historique et, par conséquent, délivrer des soins plus adaptés.

Ce travail de thèse aborde deux défis principaux du traitement des données médicales : (1) Représenter la séquence des observations médicales à échantillonnage irrégulier et (2) optimiser l'extraction des événements médicaux à partir des textes de notes cliniques. Notre objectif principal est de concevoir une représentation multimodale de la trajectoire de soin du patient afin de résoudre les problèmes de prédiction clinique.

Notre premier travail porte sur la modélisation des séries temporelles médicales irrégulières afin d'évaluer l'importance de considérer les écarts de temps entre les visites médicales dans la représentation de la trajectoire de soin d'un patient donné. À cette fin, nous avons mené une étude comparative entre les réseaux de neurones récurrents, les modèles basés sur l'architecture Transformer et les techniques de représentation du temps. De plus, l'objectif clinique était de prédire les complications de la rétinopathie chez les patients diabétiques de type 1 de la base de données française CaRéDIAB (Champagne Ardenne Réseau Diabète) en utilisant leur historique de mesures HbA1c. Les résultats de l'étude ont montré que le modèle Transformer , combiné à la représentation 'Soft-One-Hot' des écarts temporels a conduit à un score AUROC de 88,65% (spécificité de 85,56%, sensibilité de 83,33%), soit une amélioration de 4,3% par rapport au modèle basé sur l'architecture LSTM . Motivés par ces résultats, nous avons étendu notre étude à des séries temporelles multivariées plus courtes et avons prédit le risque de mortalité à l'hôpital pour

les patients admis en soins intensifs présents dans la base de données MIMIC-III. L'architecture proposée, Hierarchical Time-aware Transformer (HiTT), a amélioré le score AUC de 5% par rapport à l'architecture de base Transformer .

Dans la deuxième étape, nous nous sommes intéressés à l'extraction d'informations médicales pertinentes à partir des comptes rendus médicaux afin d'enrichir la trajectoire de soin du patient. En particulier, les réseaux de neurones basés sur le module Transformer ont montré des résultats encourageants dans l'application d'extraction d'informations médicales. Cependant, ces modèles complexes nécessitent souvent un grand corpus annoté. Cette exigence est difficile à atteindre dans le domaine médical car elle nécessite l'accès à des données privées de patients et des annotateurs experts. Pour réduire les coûts d'annotation, nous avons exploré les stratégies d'apprentissage actif qui se sont avérées efficaces dans de nombreuses tâches, notamment la classification de textes, l'analyse d'image et la reconnaissance vocale. En plus des méthodes existantes, nous avons défini une stratégie d'apprentissage actif, nommée Hybrid Weighted Uncertainty Sampling (HWUS), qui utilise la représentation cachée du texte donnée par le modèle Transformer pour mesurer la représentativité des échantillons. Une simulation utilisant l'ensemble de données du challenge i2b2-2010 a montré que la métrique proposée réduit le coût d'annotation de 70% pour atteindre le même score de performance que l'apprentissage supervisé passif.

Enfin, nous avons combiné des séries temporelles médicales multivariées et des concepts médicaux extraits des notes cliniques de la base de données MIMIC-III pour entraîner une architecture multimodale. Les résultats du test ont montré une amélioration de 5,3% en considérant des informations textuelles supplémentaires.

Remerciements

Tout d'abord, Je souhaite commencer par remercier ma mère Souad, qui m'a encouragé tout le temps, en me dédiant beaucoup d'amour et de soutien pour que je puisse réaliser ce travail. Je remercie aussi mon père Hassan, pour m'avoir accompagné durant toutes ces années d'études.

D'autre part, je remercie mon encadrant de thèse, Djamal Zeghlache, pour son soutien et son accompagnement. Merci de m'avoir accordé la chance de poursuivre le parcours de doctorat qui m'a permis d'approfondir mes connaissances et de m'avoir donné goût à la recherche. Je tiens aussi à remercier les membres de mon jury de thèse. Merci à Zahia Guessoum et Adlen Ksentini d'avoir accepté d'être rapporteurs de ma thèse et à Didier Schwab, Frédéric Blanchard et Sylvain Lecorff d'avoir accepté de l'évaluer. Je remercie Sara Barraud, Brigitte Delemer et toute l'équipe de l'hôpital Robert Debré (CHU de Reims) de m'avoir accueilli dans leur laboratoire. Ce partenariat a été très enrichissant, pleins de belles rencontres et m'a permis de mieux comprendre les enjeux de la donnée de santé. Merci à Bruno Defude pour ses conseils, son écoute et son aide pour la gestion administrative et logistique de mon travail de thèse.

Je remercie Chris Green, Nimalan, Shivam et toute l'équipe de Twitter Cortex de m'avoir accueilli pendant un stage de trois mois. Cette expérience américaine a été très formatrice et m'a permis d'initier des liens amicaux et scientifiques durables. Un grand merci aussi à tous les collègues de NVIDIA pour une expérience de stage pleines de belles rencontres et de la confiance qu'ils m'accordent pour rejoindre l'équipe de recherche Merlin-team . En particulier, je remercie Even Oldridge, mon manager, pour tout son soutien et ses conseils avisés qui m'ont permis d'avancer sur mes travaux de thèse lors de la période très difficile de la pandémie du COVID-19. Je remercie aussi Mohamed Ndaoud pour m'avoir donné goût à la recherche et pour le partage de son savoir sans fond.

Je remercie tous les membres de ma famille et tous mes amis qui m'ont soutenu tout au long de cette expérience, dans ses hauts comme dans ses bas. Enfin, merci à Mehdi Iraqi, sans qui je ne serai pas arrivée au bout. Merci de m'avoir accompagné dans ce voyage difficile.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, the availability of medical data has increased thanks to the wide adoption of Electronic Healthcare Records (EHR) in hospitals' information systems. These records store all the transactions information between patients and healthcare providers during their visit or stay in the hospital. EHRs combines three types of features: The first type is structured data (such as the patient's age, the admission date, the duration of stay, measurements, and discrete medical codes), The second class is semi-unstructured data and consists of a short free-text column storing specific information (such as doctors comments, and the description of other non-standardized conditions in the database system), and the third type is unstructured data, which refers to narrative clinical notes written by medical practitioners to report the patient's state and the medical events occurred during his stay or visit (including family history, diagnoses, diseases, procedures, and medications). The first role of electronic records is to provide up-to-date information about the patient in less time and assist medical practitioners in delivering a higher quality of care by facilitating the information exchange between health practitioners. Moreover, These centralized systems are often deployed in one or a group of medical centers over a long period leading to an extensive database of patient records with several years of medical history. These datasets are a rich source of information for large-scale statistical analysis and represent an opportunity to bridge the gap between medical

analysis and machine learning techniques that often require a large set of observations to train and achieve the best performance. Consequently, several research studies have found a secondary use of this data for conducting predictive analysis to better understand diseases evolution and build health monitoring systems helping doctors deliver better care for their patients [1].

In particular, deep learning methods have become a relevant choice for conducting prediction tasks in various domains such as natural language processing (NLP) [2, 3, 4], image analysis [5, 6] and time series modelling [7, 8, 9]. Moreover, the construction of large private data warehouses [10, 11, 12] and the publication of open-source medical databases such as MIMIC III or i2b2 [13, 14] allowed researchers to adopt and adapt these methods for solving clinical prediction problems such as risk prediction [15, 16, 17], intervention recommendation [11, 18], disease progression [19, 20] or patient sub-typing [10, 21]. Majority of these methods focused on modeling one type of input data, either tabular, textual, longitudinal, or image, whereas other methods [22, 23, 24] combined several types and showed that a comprehensive patient representation helped achieve higher scores.

1.1 Challenges of medical data

Effective modeling of medical prediction tasks must consider the challenges of processing the highly variable observational data contained in real-world clinical databases. We summarize these challenges in six categories: patient privacy-preserving, small and incomplete datasets, cost-effective annotation process, non-standardized data structures, irregular health trajectories, and multimodal data. Indeed, EHRs comprise highly sensitive personal information about patients and their conditions. Leveraging these data for research requires a de-identification step that protects patients' sensitive attributes while sharing informative data relevant to deep learning research studies. Furthermore, the protection of sensitive data and the non-existence of a centralized system collecting data from multiple medical centers lead to the definition of small datasets specific to each hospital. These small datasets limit the capacity of research to define high-performance and generalizable deep learning-based predictive model [25]. One particular requirement of deep learning methods is the collection of annotated data that guides the supervised learning process of such complex

models. This process is cost-effective when considering clinical data as it requires high experts with sufficient medical knowledge leading to even smaller training sets. On the other hand, hospitals use different standards to organize the medical data in their information systems and different biomedical ontologies [26, 27] to classify the concepts such as diseases, procedures, and treatments. These systems' differences represent an additional barrier for designing generic models ready for deployment in diverse health systems. Even when considering a unified data source, time irregularities are another common phenomenon of this real-world clinical data. Indeed, the medical observations are recorded episodically, depending on patient visits to the hospital, producing a history of care that varies from one patient to another and depends on each patient's health state and local habits. Consequently, the produced data includes irregular health trajectories with variable lengths and different periods between consecutive observations. Lastly, each time point in the health trajectory represents an episode of care with various types of data (such as text reports, treatment prescriptions, lab test orders, lab results, and records of medical parameters, diagnostics, and administrative codes) produced by several care providers during the patient's management. Leveraging all these various types in one general model is challenging as it requires the design of a multimodal system capable of learning important information from each entry and avoiding data redundancy.

1.2 Research context

Several research works were published to address medical data challenges when modeling a single or multiple types of data, either image, text, or longitudinal. To protect the patient's personal information, Andrew et al. [28] analyzed a wide range of privacy-preserving techniques (such as homomorphic encryption and differential privacy) applied to structured EHRs data for computing deep learning prediction scores. Moreover, preserving the sensitive information within the text of clinical notes is also an active research area, and yang et al. [29] devised a systematic review summarizing the deep learning methods proposed for automatic de-identification. To address the limitation of small datasets, several works proposed techniques based on Transfer Learning [30, 31, 32] to leverage the knowledge learned by a pre-trained model and extend it to the new set of

data. Another common solution is multi-task learning [31, 33, 34] that improves generalization by leveraging the domain-specific information contained in the training signals of related tasks.

Re-ordering the information contained within the EHRs data is critical to assess the patient's care pathway and to understand the evolution of the disease. The heterogeneity of data and the irregular health trajectories are two main challenges for defining an accurate temporal representation of the patient's timeline. Most of the existing works rely on the longitudinal data in the form of administrative codes, and numerical values generated at every care episode [10, 35, 36, 33]. Hence, they defined the timeline as a multivariate time series. At the same time, other works [22, 11, 23] considered the clinical notes produced during each admission to enrich the patient's care pathway. The first group [23, 22] learned an embedding representation of the whole note and added the resulting vector as an additional feature to the time series to leverage text information. On the other hand, the second group [11] defined a hybrid model where the first stage is an NLP model that extracts the medical concepts. These concepts are then added to the time series for the second-stage learning that represents the patient's health trajectory. Recent advancement of the NLP field has led to the definition of robust architectures enabling to learn contextual embeddings of words and achieve high-performance scores in downstream tasks such as concept extraction. Especially, ClinicalBERT [37] and BioBERT [38] have adapted the prominent NLP model BERT [39], which is based on the Transformer architecture [40]. First, they pre-trained the model on large corpora of medical texts to get the contextual representation of words, then fine-tuned this pre-trained architecture on various supervised downstream tasks. In particular, these models have shown improved performance scores in the medical concepts extraction tasks (ranging between 78% and 94%).

However, defining such performant models relies on the availability of extensive annotated clinical notes with a consequent number of examples of each class of interest. Research efforts [41, 42, 43, 44] to build such annotated corpora have been rising during the last decade, and several works published the annotation guidelines that allow them to produce high-quality labeled data. This process is time-consuming and costly because it often requires manual annotation by medical

experts that have limited availability. Active learning [45] is a promising research direction that has shown its effectiveness in image annotation and was extended to several applications, such as medical text annotation. The objective is to reduce the amount of training data to be manually annotated by selecting the examples that accelerate the learning iterations of the deep learning model. It places the medical expert in the heart of an iterative process by allowing him to correct the model's predictions then re-train the model taking into account his feedback. The core component of the active learning strategy is the definition of a metric, often referred to as utility function, that ranks the predictions of the model and selects the most informative examples for the next re-training iteration. The two prominent sampling strategies are uncertainty-based, and density-based [46, 47, 48].

The reconstructed timeline is episodic, and the time gaps between consecutive observations vary from one patient to another and even within the health trajectory of the same patient. Most of the current literature [49, 50, 51, 52] is based on a statistical analysis of periodic snapshots of longitudinal medical events with a fixed time interval, monthly or semi-annually. These models require the availability of temporal equally spaced medical events. Consequently, conducting statistical post-analysis of this data involves imputation methods to fill in the missing values. The performance of these methods is highly dependent on the completeness of the patients' times series and the accuracy of the imputation methods. Instead of using data imputation methods to fill the gaps between actual observations, irregularity is also valuable information that we should consider for learning the evolution of the patient's health status. Following that line of thought, recent studies [12, 53, 54, 50] took advantage of the advancement made in sequence modeling and used recurrent neural networks (RNNs) coupled with the representation of the time gap between two consecutive event points to conduct downstream medical tasks such as risk prediction, procedures recommendation, and patient phenotyping.

1.3 Objectives

The main objective of this thesis is to build a multimodal deep learning architecture that leverages various types of information contained in the EHR data and learns to represent the patient's timeline. A subsequent objective is to validate this architecture on real-world clinical application by considering the challenges of such a setting. Mainly, from the challenges presented in section 1.1, we focused on designing a framework to represent time irregularity in a patient's health trajectory and proposed an active learning strategy to reduce the annotation cost of deep learning-based medical concept extraction model.

1.4 Research questions

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

- **RQ1**: How to model irregular time observed in patients' health trajectories?
- **RQ2**: Is it possible to design a generic framework for irregular medical time series modeling?
- **RQ3**: How to represent information within the clinical notes to enrich patients' health trajectories?
- **RQ4**: What is the best Active Learning strategy that reduces the annotation cost of the Transformer-based medical information extraction approach?
- **RQ5**: 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 irregular recording of events and the cost of medical texts annotation.

1.5 Contributions

The first 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 end-to-end classification of irregular time series**. The framework processes numerical and categorical medical events and supports the patient's metadata. Besides, it gathers the state of the art sequential deep learning models and time representation techniques. Using this framework, we conducted **An empirical study of diabetic retinopathy prediction in type 1 diabetes patients** based on a comparative study of 12 temporal neural-based approaches. The data was gathered from the French database CaRéDIAB [52] and consisted of 1,207 highly variable uni-variate medical time series of HbA1c records of type 1 diabetic patients.

In the second part, we represented the information contained within the clinical notes and evaluated their importance in predictive modeling. To that end, we conducted **a comparative study between deep learning and conventional machine learning methods for medical text classification** [15]. The results showed the high effectiveness of DL-based methods when applied to predict the Health Acquired Infection (HAI) from patients' clinical notes. However, the error analysis shows that missing positive cases were due the missing of time management in our model. These findings motivated us to explore information extraction architectures for selecting relevant medical events from each clinical note to enrich the patient timeline. Those techniques often require a large amount of labeled data which is highly cost-effective when processing medical reports. Therefore, our second work aimed to define a **Deep Active Learning strategy to reduce the annotation cost of clinical notes for medical events extraction**. Specifically, we evaluated active learning strategies for transformer-based medical event extraction models.

Finally, the third work consisted of designing a multimodal architecture, Multi-HiTT: **Multi**-modal **Hi**erarchical **T**ime-aware **T**ransformer-based. This architecture leverages all information contained in the patient medical records by combining multivariate event time series, patient static information and the medical concepts extracted from clinical notes to build an accurate patient representation for clinical prediction tasks. The main contribution of this work was **the design of a hierarchical**

temporal and multi-modal patient representation combining structured features and free-text medical concepts. Using the implemented temporal framework, we validated the proposed Multi-HiTT architecture by **studying the in-hospital mortality prediction of patients admitted in critical care units**. We particularly considered 5120 irregular multivariate time series provided by the open-source dataset MIMIC-III [13].

1.6 Outline

This thesis is organized into three 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 and active learning strategies for named entity recognition is detailed in the second chapter. On the other hand, the second part exposes our three main contributions. The first chapter describes the implemented temporal framework that allows medical research teams to conduct comparative studies and select the best DL model for classifying IMTS based on their dataset and prediction task. The second chapter defines a new active learning strategy, Dynamic Hybrid Weighted Uncertainty Sampling (Dynamic-HWUS), that aims to reduce the annotation cost of clinical notes for training Transformer-based named entity recognition models. The third chapter introduces Multi-HiTT architecture that aims to combine the different levels of temporality and types of input data for building an accurate representation of the patient. Lastly, the third part includes three studies that validate our proposed methods and use real-world clinical databases, discusses their results, concludes the thesis work, and introduces recommendations for future work.

1.7 Notations

We note $p \in \mathbf{P}$ the set of patients considered in a medical study. We define the multivariate medical time series of patient p as follows:

- Multi-variate time series $(x_{p,t})_{1 \leq t \leq N}$ consists of a sequence of states $x_{p,t}$ where $x_{p,t} \in \mathbf{R}^q$ is a

dense embedding vector that represents the events with different types observed at a discrete time step t, q is the vector space dimension and N is the number of steps generally equal to the number of patient's visits.

- The state vector at a discrete time step t can be represented as a combination of three vectors: $x_{p,t} = [n_{p,t}, u_{p,t}, z_{p,t}, d_p]$, where $n_{p,t}$ is the representation vector of text notes, $u_{p,t}$ denotes the vector of numerical values, $z_{p,t}$ is related to the encoded ids of medical events, d_p corresponds to static patient information such as demographics. For simplifying the notations, we refer to the patient vectors at timestamp t as $x_t = [n_t, u_t, z_t, d_p]$

  The purpose of multi-variate medical time series representation learning is to define a dense embedding $R_p \in \mathbf{R}^d$ that comprises the temporal dynamics and the relevant medical information of $(x_t)_{1 \leq t \leq N}$. This representation is then validated on predictive supervised tasks by finding optimal f^* for $f^*(R_p) = y_p$, with $y_p \in \mathbf{Y}$ is the true label to predict for patient p.

Part I

Background and related work

Chapter 2

Irregular Medical Time Serie (IMTS)

2.1 Introduction

Our work is inspired by two lines of related research: sequence modeling using deep learning networks (Section 2.2) and the representation of time irregularity in highly variable and episodic medical time series (Section 2.3). This chapter describes the different 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 (Table of Appendix E classifies the existing works according to the data and proposed DL architecture).

2.2 Neural networks for medical time series

Researchers in natural language processing (NLP) have proposed novel deep learning architectures for language modeling using the sequence of words and inspired many researchers and practitioners

to use and adapt these methods to event time series modeling [55, 56, 7, 8, 9, 50]. These neural architectures offer a more flexible end-to-end learning approach when compared with standard time series and machine learning models. In this section, we present the two most used neural-based models for event time series representation: recurrent neural networks (RNN) [57] (including long short-term memory (LSTM) [58] and Gated Recurrent Unit (GRU) [57]) and self-attention based models so known as Transformer) [40]. They offer the flexibility of processing the raw sequence of events as input to the model without any feature engineering or aggregation methods. They also enable processing sequences of multi-dimensional vectors and represent the temporal dynamics and the information flow in time series.

2.2.1 Recurrent neural networks

The vanilla recurrent neural network is an extension of the hidden Markov model that learns to represent the hidden state of a given sequence. At each time step t, the hidden representation, h_t, of the event is computed based on its current value and the value of the previous hidden states $h_{j<t}$ as follow:

$$h_t = g(Wx_t + Uh_{t-1}),$$

where g is a smooth activation function such as sigmoid function and W, U are learnable parameters. This internal state offers the possibility to learn past information dependencies. However, RNN has limitations on learning and predicting with long sequences as it cannot capture the long-range dependencies. These problems are often referenced as vanishing, and exploding gradient [59]. The LSTM [58] addresses these problems by adding to the RNN unit a memory cell state that captures the relevant information to retain from the past and the output gate that decides by how much to update the current hidden state of the unit. The update of the internal state is defined in five steps:

- The forget gate: $f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$
- The input gate: $i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$

- The cell sate: $c_t = f_t * C_{t-1} + i_t * \tilde{C}_t, where\ \tilde{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_f)$
- The output gate: $o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$
- The updated hidden state: $h_t = o_t * tanh(C_t)$

GRU [57] is a simplification of LSTM removing the memory cell and introducing two parametric gates: An "update gate", combining the forget and input gates, and a "reset gate". The update process is reduced to three steps that are expressed as follows:

- The update gate: $z_t = \sigma(W_z.[h_{t-1}, x_t])$
- The reset gate: $r_t = \sigma(W_r.[h_{t-1}, x_t])$
- The updated hidden state: $h_t = (1 - z_t)h_{t-1} + z_t * \tilde{h}_t$, where $\tilde{h}_t = tanh(W.[r_t * h_{t-1}, x_t])$

The empirical study conducted in [60] showed the advantages of the gating units over the more traditional recurrent units regarding faster convergence and better performance scores. However, the comparison results of LSTM and GRU conducted by this work [60] were not conclusive, suggesting that the type of gated units highly depends on the dataset and the prediction task.
We simply denote the updated hidden state at time step t returned by LSTM and GRU as $h_t = LSTM(x_t, h_{t-1})$ and $h_t = GRU(x_t, h_{t-1})$, respectively. The patient representation R_p is then generally set to h_N, the hidden representation of the last observation, as it captures the combination between long-range past information observed in the patient health trajectory and the most recent medical context.
For a better representation of both contexts, left-to-right and right-to-left, in a sequence, Bi-directional recurrent neural networks learned a forward $\overrightarrow{h_f}$ and backward $\overleftarrow{h_b}$ hidden representations. The forward RNN reads the input sequence from x_1 to x_N (old to recent) and calculates a sequence of forward hidden states $\overrightarrow{h_1}, ..., \overrightarrow{h_N}$. The backward RNN reads the visit sequence in the reverse order, i.e., from x_N to x_1 (recent to old), resulting in a sequence of backward hidden states $\overleftarrow{h_1}, ..., \overleftarrow{h_N}$. By concatenating the forward hidden state and the backward one, we obtain the final hidden representation as $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$.

2.2.2 Attention mechanism

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:

$$R_p = Agg(h_1, ..., h_N),$$

where the most straightforward aggregation function Agg is the average pooling. As not all visits have the same impact on the patient overall health state, learning the importance weight α_j of each time observation x_t represented by h_t is very crucial for an accurate patient representation that could be then expressed as

$$R_p = \sum_{j=1}^{N} \alpha_j h_j$$

The introduction of the attention mechanism [61] in NLP has shown to be a powerful tool in developing DL for text classification [62, 63]. In a nutshell, vanilla attention in NLP is defined as a vector of importance weights of words. To predict the score of the downstream task, the attention vector estimates how each word is strongly correlated with the given task, and the final prediction is based on the sum of the representations of the words weighted by their attention scores. Transposing the attention mechanism to medical time series, the weights α_j can be expressed as:

$$\alpha_j = \frac{\exp(a_j)}{\sum_{k=1}^{N} exp(a_k)},$$

where: $a_j = score(h_j, Q)$ and $Q \in \mathbf{R^d}$ is the parametric context vector that can be viewed as a fixed query asking for the "most informative visits" from the input sequence. Different scores functions were proposed in NLP [64, 61, 65, 40].

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 γ_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:

$$R_p = \sum_{j=1}^{N} \sum_{i=1}^{K_j} \alpha_j \gamma_i^j h_j$$

2.2.3 Transformer: self-attention mechanism

Instead of computing absolute importance α_j of the position j related to the entire input sequence, the self-attention mechanism [40], also known as intra-attention, defines a local score. This score measures the relative importance of the position j taking into account the surrounding context positions $k < j$ and/or $k > j$ of a single input sequence. Their proposed score function computes how each position j (value vector $v_j \in \mathbf{R}^d$) is strongly correlated with the given output at j (query vector $q_j \in \mathbf{R}^d$) taking into account all surrounding positions (key matrix $K \in \mathbf{R}^{N \times d}$). In sequence modeling, both the keys and values are the neural network hidden states. The score function of timestamp j returns relative weights overall positions and is formulated as:

$$a_j = \frac{1}{\sqrt{d}} q_j^T K \in \mathbf{R}^N$$

The Transformer, an encoder-decoder architecture, is entirely built upon self-attention mechanisms and was proposed by Vaswani et al. [40] as a replacement of 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.

2.2.4 Differences between the RNN and the Transformer architectures

One of the computational bottlenecks suffered by RNNs is the sequential processing of text, and the computational cost of hidden states updates grows with the increasing length of the sequence. On the other hand, the Transformer block computes the contextual representation of each word independently, allowing for more parallelization than RNN, making it possible to efficiently train huge models on large amounts of data on GPUs. Consequently, a new SOTA of performances is reached on various NLP downstream tasks using a large stack of transformer blocks [66].

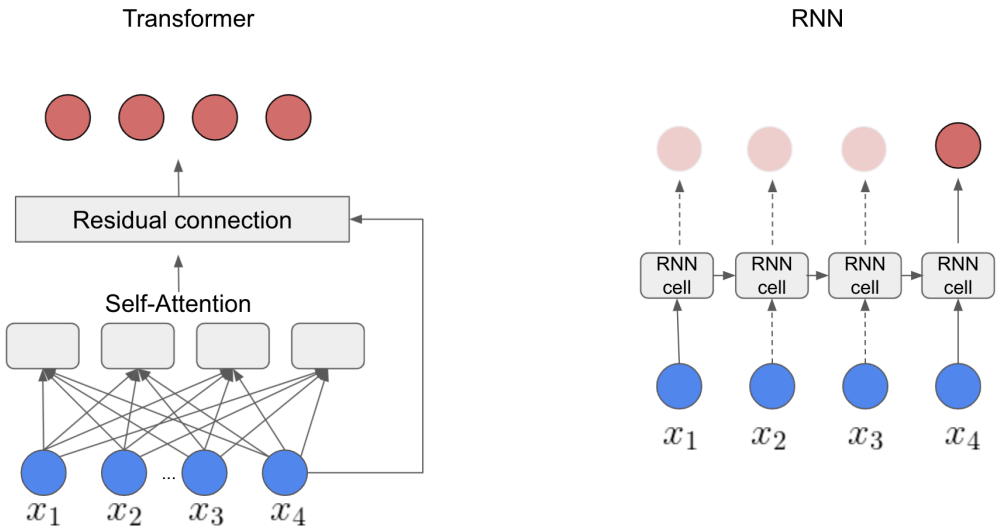


Figure 2.1: Schematic representation showing the difference between the recurrent mechanism that processes one entry at each step and the self-attention that computes the hidden representation of all entries in parallel.

2.3 Modeling temporal mechanisms of irregular medical time series

Rather than defining the longitudinal medical data as a sequence of time intervals using a finite time window segmentation, the model should account for the irregular periods between visits as additional information to the event itself [67]. 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. [68, 69] (2) Extending the original neural architecture by injecting the time gaps as an additional parameter of the model [69, 53, 50, 70, 71].

2.3.1 Time as an additional input variable

The work [69] introduced two methods for representing the time-gaps and tested it on five datasets from various domains, among them the MIMIC II dataset that includes 650 sequences of encoded medical events (204 disease codes) recorded in the intensive care unit. The two representations are

TimeMask: The authors considered the continuous elapsed time as a signal that gives context information about the medical event. Formally, they transformed the numerical value into a probability score vector with the same shape as the event vector representation and computed their element-wise multiplication to *contextualize* the medical event embeddings.

TimeJoint: Instead of using time as a signal, elapsed time contains relevant information as important as the event itself. For that reason, they proposed to represent the time as an additional vector fed to the neural methods, both vectors, the event, and time.

This work shows that none of the methods for using time can improve accuracy on the MIMIC II dataset. In contrast, for all remaining datasets, *TimeJoint* enables a significant gain of performance compared with simply using the scalar value of time in sequential models. We argue that the results observed on MIMIC-II may suffer from the small size of the dataset, limiting the performance of more complex architectures. We also noticed that all tested methods returned the same range of performances.

2.3.2 Temporal based neural architecture

Instead of considering the elapsed time as an additional variable, studies focused on modeling the temporal structure of the time series as another channel inherent to the neural architecture. Researchers have proposed temporal-based architectures that we categorize into four classes.

The first group of research considered a ***time decay*** formulation to mimic the decrease of acute conditions' effect through time. Following that line of thought, C-LSTM [53] and T-GRU [72] represent time using exponential decay function within the forget and the memory cells of LSTM and GRU, respectively. The mathematical formulation for irregular time intervals Δ_t between consecutive events is defined as: $w_{\Delta_t} = \frac{1}{log(e+\Delta_t)}$. On the other hand, GRU-D [73] defined the time decay function as an additional model jointly trained with GRU network. This decay rate is then used to weight the previous hidden states when computing the current one. It is defined as: $\gamma_t = \exp^{-max(0, W_\gamma \Delta_T + b_\gamma)}$, where W_γ and b_γ are learnable parameters.

However, the healthcare trajectories of patients comprise different temporal patterns. In fact, during follow-up, the patient could improve the states of some health conditions while others got worse over time and could also develop chronic diseases. The *time decay* does not factor in all these temporal patterns, and C-LSTM proposed a flexible ***parametric time*** introduced in the forget gate to learn the temporal dependencies between these different patterns. Similar to GRU-D, this trainable parameter is expressed as $Q_f q_{\Delta_t}$ and is summed up to the forget gate parameters: Q_f is the weight matrix and q_{Δ_t} is a temporal vector derived from time difference Δ_t.

The third research direction, ***temporal structure segmentation***, addresses the same problem exposed in *parametric time*. Nevertheless, instead of defining one parametric network to learn the overall temporal structure of longitudinal EHR data, researchers aim to explicitly model each temporal dependency component. Particularly, Lee *et al.* [50] segmented the temporal trajectory of patients into three modules: *neural abstraction module* that captures longer-term distant past, *recent context module* that embeds the recent event information using a discriminative projection and finally *periodicity mechanism* that represents periodic events. We note that the explicit modeling of these structures requires a pre-processing step that consists of segmenting the medical events at

a certain temporal granularity (related to temporal window size) and computing the time gaps only between periodic events.

In the Transformer architecture, a positional encoding vector was introduced and summed up to the input words' vectors before computing the self-attention scores to account for the sequential order of words within the sentence. Motivated by modeling irregularity in time series and leveraging the Transformer architectures, recent works [22, 70] focused on defining ***functional time representation*** that replaces the base positional encoding of self-attention to leverage the elapsed time between observations.

2.4 Clinical applications

2.4.1 Downstream tasks

Motivated by the advancement of sequence modeling observed in NLP [57, 61, 40, 39] and their application to time series classification [7, 9], several works were published in the medical domain showcasing the efficiency of using sequential-based neural networks to represent the patient's healthcare trajectory and conduct predictive clinical studies. RNN-based, Attention-based, and Transformer-based models have been successfully applied to many clinical events prediction tasks that we categorized into five major families. The first category is binary classification [53, 74, 73, 67, 75, 76, 72, 33, 76] where the objective is to predict the presence, the absence or the future risk of incidence of a given clinical outcome such as heart failure onset [75, 17, 16, 77], risk of in-hospital mortality [72, 73, 67, 78, 79] and patient readmission [53, 76, 80, 81]. Another type of classification is the multi-label classification, where the objective is to predict multiple outcomes such as the categories of the future diagnosis [82], the severity levels of a given disease [83], and diseases classification [11]. The regression task predicts a real-valued attribute and is used in the medical domain to estimate the temporal progression of the patient's health, filling the missing values of numerical indicators and predicting their future values. For example, the works [84, 22, 23] estimated the length of hospital stay in days, Qingxiong et al. [72] filled the missing numerical rates based on surrounding context and Zhengping et al. [73] predicted the future values of multi-variate continuous outcomes. Another clinical application is medical event prediction [12, 85, 50] that helps doctors in understanding the evolution of the patient's health trajectory and anticipating the occurrence of adverse events for better patient management. Finally, recommendation systems [86, 18] were proposed to assist health practitioners in selecting the most appropriate treatments or procedures based on the patient's history.

2.4.2 Deep Learning architectures

The RNN-based models (LSTMs and GRU networks) are the most used architectures in modeling medical time series [12, 78, 80, 73, 83] and have shown very promising results when compared to conventional machine learning algorithms. Specifically, Choi et al. [17] designed a two-stage process. They first defined the embedding vectors of medical events codes using the skip-gram model [87] then they fed the sequence of medical event embeddings to a GRU layer and generated the patient hidden representation for predicting the risk of heart failure. Esteban et al. [83] used an MLP projection layer to represent constant patient information and model the temporal dynamics of his care history using a GRU network. Then, they concatenated the two hidden representations, static and dynamic, and applied a softmax layer to predict the outcomes of the kidney transplantation procedure. [78] used bi-directional LSTM (BiLSTM) to model the sequence of medical event embeddings and generate the patient representation from the hidden states. They compared their approach with other aggregation methods (average and self-attention pooling methods) and showed that using the hidden representation returned by the BiLSTM yields the best results. More recently, research teams explored the effectiveness of Transformer-based models [33, 88, 79] for modeling the medical time series and showed outperforming results over GRU/LSTM based models. Song et al. [33] represented each episode with a multi-dimensional embedding vector grouping the information of all medical events. Then, they passed the sequences of embeddings through a Transformer encoder with causal attention (the observation t can only attend to past information $j < t$) and a dense interpolation layer for representing the time gaps between consecutive observations. Tipirneni [79] proposed a Transformer-based model with a novel Input Triplet Embedding component that represents the time of the observation t, the features $(f_j)_{1 \leq j \leq K}$ observed at t and their values $(v_j)_{1 \leq j \leq K}$.

2.4.3 Attention mechanism from medical time series

Medical researchers also explored different attention mechanisms proposed in NLP sequence modeling to mimic how doctors attend to a patient's needs and explore the patient record. Usually,

there is a focus on specific clinical information (e.g., key risk factors and medical antecedents) working from the recent events to the further records. The first published works [75, 86, 77, 23] used global attention to define the patient representation as a weighted average over the set of visits' embedding vectors returned by the sequential module. Multi-level attention mechanisms were also proposed to attend to different medical aspects within the same multivariate medical time series [82, 81, 76]. The multi-head self-attention mechanism can automatically capture all nested dependencies between events occurring in the same sequence. Consequently, medical researchers [33, 22, 79] leveraged it for the definition of more fine-grained representation of the patient, achieving better performances on downstream tasks.

2.4.4 Irregular time modeling of clinical data

According to the outlined medical studies, they aim to develop a deep learning system that accurately represents a patient's medical history and uses it in various downstream tasks, including risk prediction, phenotyping, and intervention recommendations. They used sequential deep learning methods like GRU, LSTM, or Transformer to derive the patient's health state vector at a specific point in time, t, by taking into account the evolution of past medical events, such as diseases, procedures, treatments, and numerical indicators. The majority of the proposed works ignored the time irregularity between consecutive visits, however, the recent studies [12, 53, 74, 73, 75, 84, 54, 76, 72] that integrated it have demonstrated its importance in capturing the contextual relationships between visits leading to a better representation of the evolution of the patient. With the exception of the TAPER architecture [22], all of these methods focus on one type of data (numerical, categorical, or text) to represent a patient's timeline. In fact, TAPER [22] included a combination of categorical medical codes, patient demographics, and clinical notes. A limitation of their work is representing the whole clinical text as one piece of information and learning a vector representation of it instead of extracting the relevant concepts that highly impact the downstream clinical application and therefore suppressing all redundant information such as medical sections titles and hospital information without affecting the downstream clinical application. In light of this observation, we

investigated Named Entity Recognition methods to create a deep learning architecture capable of extracting relevant concepts from clinical notes.

Chapter 3

Neural-based architectures and Active Learning strategies for medical events extraction

3.1 Introduction

The importance of the information contained within the clinical notes motivated researchers to build systems for extracting the relevant medical concepts. Several methods were proposed, ranging from rule-based algorithms to complex neural-based architectures. The recent progress made in NLP with the publication of the Transformer [40] architecture has led to the definition of high-performing models for medical named entity recognition tasks. However, these models contain millions of parameters and require a large annotated corpus for training. This requirement is hard to achieve in the medical field as it necessitates access to private patient data and high expert annotators. In comparison, an effective Active Learning algorithm can theoretically achieve exponential acceleration in labeling efficiency and thus reduce the annotation time for medical

experts. First, we present the neural-based methods proposed to solve NER tasks (Section 3.2) and the Section 3.3 exposes the existing AL strategies.

3.2 Named Entity Recognition task (NER)

The NER term stands for named entity recognition and entity extraction, a technique used in natural language processing (NLP) that automatically identifies named entities in a text and categorizes them into pre-defined categories. In medical domain, named entity recognition is used to organise unstructured medical text into a form that can be interpreted by downstream computer algorithms. Examples of downstream applications are automatic extraction of medical codes, build a concise summary of clinical narratives and disease prediction.

Neural-based architectures for NER

Unlike structured data, most existing data are heterogeneous textual notes created to support the primary purpose of care. This unstructured data requires additional processing to make it more suitable for conducting epidemiological research studies or designing administrative support tools. These clinical notes remain a rich source of relevant information about the patient, such as symptoms, the reason for diagnoses, illness evolution trajectory, social situation, care timeline, and medical history. The main objective of Natural Language Processing (NLP) techniques in medicine is to achieve a good accuracy of automatic extraction of these medical concepts [89, 90, 91, 92]. Due to the strong need for effective information extraction methods in the clinical domain, shared datasets, such as i2b2, have been developed, leading to the development of new, more efficient methods for extracting medical information.

Over the last years, NLP researchers have focused on building features from clinical notes relevant to named entity recognition prediction tasks. The first proposed approach was to develop a rule-based algorithm that combines syntactic properties of natural language and domain-specific rules [93, 94, 95, 96]. These methods enabled linguists and clinicians to work together to define each specific rule, leading to a significant modeling time and requiring a deep understanding of NLP and

medical domains. Thanks to the marked progression of language modeling methods in the generic NLP domain, medical teams saw a new alternative for automating the definition of features relevant to their extraction tasks while requiring less expertise and analyses.

The language modeling task consists of learning high-quality representations of words in a vector space from a large amount of unstructured text data. These representation vectors are then used as input features for predictive models of various NLP tasks such as Named Entity Recognition (NER) [97, 98] or text classification [15, 51, 99] . Mikolov et al. [87] introduced Word2Vec, a single-layer deep neural network that uses a novel training objective, "Skip-Gram", to conduct unsupervised learning of words representations. Specifically, The Skip-Gram uses the surrounding local context of individual words to update their representation vectors. Following the same line of thoughts, Joulin et al. [100] developed FastText, which is based on the Word2Vec model but is additionally learning the representation of the combination of adjacent characters that form the words, called 'n-grams'. These n-grams allow the model to consider all the language variability and build a robust representation of rare and unseen words. Given the rich and highly variable information contained within clinical notes, medical research teams pre-trained medical words embeddings and used them as input features for various medical predictive models [51, 15, 101], leading to significant improvement of predictive performance. For a more detailed overview of existing approaches and their benefits in the biomedical domain, Chiu et al. [102] conducted a comprehensive study that assesses the quality of biomedical embeddings with respect to a set of parameters and various tasks.

In clinical notes, the patient's condition is often described using abbreviations, fragmented phrases, and domain-specific jargon. Accordingly, words and abbreviations in clinical records can differ in meaning depending on their local context and the specific type of the note. For example, the term "Cold" could refer to three different things: a temperature, an unfriendly character, or a symptom. The fixed representation vector fails to comprehend all these meanings in one single vector as it does not consider the word's position with respect to its surrounding context. To address this problem, Peter et al. [103] pre-trained a language model (ELMo) that uses a set of Bidirectional LSTM layers. They generated the contextualized word representation as a concatenation of the

input sequence's left-to-right and right-to-left hidden states. In the medical domain, Yuqi Si et al. [92] demonstrated the importance of the contextualized embedding vector provided by Elmo in four medical concept extraction tasks. The F1-score of the four tasks was improved, on average, by 4%.

Based on the results achieved by ELMo's bidirectional architecture, Devlin et al. [39] introduced the BERT model, which replaces the sequential LSTM neural network with a stack of Transformer encoders [40]. In particular, they emphasized the limitation of sequential language modeling, which cannot combine both left and right contexts simultaneously. In addition, they developed a novel pre-training objective known as the Masked Language Model (MLM). Using this approach, a subset of words is randomly selected with a probability of 15% and are replaced with the special symbol [MASK]. Next, the Transformer stack generates a prediction for the masked words based on both left and right surrounding context. In brief, BERT is a two-stage training approach that combines, in a single model, the unsupervised pre-training of contextual embeddings using MLM and the fine-tuning of model's parameters with respect to a specific supervised downstream task. In particular, BERT outperformed ELMo scores on eleven major generic NLP tasks, including named entity recognition.

As all the relevant information is contained within the pre-trained embeddings, several works in the medical domain [38, 37, 104] have demonstrated the need for pre-training BERT on domain-specific datasets to achieve the state of the art scores on medical NLP tasks. Lee et al. [38] were the first to introduce a biomedical-specific language representation, they called BioBERT. It is a BERT-based model pre-trained on a large-scale biomedical corpus (4.5 billion PubMed abstracts and 13.5 billion PMC Full-text articles). They enhanced the performance of three biomedical NLP tasks: biomedical named entity recognition (0.62% F1-score improvement), biomedical relation extraction (2.80% F1-score improvement), and biomedical question answering (12.24% MRR improvement). BioBERT demonstrated the need for pre-training BERT-base models on domain-specific data through these different comparison studies. As routine clinical data is different from published biomedical texts, Alsentzer et al. [37] extended the work done by Lee et al. by additionally pre-training BERT-base and BioBERT models on over two million clinical notes from the

MIMIC-III database [13]. They demonstrated the utility of using domain-specific BERT model for a subset of clinical NLP tasks: i2b2 named entity recognition challenge [105] (3.65% F1-score improvement) and MedNLI natural language inference [106] (5.1% F1-score improvement). Beltagy et al. [104] pre-trained a new domain-specific representation model SciBERT using 1.4 million cross-domain scientific texts: 18% computer science papers and 82% biomedical papers. The major contribution of SciBERT is the usage of "SciVocab", a new WordPiece vocabulary [107] built from their scientific corpus. The study of the effect of SciVocab demonstrated the importance of using a domain-specific vocabulary to pre-train contextual representation models.

These works covered building biomedical and clinical-specific BERT resources and applying these resources to medical information extraction models. However, it is worth noting that the pre-training time of such domain-specific BERT range from 7 days to 23 days, depending on the GPUs resources and datasets size. Furthermore, access to a large set of medical notes for pre-training and an annotated corpus for medical concept detection are the key to defining such performing models. These two requirements limit the application of similar models in real-world private medical datasets where the annotations are often missing, and their acquisition cost is nonnegligible. In our work, we considered the pre-trained ClinicalBert as the core architecture for the medical concept extraction task and focused on solving the annotation cost by defining a novel active learning sampling strategy for Transformer-based architectures.

3.3 Review of Active Learning

The key idea behind active learning is to select the training observations that would accelerate the convergence of the supervised algorithm. In other terms, it controls how a learning algorithm could perform better with less training data. Active learning is well-motivated in several supervised learning problems where the acquisition of labeled data is costly. In particular, in the medical domain where the annotation of clinical notes requires additional expertise. The AL algorithm relies on the scenario and the query strategy. The scenario builds the pool of unlabeled data from which the learner is allowed to ask queries. On the other hand, the query strategy defines the

metric that ranks the observations and selects the samples to visualize to the annotator oracle.

3.3.1 Scenarios

- **Membership query synthesis**: The learner generates synthetic examples based on its knowledge and requests the oracle for labels [108]. However, labeling such artificial examples can be misleading for medical annotators as this data is different from actual human-defined examples.

- **Stream-based sampling**: Also known as selective sampling [109]. In this setting, the learner iterates over the unlabeled dataset. For each example, it decides whether it should be queried for human annotator validation or not. On the other hand, training complex deep learning models often requires many samples. Therefore, drawing an instance one at a time from the data source to prepare the samples of the next iteration is unrealistic and induces a large training latency.

- **Pool-based sampling**: Instead of applying the selection strategy to each sample in sequential order, the pool-based strategy [110] ranks the entire unlabeled pool before selecting the best queries for annotation. To ensure training efficiency, this scenario is the most common in the medical concept extraction application [111, 112, 48, 113].

3.3.2 Query strategy techniques

The query step consists of selecting samples based on a pre-defined strategy for a given deep learning model with parameters θ, applied to an unlabeled pool of data U. The labels are then collected from the oracle to form a new annotated training set L. Lastly, the set L is used to update the model's parameters θ and the pool U simultaneously. This section summarizes the four most common strategies defined for deep active learning modeling. For a thorough overview of all proposed approaches to select the most informative samples, we refer the reader to literature surveys on the topic [114, 115]

- **Random Sampling**: This strategy is used as a baseline in almost all scientific researches. It randomly selects the sample to annotate from the pool U. It is also called passive learning, as there is no active way of selecting samples at each learning iteration.

- **Uncertainty Sampling**: It is the most intuitive and common strategy to rank the informativeness of unlabeled observation [110]. In this setting, the learner queries samples with the highest uncertainty score. In our work, we used entropy-based uncertainty metric [116]. For a given probabilistic deep learning classifier with C classes and parameters θ, uncertainty is measured through the entropy formula:

$$x^* = \arg\max_x - \sum_{i=1}^{C} P_\theta(y_i|x) log(P_\theta(y_i|x)),$$

  where $P_\theta(y_i|x)$ is the probability that the sample x belongs to the class y_i. One limitation of the uncertainty sampling strategy is the sensitivity to select outliers as query samples. Outliers are likely to have high uncertainty scores but, at the same time, are not "representative" of other instances in the distribution. Including them in the training set will not improve the model accuracy on the data as a whole. Furthermore, it could easily lead to insufficient diversity of batch query samples (such that the data representation space is not fully queried) and consequently would not help DL model generalization through training iterations.

- **Diversity-based Method**: In order to prevent the outlier selection problems, Zhdanov et al. [117] proposed representiveness sampling strategy that aims to build a batch sample with diverse examples from all over the feature space. The sampling objective aims to find the optimal set $S \subseteq U$ and such as:

$$\max_{S \subseteq U} \sum_{x_i \in U} \min_{x_j \in S} d(x_i, x_j),$$

  Where $d(.,.)$ is a distance metric measuring the similarity between two samples. Used in isolation, diversity sampling methods will often privilege the samples farther away from the

decision boundary, therefore not selecting items that are likely to be mislabeled.

- **Density-Weighted Method**: This strategy combines uncertainty and diversity-based approaches to ensure the exploration-exploitation trade-off. The main idea behind this strategy is that informative instances should be not only those which are uncertain (good at exploitation) but also those which are "representative" of the true data distribution (exploration of the feature space). The strategy metric is defined as follows:

$$x^* = \arg\max_x \phi_u(x) \cdot (\frac{1}{U} \sum_{u=1}^{U} sim(x, x_u))$$

where $\phi_u(x)$ represents the informativeness of x according to the uncertainty sampling approach. The second term calibrates the informativeness of x by its average similarity to all other instances in the pool set U.

Part II

Methodology

Chapter 4

Time-aware deep learning Framework for IMTS classification

4.1 Introduction

Our objective was to design an end-to-end approach that addresses the downstream clinical tasks related to modeling time series with irregular observations. This approach should address the main challenges of multivariate medical time series: irregular time modeling, learning the dependencies between history spans over a long period, representing different evolution of medical states (worsens, improves, or chronic), and considering variable types of medical observations with multi-event time points. We designed a 3-step generic approach for end-to-end medical time series classification that integrates state-of-the-art architectures depicted in Chapter 2. The developed framework (Section 4.1) aims to facilitate the definition of empirical predictive analysis for medical researchers. We also designed a comparative study pipeline (Section 4.4) that enables research teams to test different variants of models and select the best architectures based on their dataset and prediction task.

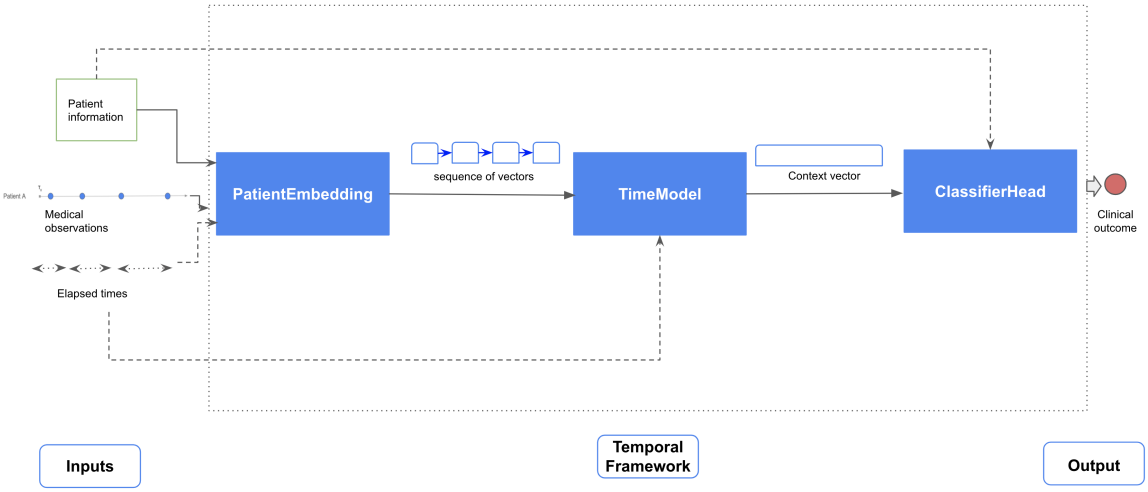


Figure 4.1: Three-steps modeling of irregular medical time series modelling

4.2 Framework modules

The architecture of our Framework is illustrated in Figure 4.1. The `PatientEmbedding` module aims to embed the raw information recorded at each time-stamp t and the related time value (such as the duration since last visit or the absolute time in days). The output, a sequence of dense vectors, is passed through the `TimeModel` component to encode the context and the temporal dynamic for each visit. The module also combines these hidden states and embeds the patient's health trajectory. Finally, the `ClassifierHead` module, a feed-forward network, is defined to make the final prediction. We note that the parameters of these three modules are simultaneously adjusted during the end-to-end supervised training.

4.2.1 `PatientEmbedding` module

The module can process temporal indicators, either numerical or categorical, as well as static patient information. It also supports different aggregation of these representations into one single sequence of health state embeddings. The possible inputs to the module are encoded categorical variables, normalized numerical values and the time of the observation. The embedding $e_i \in \mathbf{R}^q$

of the raw information x_i recorded at visit i is computed by aggregating four components: (1) *Categorical embedding* e_i^z, (2) *Continuous embedding* e_i^u, (3) *Time representation* e_i^t, and (4) *Static features embedding* e^d. The aggregation operator, Agg, can be either "concatenation", "mean" or 'attention-based average". The resulting embedding is expressed as: $e_i = Agg(e_i^z, e_i^u, e_i^t, e^d)$. In the context of deep learning, the embeddings are defined as low-dimensional, learned continuous vector representations of discrete variables. However, studies [69] extended this definition to continuous numerical variables. This work defines three methods for representing numerical features (time and numerical events). The first method is `Identity` that conserves the original value. The second is `Linear` where a shallow one layer multi-perceptron neural network (MLP) [118] converts the single value to a fixed-length vector. The last technique is `Soft One-hot encoding` [69] that mimics the embedding lookup table of categorical variables, except here, each continuous value is a weighted sum of the entire embedding table instead of a one-hot index lookup. For the time feature, we included two additional encoding methods: `TimeEncode` that learns a temporal positional encoding using a functional kernel [70], and `TimeMask` [69] that converts the time value to a vector of weights between 0 and 1 to "contextualize" the medical events.

4.2.2 `TimeModel` module

The sequence of embedded visits $(e_i)_{1\leq i\leq}$ are then passed through `TimeModel` that learns the hidden representation $(h_i)_{1\leq i\leq N}$ of all visits and aggregates them into one medical context vector $R_p \in \mathbf{R}^d$ of the patient p:

$$R_p = Agg([h_i]_{1\leq i\leq N}),\ where:\ [h_i]_{1\leq i\leq N} = Encode((e_i)_{1\leq i\leq N})$$

for the sequence encoder module "*Encode*", we integrated five neural-based architectures described in Chapter 2: GRU [57], LSTM [58], BiLSTM [119], Transformer [40], C-LSTM [53] and Time-aware transformer [70]. It is worth noting that both architectures, Time-aware transformer and C-LSTM, requires the time value of each visit as an additional input. The aggregation module "*Agg*" is identical to the one used in `PatientEmbedding` component.

4.2.3 `ClassifierHead` module

The module consists of feed-forward neural network [120], combined with a softmax output layer, to conduct classification tasks. It takes as input the contextual representation R_p and computes the probability scores over all pre-defined target classes. The output can be expressed as: $\hat{y} = sigmoid(w_o^T R_p + b_o)$. Note that the static features embedding e^d can be processed separately from the time series then concatenated to the the contextual representation R_p right before computing the classification scores as follow: $\hat{y} = sigmoid(w_o^T [R_p, e^d] + b_o)$.

4.3 Technical features of the Framework

The Framework's design aims to facilitate the definition of deep learning architectures for modeling the irregular sampled multivariate medical time series and allow medical research teams to quickly test these complex architectures on their private data and specific use cases. To achieve that goal, the Framework implements three standards features: (1) **YAML configuration files** that facilitate the definition of the model architecture and the setting of the argument of training and hyper-parameters optimization experiments. (2) **Python scripts** that automatize the set-up and execution of experiments (3) **Logging and file reports** that saves experiment's outputs for comparing the performances and visualizing predictions

4.3.1 YAML configuration files

For hyper-tuning experiments and training of resulting best models, we designed two structures of YAML configuration files. For both configurations, five major sections need to be set up: data paths, training and optimization arguments, patient embeddings modules, temporal model optimization parameters, and classification head parameters. Appendix outlines an example of this configuration file.

4.3.2 Python scripts

We also developed three main scripts to automate the execution of experiments using command-lines.

- `run_hyperparam.py`: The script processes the YAML file containing the search spaces of the hyper-parameters and launches Bayesian optimization trials using the Optuna package [121]. It is also linked to the "Weight And Biases" platform [122] to track the evolution of the performance metrics.
- `train_model.py`: The script processes the YAML file detailing the parameters and instantiates the related model architecture. Then, it runs the training experiments, logs the training progression, and finally saves the results of test data on disk. The output directory also includes the model's checkpoints and training history. Note that the test data can be related to a fixed set of patients provided by the user or automatically generated if the k-fold cross-validation option is enabled.
- `do_test.py`: The script takes the path to the trained model's checkpoints and the set of test patients, then computes the prediction scores along with metadata such as the attention weights.

4.3.3 Logging's reports

To conduct empirical studies and analyze their results, we have defined three types of logging: (1) A python logger storing the history track of training and testing phases. (2) CSV files storing the predictions for each cross-validation test set. (3) A pickle file storing the attentions weights returned by the model for each patient in the test data.

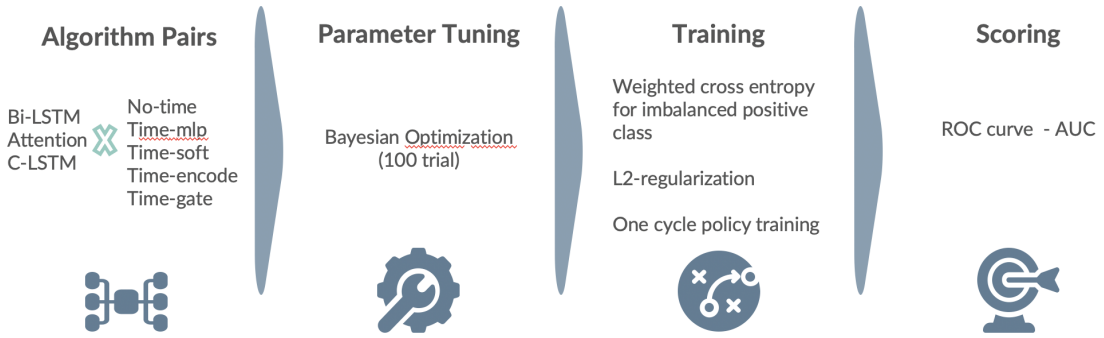


Figure 4.2: Pipeline of the comparison study proposed by the Temporal Framework. It provides the necessary modules and tools to efficiently train and compare different sequential models applied to medical data

4.4 Comparison study

4.4.1 Objective

The purpose of implementing a generic comparison process is to unify the experimental results of different models and ensure an unbiased comparison. The Figure 4.2 outlines the four steps of proposed the process. The main question addressed by the pipeline is: What is the best DL model for classifying IMTS given specific dataset and downstream task?

4.4.2 Algorithms: sequential models and time representations

The first step of the pipeline aims to include all state of the art models proposed for modeling irregular medical time series. We implemented two deep learning models for sequential inputs, Transformer (denoted in our experiments by "attention") and BiLSTM, alone, as well as combined with different time representation modules: TimeMask, TimeEncode, MLP projection, and Soft One-hot embeddings. Note that we used a bi-directional variant of LSTM to allow for using information from both contexts, left-to-right and right-to-left, and to ensure a fair comparison with Transformer architecture. We also tested a time-based model, C-LSTM, that directly incorporates

the timing irregularity in its parameters. For C-LSTM, we defined three ways of representing time gaps. The first option is *forget* and is similar to the original implementation. It considers time-gaps a parametric forgetting matrix that captures the evolution of health conditions and chronic diseases. However, the C-LSTM was designed for risk prediction in intensive care units and relied on a short medical history. In our case, we wanted to enable learning to represent the temporal patterns over a long period of the history of chronic disease and consider the time irregularity as an additional risk factor for developing health complications. For that purpose, C-LSTM was extended to include time-irregularity in the output cell as well, and two options were tested: *output* and *forget_output*. The first one considers temporal irregularity as a parameter that controls the information flow over historical data points. While in the second option, time irregularity controls both the forgetting memory and the output information. In total, the comparison study comprises 12 different models' variants where each variant is composed of a sequential model and a time representation.

4.4.3 Parameters tuning experiments

Using the Optuna Python package, for each group of experiments with deep learning and a time representation, the user can perform a bayesian hyperparameter optimization with 100 trials to optimize a predefined performance score. The hyperparameter tuning process uses 10% fixed validation set from the input data. To track the evolution of trials, the results of hyper-tuning trials are automatically logged in Weight&Bias platform [122].

4.4.4 Training optimization

To prevent over-fitting, we enabled L2 weight decay regularization to all models. Besides, we defined weighted cross-entropy loss class for training with imbalanced classes distribution. We also used once-cycle learning policy from fast.ai library [123].

4.4.5 Performance scoring

We used AUROC, defined as the area under the ROC curve, to evaluate the quality of predictions. It quantifies the probability that the classifier will rank a randomly chosen positive example higher than a random negative one. The advantage of the AUROC over the F1-score is that it considers all the possible classification thresholds to measure the quality of the model's positive and negative predictive values. AUROC also provides a more accurate performance profile of models for imbalanced datasets [124]. We note that the F1-score, precision, recall and accuracy metrics are also implemented in the proposed framework.

Chapter 5

Hybrid deep active learning strategy

5.1 Preliminary work

Clinical reports are a rich source of information as they contain detailed descriptions about the patient's health state and all the administered treatments along with his hospital visit or stay. Deep learning methods [15, 101] were proposed to learn clinically meaningful patterns from these texts to guide clinical decisions, including delaying or preventing disease onset. Particularly, we conducted an empirical study [15] that compares words embedding techniques [87, 125], classification machine learning algorithms, and CNN-based deep learning architectures when applied to french clinical reports for predicting the health-acquired infection onset.

We studied a cohort of 1,531 patients who visited three French university hospitals (Lyon, Nice, and Rouen) between October 2009 and December 2010. The input data consisted of the concatenation of all de-identified free-text medical reports of different types (such as discharge summaries, imaging reports, surgery reports, and consultation reports.) generated during the patient management. The best performing model was the CNN-based one with an increase of

14.1% over the best performing machine learning model. This performance gap consolidated the effectiveness of deep learning methods compared to machine learning ones. Besides, we defined a metric for the CNN model's output interpretability. The metric scores showed that 80.2% of the most important 3-words phrases selected by the model to compute its predictions included clinical terms relevant to infectious signals. On the other hand, the analysis of the CNN errors showed that 50% of the false positive classifications were due to the absence of temporality management.

The results analysis of our empirical study underlined the importance of taking into account temporality when modeling the incidence or the evolution of clinical outcomes. Furthermore, it showed that only a subset of medical concepts is important to downstream clinical tasks. These findings motivated us to explore information extraction models capable of filtering the relevant information from the clinical note.

Although active learning methods have been shown to be effective in many tasks, including text classification, information extraction, and speech recognition [115], there are limited explorations of AL techniques for clinical and biomedical NLP tasks. Besides, All the strategies proposed for optimizing the medical concept annotation process are applied to conventional machine learning models. On the other hand, Transformer-based architectures have shown encouraging results in medical information extraction problems [38, 37]. This finding motivated us to define a Hybrid Weighted Uncertainty Sampling (HWUS) that takes advantage of the contextual embeddings learned by the Transformer-based approach to measuring the representativeness of samples. The chapter formulates the problem, defines the proposed metric, and exposes the experiment setup and results.

5.2 Objective

In this chapter, we define our proposed active learning strategy that takes advantage of the contextual representations given by the Transformer-based NER models to define the sample representiveness strategy. Additionally, we define a decayed control parameter β to enable a dynamic calibration between uncertainty and sample representiveness based on the AL-based training stage. The main question addressed by the proposed strategy is: How to optimize annotation cost of

Transformer-based NER models ?

5.3 Problem Formulation

We define by $\mathcal{X}$ the set of all examples $x \in \mathcal{X}$, by $\mathcal{L} \subseteq \mathcal{X}$ the set of labeled data and by $\mathcal{U} \subseteq \mathcal{X}$ the set of N remaining unlabeled examples. Every step of active learning consists of selecting relevant $B \leq N$ examples from $\mathcal{U}$, collecting their related labels from the human annotator, and further training of the learning model. The process is iterative until a stopping criterion $\mathcal{C}$ is reached. We denote by $\mathcal{S}$ the set of selected examples at each iteration. We refer to the uncertainty score of a sample x by $\phi_u(x)$, and $sim(.,.)$ a distance-based metric that computes the similarity between two vectors. The objective is to define a hybrid sampling strategy for optimizing the training of a Transformer-based classifier.

5.4 Core Transformer-based architecture

ClinicalBert [37] have reached SoTA results in medical concepts extraction tasks. Therefore, we used it as the base architecture for our active learning experiments. The architecture consisted of a stack of $L = 12$ Transformer blocks, a hidden dimension of $d = 768$, and $A = 12$ self attention heads. A single linear layer is added to the model for concepts classification.

First, we downloaded the publicly available pre-trained embeddings and used them to initialize our classification model. Then, we fine-tuned the model on the i2b2-2010 task [105]. We note that we used the same hyper-parameters reported in the original paper. A learning rate $lr \in \{2.10^{-5}, 3.10^{-5}, 5.10^{-5}\}$, a batch size $bs \in \{16, 32\}$, and epochs $e \in \{3, 4\}$. The maximum sequence length was set to 150.

5.5 Dynamic Hybrid Weighted Uncertainty Sampling Strategy

Settles et al. [126] proposed an effective active learning strategy for sequence labeling task, called information density (ID), where the informativeness of a sample x is weighted by its average similarity to all other samples, subject to a parameter β that controls the relative importance of the density term. Following the same line of thought and in order to ensure the trade-off between uncertainty and diversity of the set $\mathcal{S}$, we formulated the following selection criteria:

$$x^* = \arg\max_{x \in \mathcal{U}} \phi_u(x) \times (\frac{1}{N} \sum_{u=1}^{N} sim(h_x, h_{x_u}))^{\beta},$$

where the second term measures the similarity between the hidden representation of x (extracted from the Transformer-based model) to all other representations of the pool $\mathcal{U}$.

5.5.1 Sample representiveness

The input sample x, a sequence of K tokens, is fed through the ClinicalBert classifier and we extracted the contextual hidden representations returned by the Transformer block $(h_i)_{1 \leq i \leq K}$. Then, we set the representiveness h_x to the hidden representation h_j of the token to be classified x_j. Then, we used the cosine distance to compute the similarity between representations: $sim(h_x, h_{x_u}) = \frac{1}{d} \sum_{i=1}^{d} h_x^i . h_{x_u}^i$.

5.5.2 Uncertainty metric

The uncertainty is measured by the entropy of the classifier softmax layer:

$$x^* = \arg\max_{x} - \sum_{i=1}^{C} P_\theta(y_i|x) log(P_\theta(y_i|x)),$$

5.5.3 Decayed control parameter

To give more weights to the pre-trained embeddings of the Transformer model, we introduced a dynamic parameter β based on a step-wise decayed rate that ensures decreasing relative importance of the density term through AL iterations. The dynamic parameter is formulated as $\beta_t = \frac{\beta_0}{1+kt}$, where $\beta_0 \in \{1, 2, 3\}$ is the value at the initial iteration, k is a hyper-parameter and t is the iteration number. We ran experiments with three values of $k \in \{0.25, 0.5, 0.75\}$. Setting a dynamic rate is motivated by the fact that the initial iterations of the trained model do not ensure high-quality prediction scores, and thus we privileged the density-based term in the early training stages to account for the medical knowledge contained within the pre-trained embeddings. Additionally, we examined a exponentially decayed rate, but our experiment results were inconclusive.

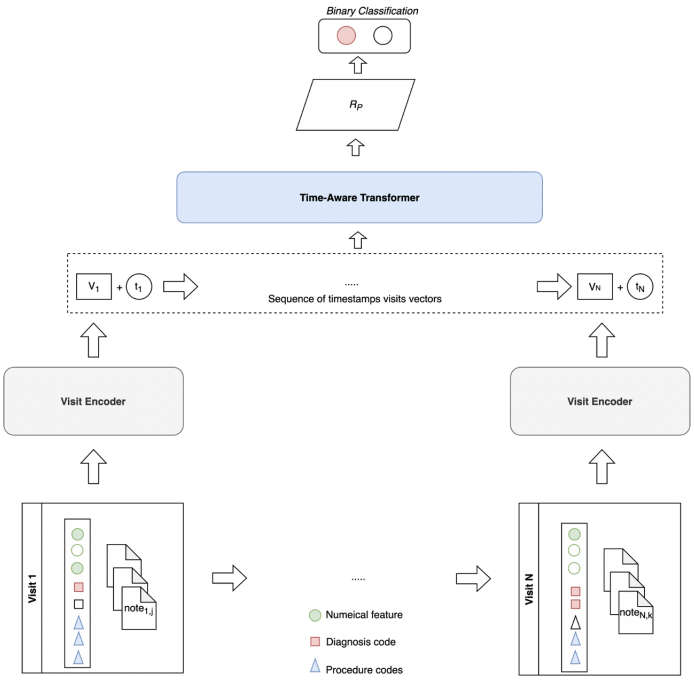
Chapter 6

Multi-modal Hierarchical Transformer-based approach for IMTS classification (Multi-HiTT)

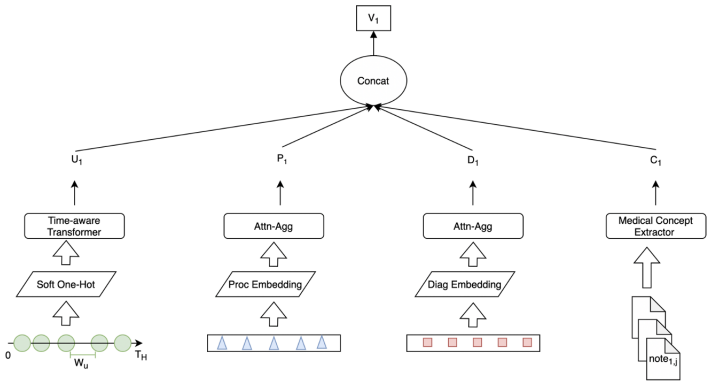
6.1 Motivation and objective

Transformer-based architecture combined with time representation showed promising results for modeling irregular medical time series with one type of input data. In this work, we aim to design a hierarchical transformer-based architecture for modeling highly variable multivariate time series and evaluate the impact of representing the time irregularity between visits on the performance of clinical prediction tasks. Additionally, we evaluate the effectiveness of combining temporal features, patient static information, and the extracted medical concepts from clinical notes type on the performance scores. The main question addressed by the proposed architecture is: How to combine different types of medical records to improve downstream tasks performance?

6.2 Model architecture



(a) General architecture of the HiTT model



(b) Architecture of the `VisitEncoder` sub-module

Figure 6.1: Diagram of the Multi-HiTT architecture

The architecture of Multi-HiTT is illustrated in Figure 6.1a. The first level of our Multi-HiTT network is to model the latent correlations among recent observations and importantly improve the capability of patient's recent health condition in each hidden state. This modeling process is represented in the Figure 6.1b . Take the current t-th episode for instance, the objective of the module is to express the episode as a set of four independent embedding tensors: $V_t = [U_t, X_{t,p}, X_{t,d}, Nt]$. To that end, `VisitEncoder` can be decomposed in three components: numerical time series module, categorical embeddings module and clinical concept extractor module.

The second level of Multi-HiTT is to combine the long-term dependency and patient's short-term condition to obtain current patient's overall health state. The modeling process is illustrated in the Figure 6.1a. The construction of contextual attention-based hidden state h_t is ensured by the Time-aware Transformer encoder. It takes as inputs the sequence of visits embeddings $(V_t)_{1 \leq t \leq N}$ and the related time gaps Δ_t. Then, it computes temporal position encoding with `T`, sums it up to V_t and propagate the information through a stack of self-attention and point-wise feed-forward layers.

The final patient representation R_p is set to the last hidden state h_N. This vector, concatenated with patient demographics embedding, is passed through a feed-forward network to make the final prediction.

6.2.1 `VisitEncoder` module

Numerical Time Series Representation Module

The first component is taking as input the window-based segments of the time series C_t^u, a context matrix consisting of recent N_{w_u} numerical values, where w_u is the temporal window width (in our pre-processing w_u is equal to 6 hours). Then, it projects the segments into a latent space using `Soft One-hot` module and returns a contextual hidden state U_t using a Transformer encoder.

Categorical Embeddings Module

The second component comprises two embeddings modules `DiagEmb` and `ProcEmb` for generating the Bags of Embeddings related to diagnoses ($X^i_{t,d}$) and procedures ($X^j_{t,d}$), respectively. Then, it aggregates them using an attention-based weighted average operator.

Text Representation Module

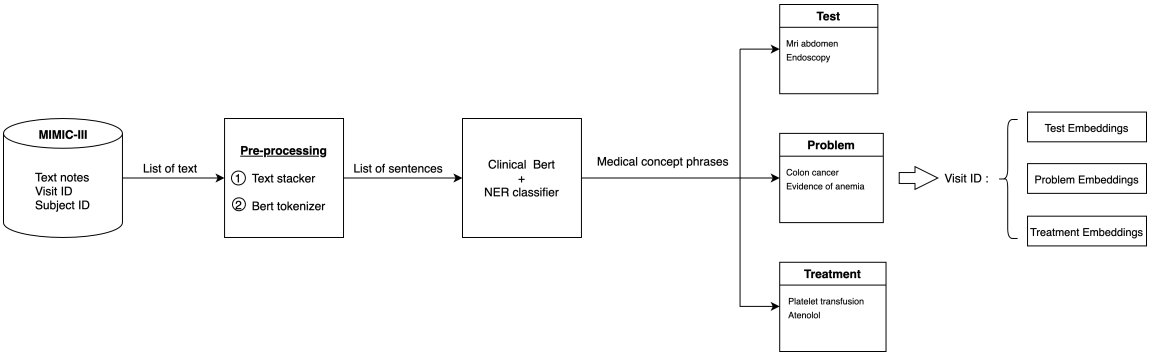


Figure 6.2: Workflow to prepare the pre-trained embeddings of medical concepts extracted from all clinical notes generated during a Visit ID

Finally, we defined a text representation module to process the clinical notes $(n^i_t)_{1\leq i\leq k}$ generated during the visit t. Each note will be fed to the fine-tuned ClinicalBert to extract the medical concepts phrases (1). Then, each phrase will be presented by the average of contextual words embeddings (2). The resulting phrases embeddings are then gathered based on their predicted entity class (Problem, Test, Treatment) (3). Lastly, the three vectors are concatenated with the processed procedures, problems, and continuous embeddings to form the encoded visit vector V_t (4).

6.3 Baselines models

Previous studies [72, 33] have shown that deep learning methods outperformed conventional machine learning algorithms for the in-hospital mortality task. Consequently, we compared our model with the following deep learning-based baselines:

- **Vanilla BiLSTM**: The embeddings of medical events within the same episode are averaged and passed through a BiLSTM to encode the sequence of visits. The patient representation is set to the last hidden state of the recurrent network and fed to the classifier module.
- **Vanilla Transfomer**: The model architecture is similar to the first one, but we replace the sequence encoder "BiLSTM" with a Transformer architecture.
- **Hierarchical Transformer - Time**: We deactivate the temporal module that represents the time gaps between consecutive visits. In other terms, we considered the sequence of episodes as an equal-spaced multivariate time series.
- **Hierarchical BiLSTM**: We replace the two Transformer encoders with "BiLSTM" ones and use the "Linear" embedding module to represent the time gaps between consecutive visits.

6.4 Variant of Multi-HiTT architectures

Besides, for isolating the performance gain of each representation module defined within the `VisitEncoder`, we defined four variants of the Multi-HiTT architecture:

- **HiTT - continuous**: Removing the continuous module that learns to represent multivariate time series.
- **HiTT - diagnoses**: Removing the `DiagEmb` module that learns to represent categorical embeddings of diagnoses codes.
- **HiTT - procedures**: Removing the `ProcEmb` module that learns to represent categorical embeddings of procedures codes.
- **HiTT + text**: Adding text representation module that represents the medical concepts described in the clinical notes.

Part III

Experiments and Results Analysis

Chapter 7

Medical sources

7.1 Introduction

This chapter presents two real-world clinical databases used to validate the proposed temporal framework (Section 7.2) and the Multi-HiTT architecture (Section 7.3). The Section 7.4 describes the i2b2-2010 challenge task we used for building the simulation study of active learning strategies.

7.2 Regional database of diabetic patients - CaRéDIAB

7.2.1 Database description

In the Champagne Ardenne area and since 2003, the database CaRéDIAB [52] (Champagne Ardenne Réseau Diabetes) stores clinical and paraclinical data of diabetic patients, regardless of their diabetes type. In particular, it includes more than 2000 adults and children with type 1 diabetes. The number of followed patients makes it a valuable database with rich structured follow-up data. It allows simultaneous access to the HbA1c as a marker of diabetes control, diabetes complications, and many other medical variables.

The data is collected during medical and paramedical consultations or hospital stays. Some

patients also benefit from retinopathy screening, using a mobile imaging unit, with a distance reading of the fundus photographs by an ophthalmologist. As with all real-life databases, there is some irregularity between visits, with some patients having periods without follow-up.

A Clinical Research Associate based at the Reims University Hospital ensures the update of the history of medical records. The database has obtained the agreement of the National Commission for Data Processing and Individual Liberties (CNIL; Commission Nationale Informatique et Libertés), number 1434306. All professionals certified by the network are subject to a confidentiality agreement, and patients signed the informed consent to include them in the database [52].

At each patient visit, the clinician registers the HbA1c rate, leading to a varying sequence of measurements depending on the frequency of patients' appointments and blood tests (B.2 visualizes the different time series profiles). The presence or absence of diabetic retinopathy is assessed from the screening performed within the CaRéDIAB network, on fundus photography, during hospitalization, or by their usual ophthalmologist. The criteria for retinopathy are defined by the international ETDRS criteria, as recommended by the French guidelines [127].

7.2.2 Target variable definition

Our database stores the information about retinopathy status in different columns depending on the report source (consultation, hospitalization, ophthalmologist). Therefore, a pre-processing step was necessary to aggregate these features and accurately fill in the target variable for the patients considered in the study. We conducted an iterative process with two steps: First, the medical experts defined the annotation rules based on eye-screening results to determine the retinopathy status. Then we generated the related algorithm and selected a random set of patients to validate the results. After three iterations, we fixed the rule-based algorithm (detailed in Appendix B.1). This problem is posed as a binary classification one, and true retinopathy labels were created by checking the retinopathy status at the following visit.

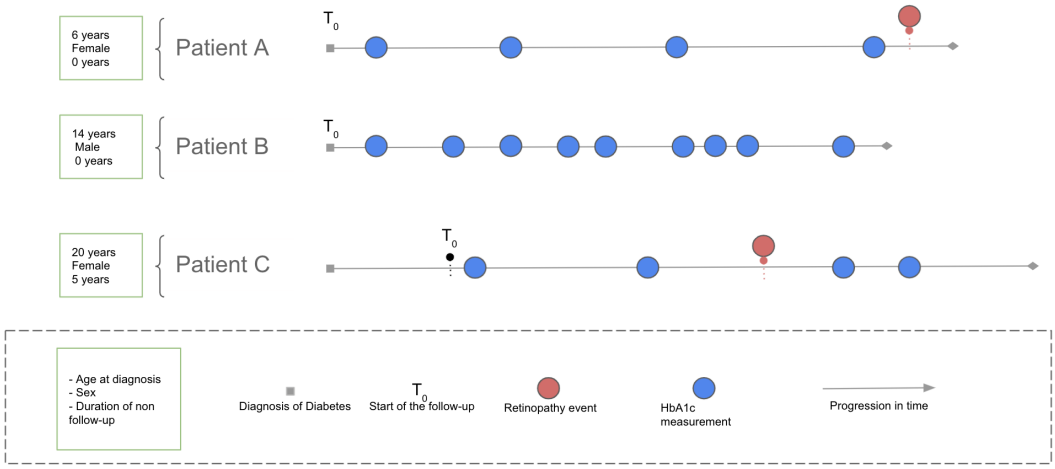


Figure 1. Variabilité de l'historique des soins des patients diabétiques de type 1

Figure 7.1: Schematic representation of the follow-up of different patients: example of three patients with type 1 diabetes and different follow-up durations and frequencies. We note that the interval between observations varies between patients but also within the follow-up of the same patient. Lastly, not all patients have diabetic retinopathy.

7.2.3 Data pre-processing

The analysis of the HbA1c records shows some data redundancy of the HbA1c levels taken within a time window of seven days. These redundancies were found for hospitalizations recorded before 2018 and are related to the system allowing different caregivers' to fill in the same data record. Since then, the information system has been updated to avoid duplicates. We grouped all HbA1c levels occurring within seven days to process duplicates by taking the most recent one. Then, we computed the sequence of time gaps between consecutive records. In addition to the sequential features, we defined the patient metadata as a list of three features: the non-follow-up duration, the sex of the patient, and age at onset of diabetes. Finally, all numerical values were normalized to meet the input format requirement for deep learning models.

7.2.4 Data statistics

The Table 7.1 describes the population of study (mean ± standard deviation). The total number of patients included in the study was 1207 patients. We noticed an imbalanced distribution between the positive class (presence of retinopathy) and the negative (absence of retinopathy). The age at the discovery of diabetes was the same for both classes. However, the duration of follow-up and non-follow-up were twice as long for patients that developed a retinopathy complication. Furthermore, those patients also had a more irregular history of records emphasized by longer time gaps between consecutive records and higher HbA1c rates.

Table 7.1: Preprocessed dataset statistics: Distribution between the patients with a retinopathy complication and those without.

Variable	**Without retinopathy**	**With retinopathy**
Patients (sex : M / F)	967 (M=476, F=491)	240 (M=126, F=114)
Number of HbA1c	21.87 ± 16.92	21.41 ± 17.36
Age of onset of diabetes (years)	17.72 ± 14.91	17.89 ± 13.27
Follow-up duration (months)	124.01 ± 87.92	247.65 ± 133.44
Duration without follow-up (months)	43.89 ± 94.47	94.55 ± 135.50
Median of time-gaps between records (days)	98.72 ± 71.90	123.85 ± 132.42
Median of HbA1c level (%)	7.89 ± 1.29	8.57 ± 1.54

7.3 Medical Information Mart for Intensive Care (MIMIC-III)

7.3.1 Database description

MIMIC-III [13] 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. As visualized in Figure 7.2, 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 stay, and survival data. The database covers the health information of 38,597 adult patients and 49,785 hospital

admission. 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.

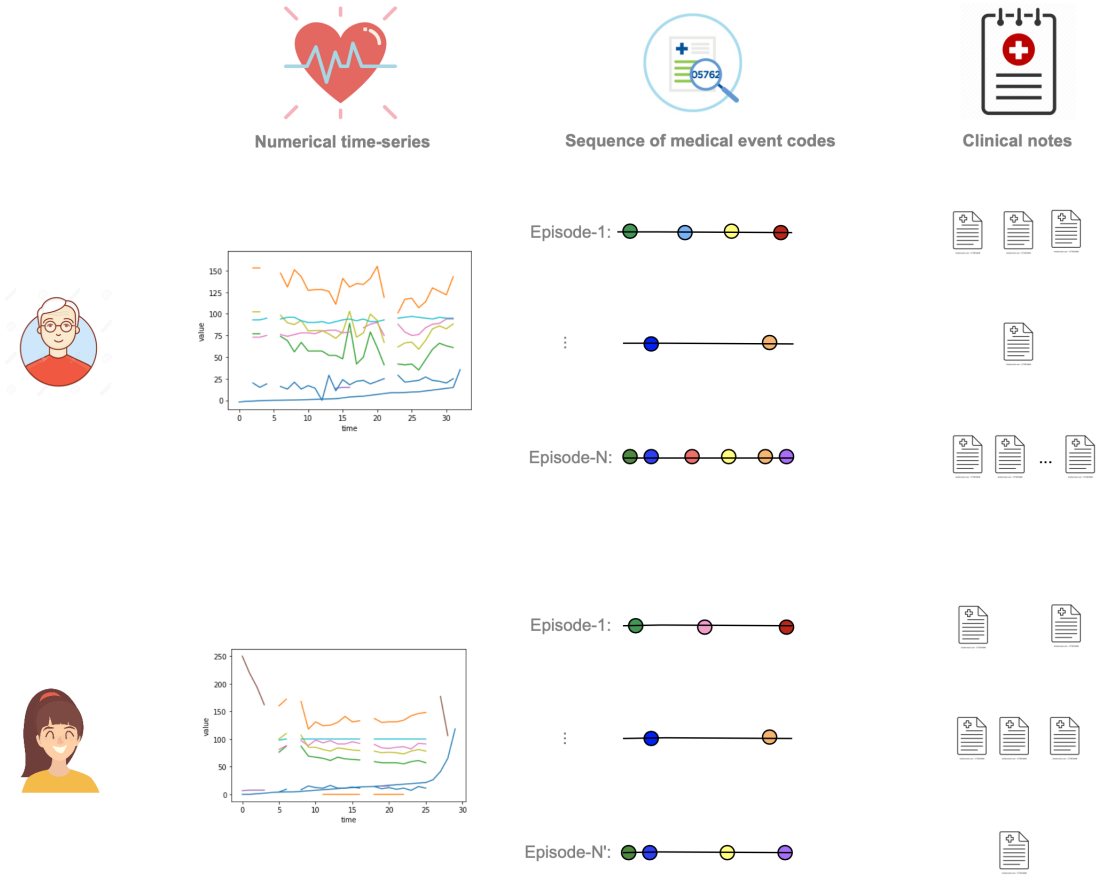


Figure 7.2: Schematic representation of two critical care patients with different duration of follow-up and records frequencies. Note that certain physiological variables are not examined at some visits, causing missing values

To prepare data for machine learning and deep learning studies, Huang et al. [128] derived from this database a public benchmark that includes four different clinical prediction tasks: in-hospital mortality, physiologic decompensation, length of stay (LOS), and phenotype classification. The resulting benchmark defined multivariate numerical time series and the related sequences of diagnoses from more than 42,000 intensive care unit (ICU) stays. It also created the associated labels for the four pre-defined tasks. Their work enabled the publication of novel modeling techniques that follow a rigorous evaluation process and a fair comparison with other existing approaches. We extended their pre-processing step to include the sequence of procedures and clinical notes generated at each stay for our experiment. We used the MIMIC-III mortality prediction task as an example of a classification task with multivariate time series. We also restricted our study to the seven most frequent numerical parameters: Diastolic blood pressure, Glucose, Heart Rate, Mean blood pressure, Oxygen saturation, Respiratory rate, and Temperature. Besides, we selected the patients with at least two visits and computed the time gaps between consecutive ones. For the IHM task, we ensure that the last visit lasts more than 24h. The dataset comprises 13718 hospital admissions, 177483 diagnoses with 4114 unique ICD-9 codes, and 51536 procedures with 1250 unique ICD-9 codes of 5120 patients.

7.3.2 Clinical objective

Mortality prediction is vital during rapid triage and risk/severity assessment. The IHM could be defined as the outcome of whether a patient dies during the hospital stay or lives to be discharged.

7.3.3 Target variable definition

This problem is posed as a binary classification one, and true mortality labels were created by comparing the date of death with hospital admission and discharge times.

7.3.4 Data pre-processing

This section explains the pre-processing steps for preparing the irregular sampled multivariate time series for the in-hospital mortality binary prediction task. We defined each visit as a sequence of three dynamic variables: sequence of diagnoses codes $x_{t,d} \in E_d$, sequence of procedures codes $x_{t,p} \in E_p$, and a matrix of seven time series $u_t \in \mathbf{R}^7$. We also define patient-level static vector d_p containing the following demographic data: ethnicity, gender, age. The categorical events were transformed using *Label Encoding* that maps the medical code to an integer value between 1 and the cardinality of events set: $e_d = |E_d|$ for diagnoses and $e_p = |E_p|$ for procedures. Finally, all numerical time series were normalized to meet the input format requirement for deep learning models.

7.3.5 Data statistics

The Table 7.2 describes the population of in-hospital mortality prediction study. We reported the averaged features in the format of "mean ± std" and specify the interquartile range of the median values. We noticed an imbalanced distribution between the positive class (occurrence of IHM) and negative (absence of IHM). The age at last admission is skewed for patients deceased during the hospital stay. However, the duration of follow-up is similar for both populations. Furthermore, the patients with in-hospital mortality stayed longer with more procedures and diagnoses. Those patients also had a more irregular history of records emphasized by longer time gaps between consecutive records.

Table 7.2: Preprocessed dataset statistics: Distribution between the patients with positive in-hospital mortality and those without

Variable	**No IHM** (74.9%)	**IHM** (15.1%)
Patients (sex : M / F)	4346 (M=2386, F=1960)	774 (M=442, F=332)
Median Number of Visits	2 (IQR=1)	2 (IQR=1)
Average age at last admission (years)	66.16 ± 16.73	89.25 ± 16.12
Average follow-up duration (days)	586.87 ± 768.23	598.44 ± 744.55
Average length of stay per visit (days)	3.63 ± 5.32	5.43 ± 7.90
Median number of diagnoses per visit	12 (IQR=9)	13 (IQR=9)
Median number of procedures per visit	3 (IQR=4)	5 (IQR=5)
Average of time-gaps between visits (days)	420.77 ± 577.79	499.08 ± 659.72

7.4 I2b2-2010 NER Challenge

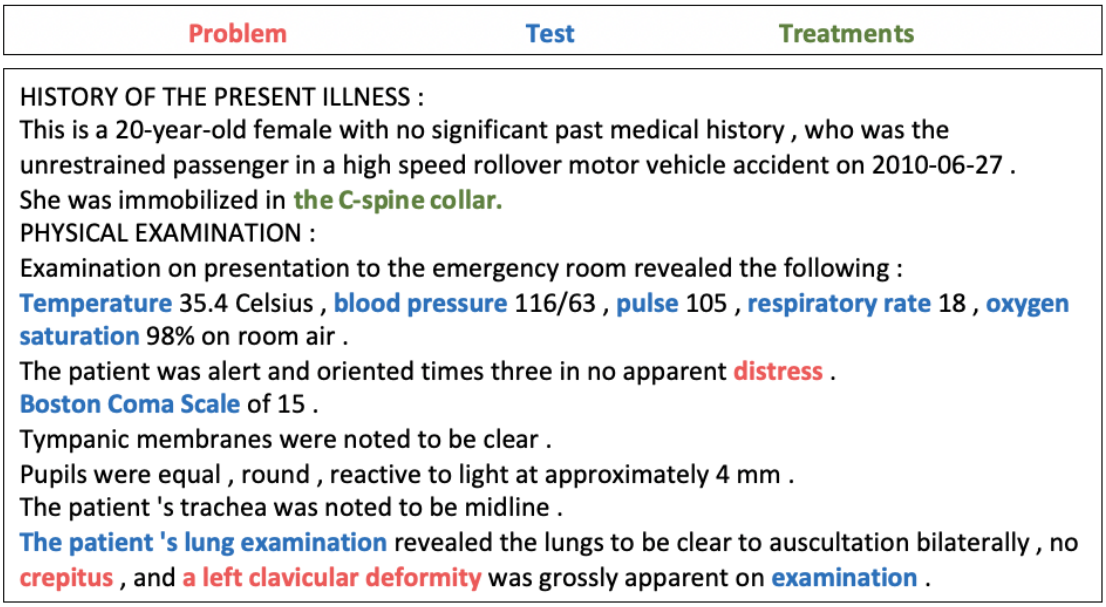


Figure 7.3: Annotated sentences with medical concepts defined in i2b2-2010 challenge

7.4.1 Challenge description

The 2010 i2b2/VA NLP challenge defined a medical concept extraction task and comprises 349 clinical documents with 20,423 unique sentences. Each sentence was annotated based on three types of medical entities: problem, treatment, and test (Fig. 7.3).

7.4.2 Clinical Objective

The vast majority of clinical data available in the EHRs takes the form of narratives written in natural language. While free text is practical to describe complex medical states, it is difficult to use for medical decision support systems, clinical research studies and statistical analysis. The goal of the i2b2-2010 Challenge is to develop NLP approaches to automatically extract key medical

concepts from clinical notes and reuse them in downstream tasks. They defined three categories of concepts: Problem, Test and Treatment.

7.4.3 Data Pre-processing

Pre-processing is necessary to prepare data for building supervised machine learning models since clinical narratives are frequently unstructured and fragmented. Tokenization and sentence segmentation are the two most important preprocessing steps. Sentence segmentation consists of detecting the sentence boundary and convert a document or a text-paragraph to a list of independent sentences. Tokenization is the process of breaking down each sentence into smaller chunks, usually words.

7.4.4 Data statistics

Table 7.3: Statistical description (counts of sentences, words, and entities, average words per sentence, average entities per sentence) of the test set and the pool of querying data using i2b2-2010 challenge data

	Pool data	**Test data**	**Total**
Sentence count	18681	3808	22489
Word count	265092	55195	320287
Entity count	39687	7999	47686
Average words per sentence	20.5632	20.8953	-
Average entities per sentence	2.1244	2.1005	-

Chapter 8

Retinopathy Prediction Use Case

8.1 Introduction

To validate our proposed framework, we conducted a comparative study to predict the retinopathy complication, using 1207 highly variable medical time series of HbA1c gathered from the French database CARéDIAB. This chapter outlines the experiment's details, the analysis of the obtained results showcasing the relevance of applying temporal deep learning methods for retinopathy prediction.

8.2 Background

Several research studies [49, 129, 130, 131, 132, 133, 131, 134, 135] have focused on analyzing EHR data to detect and diagnose diabetes complications, mainly in patients with type 2 diabetes. They applied machine learning models, mostly Random Forest, Support Vector Machine (SVM), Logistic Regression (LR), and Artificial Neural Network (ANN), to conduct uni-label classification. The task consisted of predicting the outcome separately or using multi-label classification [136] to consider the correlation between these outcomes. All studies used tabular features as input to machine learning models, including risk factors, patient demographics, and comorbidities status. They used

aggregation methods such as general computing statistics (min, mean, max) or generating rolling window statistics for finer temporal granularity to define temporal-based risk factors. The subset of works [135, 137, 138, 49, 134] that studied EHR data for detecting diabetic retinopathy (DR) have reached an AUROC score ranging from 0.726 to 0.802. Particularly, HbA1c rates were a strong predictor of DR in all predictive models that included them as risk factor [52, 139, 130, 140]. On the other hand, a recent review of deep learning methods applied to diabetes [141] showed that most related research has focused on the analysis of medical imaging to detect and diagnose multiple diabetes complications. The most used architecture is the convolution neural network (CNN) and its derivatives. While the only work taking advantage of longitudinal data [53] used a C-LSTM model to predict unplanned readmission and to recommend the next intervention. Finally, among all diabetes complication studies, only a few have modeled type 1 diabetes as outlined by the recent review [142] where authors have concluded that there is a gap in knowledge regarding the prediction of microvascular complications specifically for T1D patients.

8.3 Objective

Our work aims to combine sequential neural networks and time irregularity representations to model the series of HbA1c levels and their variability (short and long-term variability) observed for T1D patients with different disease duration for predicting diabetic retinopathy. As previous work [52] demonstrated the impact of distant past HbA1c features on the development of retinopathy, we did not include time-decay-based temporal architectures (GRU-D and T-GRU), and we focused on the parametric approach "C-LSTM".

8.4 Experiment Set-up

8.4.1 Partitioning and Cross-validation

To measure the trade-off between the number of patients and the length of the sequence of records, we defined four groups of patients based on the number of HbA1c measures. Each group X includes

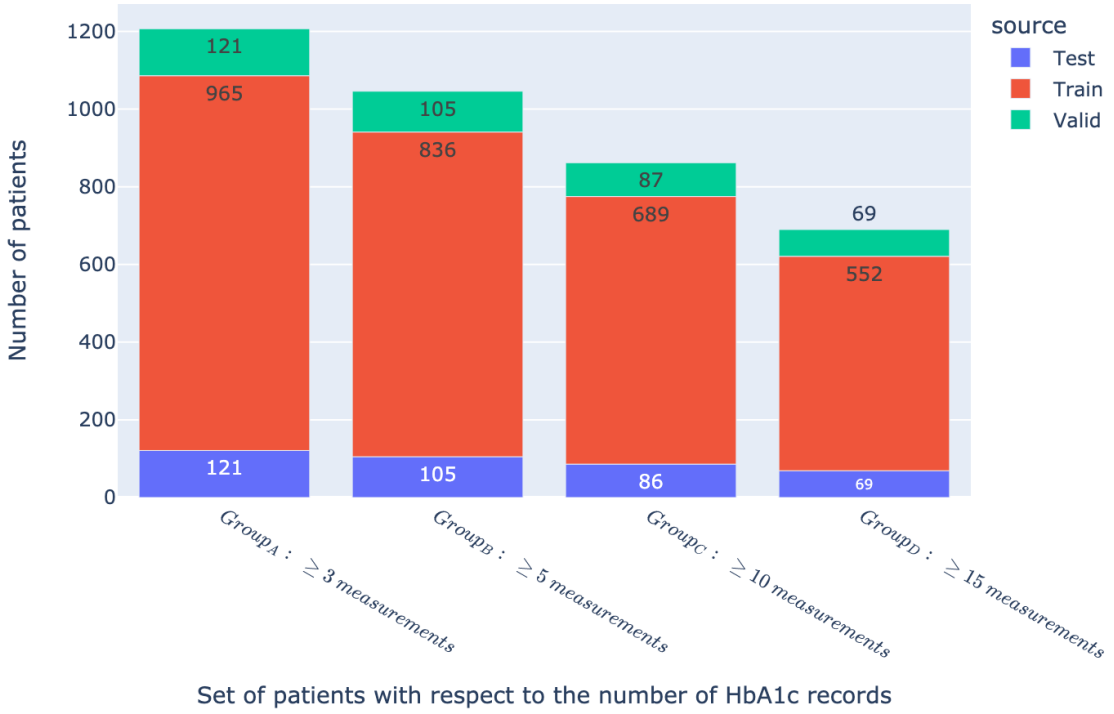


Figure 8.1: The distribution of the patients in cross-validation sets with respect to patients groups. The sets $Group_A$, $Group_B$, $Group_C$, and $Group_D$ include patients with at least 3, 5, 10, and 15 records respectively. The validation set is fixed and used to fine-tune the hyper-parameters. In contrast, the Test set is dynamically computed during the nested 5-fold cross-validation iterations applied to the remaining patients not included in the validation.

patients with at least X values recorded during their follow-up history. For example, one patient with more than 15 records will be present in the four groups. Finally, we applied the cross-validation protocol defined previously to each group.

8.4.2 Performance metrics

We used the performances metrics defined in (Section 4.4.5): The area under the ROC curve (AUROC) and the F1-score.

8.4.3 Setup + Hyper-parameter tuning

We optimized the hyper-parameters of each model using the Optuna Python package [121]; we report the search space and the best parameters in the Appendix B.3. The results of hyper-tuning trials were logged in Weight&Bias platform [122]. To prevent over-fitting, we applied L2 weight decay regularization to all models. Besides, we used a weighted cross-entropy loss for training to consider class imbalance. The weight decay and the weight of classes were also included in the hyper-parameters optimization step. We used a V-100 GPU with 32GB of memory to conduct all the experiments.

8.5 Results analysis

The hyper-parameters fine-tuning results in Fig. 8.2 show that the attention mechanism models are outperforming recurrent neural architectures for the group of patients with a minimum of 3 records. However, we observe that both approaches lead to similar performance for the other groups. The group of patients with a minimum of 3 records was the largest one. Therefore, we could hypothesize that attention-based models outperform RNN-based methods when a larger dataset is available. Furthermore, C-LSTM models are under-performing the other models for all groups, while the gap decreases for the set of patients with at least 15 records. The hyper-parameters

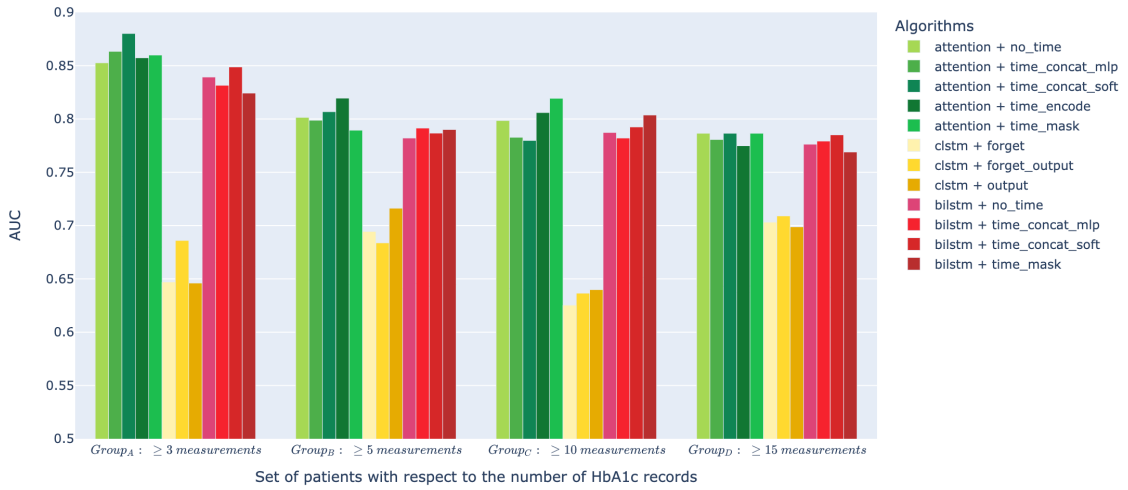


Figure 8.2: Hyper-parameters optimization results with respect to the minimum length of HbA1c history. Each algorithm corresponds to the pair (sequential model, time representation technique) that we denote by "sequential model type + time representation". The figure shows the trade-off between training with more patients and the time series length.

tuning experiment demonstrated that the best architectures reach the highest performance score for the most extensive data set. Therefore, we focused the rest of our analysis on $Group_A$ and ran the nested 5-fold cross-validation protocol using the best models returned by the hyper-parameter tuning (the table of results can be found at B.4). The objective of this second experiment was to test the generalizability and the stability of the models over new sets of patients. Similar to validation results, Fig.8.3 shows the same relative ranking of the performance of the algorithms. Furthermore, it shows that the attention-based model combined with the soft one-hot representation of the time reaches the highest score. However, it is more variable than the LSTM model with the time concatenation or the temporal-based attention-based model (attention + time encode). This variance could be explained by the training sample size (965 patients) as Transformer-based are shown to be more efficient with the increasing scale of data [143].

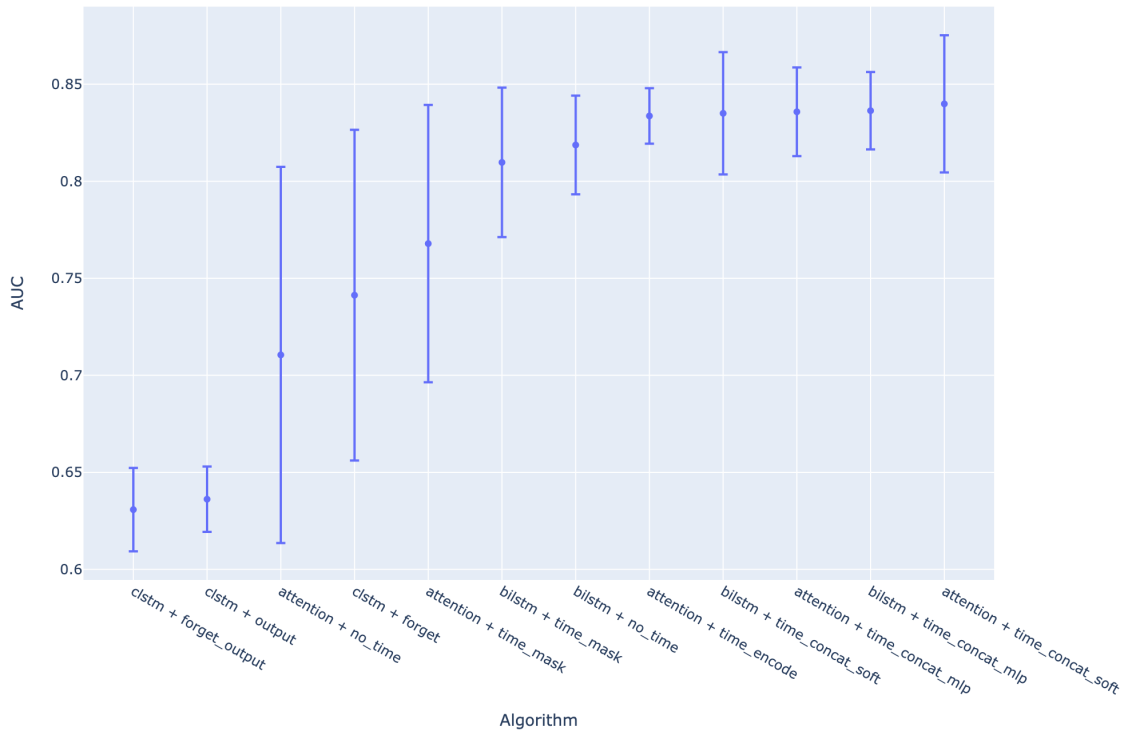
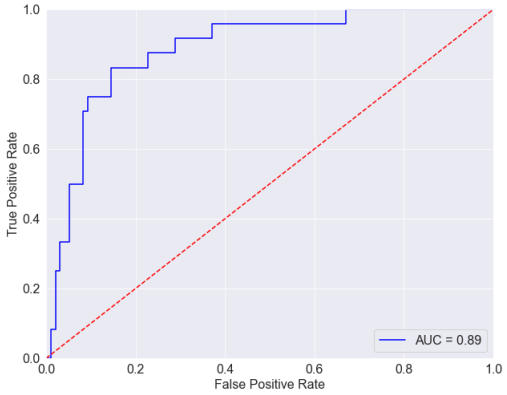
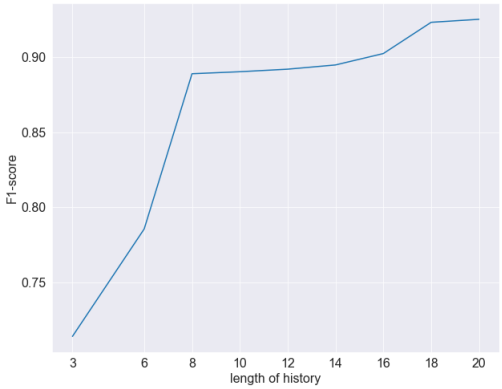


Figure 8.3: AUROC scores (mean ± std) for all the pairs (sequential model, time representation) when applied to the nested cross-validation sets of the $Group_A$ corresponding to patients with at least 3 past HbA1c observations.

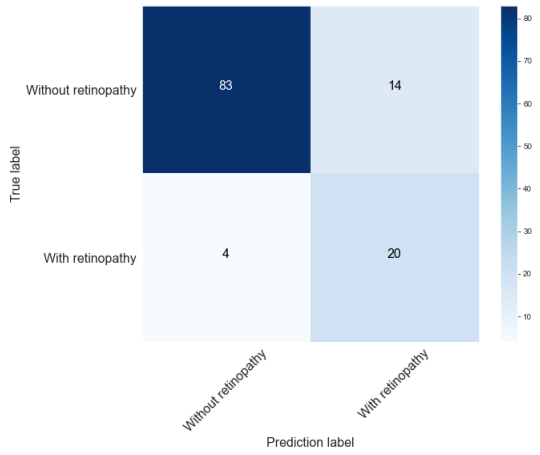
Selecting the model leading to the highest score "*attention + time_concat_soft*" (AUROC of 88.65%), we chose the cut-off threshold based on the ROC curve, such as maximizing the True Positive Rate while ensuring a low level of false alarms. The resulting F1-score is 0.8512 (a specificity of 85.56%, a sensitivity of 83.33%), and the confusion matrix shows (Fig E.3) a strong ability of the model to retrieve patients with the risk of developing retinopathy. Furthermore, we visualized the model's performance based on the sequence length of HbA1c levels. We observed that the F1 score increases when the sequence gets longer. This result suggests that the model is more confident about its predictions for patients with a longer series of HbA1c records.



(a) The ROC curve and the area under curve score (AUROC)



(b) Evolution of F-1 score with respect to patient's history length



(c) The confusion matrix of best attention model

Figure 8.4: Performances of the model *attention + time_concat_soft*: Fig (8.4a) visualizes the ROC curve showcasing the trade-off between sensitivity and specificity for each possible cut-off. Fig (8.4b) displays the evolution of F1-score with respect to the number of HbA1c records in the patients' history. Fig (8.4c) zooms in the predictions of the model and exhibits the correct predictions of each class (the diagonal), the false positives (top-right), and the false negatives (bottom-left)

We also tested the impact of the patient-level information, age at diabetes onset, sex, and the non-follow-up-duration, by removing them from the input time series. In Fig. 8.5, we report the

results of the best performing model of each group, with and without patient-level information. We noticed that for groups A, B, and C, the relative gain of performance is 21.1%, 16.5%, and 20%, respectively. While the patient information helps get a higher score, the gap is less significant (6.5%) for the group D with more than 15 records.

Finally, we measured the execution time of the best performing BiLSTM and Self-Attention models. B.5 shows the execution profile of the training loop. We can observe that the attention-based model was 1.4 times faster. This improvement is mainly due to optimized callback. The time difference is related to the backward optimization step of BiLSTM that updates the historical states iteratively due to the recurrence mechanism. In contrast, the attention-based model's states are updated in parallel.

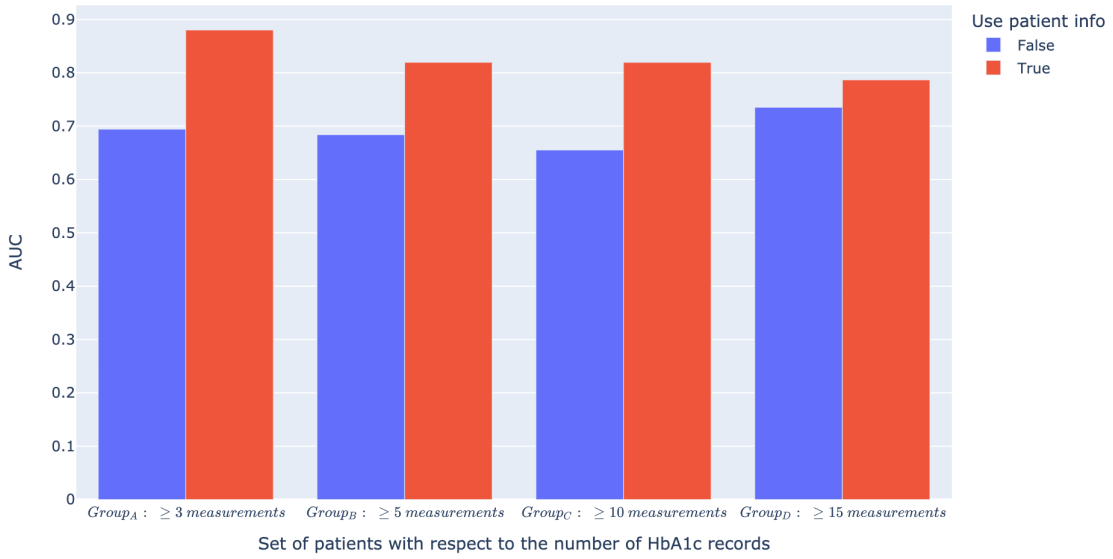


Figure 8.5: The impact of adding side-information: non-follow-up duration, age at onset, and the sex of the patient on best performing algorithm for each group of patients.

Chapter 9

Simulated Active Learning Study

9.1 Introduction

To validate our proposed active learning strategy (Dynamic-HWUS), we conducted a simulated experiments using the annotated corpus of the i2b2-2010 [46] challenge and the pre-trained ClinicalBERT model [37]. This chapter outlines the background of applying active learning strategies to medical concepts extraction, the description of the simulation study, and the analysis of the obtained results.

9.2 Objective

We aim to compare different active learning strategies with passive learning in terms of downstream performance (F1-scores) and the size of annotated data needed to reach a fixed performance threshold.

9.3 Background

The requirement of large annotated clinical notes to train machine learning NER models and the high costs and constraints associated with obtaining annotations in this domain have motivated research works to explore the application of active learning strategies. Romer et al. [144] combined active learning strategy and semi-supervised learning [145] to reduce the annotation cost required to train a Naive Bayes classifier [146]. Their approach maximized mutual information and leveraged the distribution of unlabeled and labeled data in selecting the following iteration samples. In particular, they showed that the method could be applied for fine-tuning a general model to a more specific one. A specification that is needed when applying a pre-trained model to data from another medical site. On the other hand, Kholghi et al. [147] benefited from the availability of large medical ontologies and terminologies and incorporated this external knowledge into the AL sampling strategy. Specifically, they weighted the informativeness-based metric by a domain knowledge-based importance term. This importance is expressed as a function of the concepts contained in the sequence. They showed that training a CRF model with the additional external knowledge outperformed state-of-the-art strategy by 14% in terms of annotation cost reduction. These findings highlighted the promise of integrating domain knowledge within active learning query strategies. However, these studies assumed that the annotation cost of all sentences is the same. To address this limitation, Chen et al. [112] linked the annotation cost of a sentence to its length and numbers of medical concepts. Following that assumption, they developed new AL cost-aware query strategies belonging to uncertainty-based and diversity-based approaches. Additionally, they conducted a comprehensive empirical evaluation of their proposed and well-established AL approaches using the CRF machine learning model. They discovered that uncertainty-based methods lead to a significant reduction of annotation effort compared to the diversity-based ones. Notably, their results showed that uncertainty sampling saved 66% and 42% of the whole number of sentences and tokens required annotation to reach the extraction F-measure of 0.80.

The presented studies have demonstrated the effectiveness of uncertainty and diversity-based methods using simulated experiments. These experiments were based on the corpus developed for

the i2b2/VA 2010 challenge [105], and the active learning process was simulated by incrementally training on the gold standard dataset. More recent studies ran simulations and real-life experiments to effectively estimate the actual annotation cost and compare the methods in both settings. Chen et al. [48] proposed a hybrid AL strategy, they called CAUSE, that combines the uncertainty of the model and the representativeness of the input sentence. They proposed a clustering method using Latent Dirichlet Allocation (LDA) [148] to get the sentence's feature representation and ensured a diverse sampling batch by selecting sentences from different clusters. They compared simulated experiments using Area under Learning Curve (ALC) score, and they also conducted a real-world annotation study collected from two nurses over three weeks. Their results showed a significant reduction in cost using their proposed method compared to passive learning. Interestingly, the real-life experiment results showed that AL did not guarantee less annotation time than random sampling across different users, suggesting that the sampling strategy should additionally estimate the annotation time. Following that line of thought, [113] Wei et al. modeled the annotation cost in their proposed AL strategy, called Cost-CAUSE, using a regression model. They ran a simulated study and an actual user study with nine annotators to validate their approach. Furthermore, they compared their method against random, uncertainty, and CAUSE strategies. The user study results showed a time reduction ranging between 20% and 30%.

These proposed approaches were applied to machine learning-based models such as CRF, SVM, and Naive Bayes classifiers. To the best of our knowledge, our study, introduced in chapter 5, is the first to apply AL approaches to train a Transformer-based clinical NER model incrementally. We also propose a novel hybrid strategy that dynamically calibrates the uncertainty and diversity informativeness of a given sample based on the fine-tuning stage of the model.

9.4 Simulation experiment

9.4.1 Baseline AL strategies

We set two baselines for comparing our proposed AL strategy. First, we fine-tuned the ClinicalBert model using random sampling to set the baseline of the active learning experiment. Furthermore, an active ClinicalBert with uncertainty sampling is defined as a traditional active learning strategy.

9.4.2 Named Entity Recognition task

We formulated the NER task as a sequence tagging problem using the BIO format where "B" represents the tag for the beginning of an entity, "I" for inside the entity, and "O" for outside the entity. As the task includes three types of entities, we had seven classification labels: "B-problem", "B-treatment", "B-test", "I-problem", "I-treatment", "I-test", and "O". For each token in the input sentence, the model estimates the probability of belonging to a given class.

9.4.3 Performance Metric

Following previous studies on medical NER [147, 149, 48], we evaluated the active learning strategies using learning curves that plot F-measure of the model versus number of annotated sentences. We also computed the area under the learning curve (ALC) as a global score to compare the methods.

9.4.4 Active Learning iterations

In the real-world AL context, human annotations are employed after each training iteration to assign labels to the pre-selected samples; in this study, we simulated this process based on the pre-annotated corpus of the i2b2-2010 challenge. The corpus consists of two sets of training and testing. During the AL process, the train set is used as the unlabeled pool of data from which samples are selected and labelled iteratively. The human annotations were replaced with the true labels provided by the challenge. The test set was used to evaluate the performance of the model built at each iteration of AL.

9.4.5 Experiment setup

We reserved 20% of sentences as test data and used the remaining sentences as the pool of data to be queried $\mathcal{U}$. We initially fine-tuned the ClinicalBert model using 5% of training data then ran three AL experiments. For all the three strategies, we iteratively selected a sample $\mathcal{U}$ with 5% to update the model parameters and measure the performance scores on the test sample. Table 7.3 shows the characteristics of the pool and test sets.

9.5 Result Analysis

We ran a supervised learning experiment with the 80% of training data using a cross-validation protocol. The best F1-measure score of 0.8102 was obtained with a batch size of 16, a learning rate of 5.10^{-5} and 4 epochs. Furthermore, the best ALC score of dynamic-HWUS was reached using $k = 0.25$ and $\beta_0 = 2$.

In Figure 9.1 and Table 9.1, we observed that all active learning algorithms performed better than passive learning, indicating the promise of combining AL and Transformer-based architecture in medical concept extraction. Furthermore, using an active learning strategy benefited the early stage training of the model and accelerated the learning curve to reach high-performance scores with fewer data samples.

Interestingly, the uncertainty method outperformed the hybrid weighted strategy with static $\beta = 1$ in the early stage training. We argue that the contextual embeddings learned by the pre-trained ClinicalBert are general and do not capture the specificities of the concept extraction task. Indeed, as part of the extraction task, ambiguity, polysemy, synonymy (including abbreviations), and word order variations should be addressed. However, a clinical narrative often presents unstructured,

Table 9.1: Area under learning curve (ALC) scores of the four AL strategies: Random sampling (RS), Uncertainty sampling (US), HWUS with $\beta = 1$ (Static-HWUS) and HWUS with decayed rate β (Dynamic-HWUS)

	RS	**US**	**Static-HWUS**	**Dynamic-HWUS**
ALC score	0.7070	0.7297	0.7259	0.7339

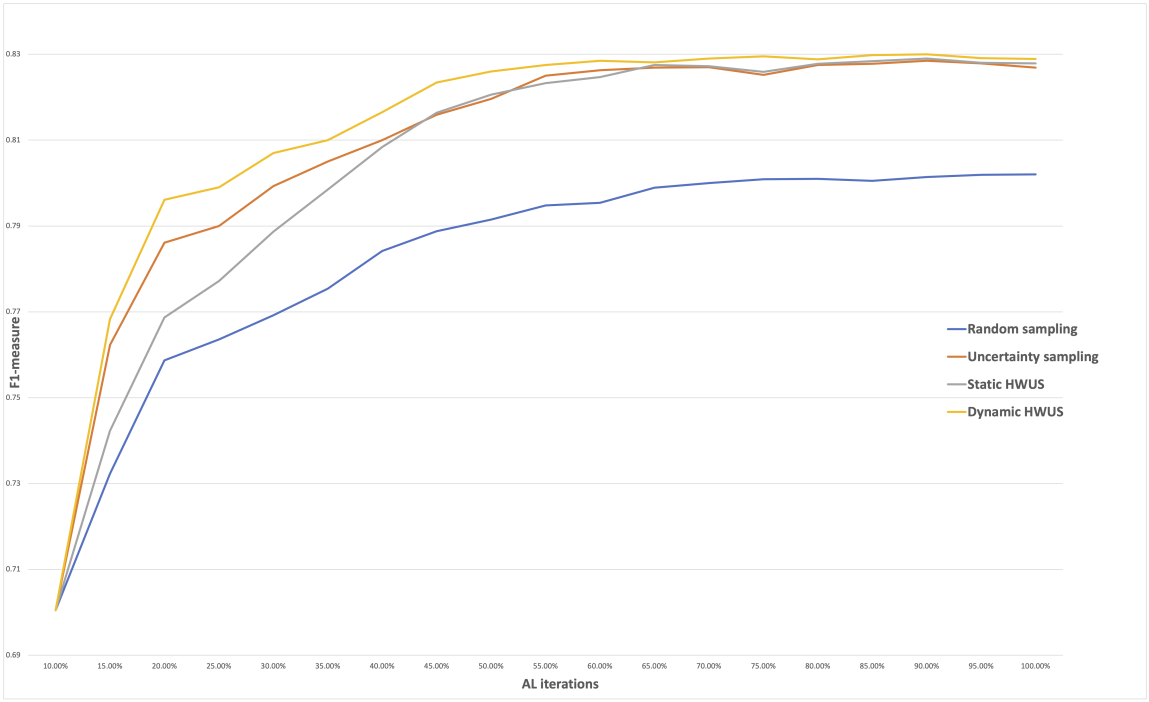


Figure 9.1: Simulated learning curves with 5% iteration step for random sampling (Random sampling), entropy-based (Uncertainty), static HWUS with constant β and dynamic HWUS with decayed β to select sampled batch

grammatically incorrect, and unorganized information. Therefore, it is crucial to fine-tune the embedded vector space based on highly variable clinical notes to learn how to represent words within the context of the medical concept extraction task. Following that line of thought, we introduced the decayed parameter β in our **HWUS** strategy to give more importance to uncertainty sampling at the beginning of fine-tuning while iteratively increasing the density-based calibration term. Figure 9.1 shows that **HWUS** performed similarly to uncertainty methods in the two first iterations before reaching higher accuracy in the following iterations, taking advantage of the calibration term in selecting effective samples.

To summarize, the simulated study demonstrated that the three approaches, Uncertainty-based, static HWUS, and Dynamic HWUS, have reduced the number of samples needed to reach the F1-measure of 0.81 by 60%, 55%, and 70%, respectively compared to passive learning.

Chapter 10

In-Hospital Mortality Use Case (IHM)

10.1 Introduction

Transformer-based architecture combined with time representation showed promising results for the diabetic retinopathy application (Chapter 8.1). In this task, the input data sample consisted of a univariate numerical time series (HbA1C measurements) that spans over a long period (average of 124 months). In this third experiments, we wanted to validate these results when applied to multivariate time series that span over a shorter period (20 months) using our proposed Multi-HiTT architecture. To that end, we used the in-hospital mortality task defined from MIMIC-III [13]. Section 7.3 gives detail about the task and the input features. A subsequent motivation is to showcase the utility of our framework in multiple clinical applications with various types of temporal inputs.

10.2 Objective

The objective was to validate the effectiveness of considering multi-modal patient timelines with hierarchical temporal structure on solving clinical prediction tasks. Our workflow primarily makes use of the multi-HiTT architecture and the implemented temporal framework.

10.3 Experiment setup

10.3.1 Cross-validation

we defined 80% and 10% of stratified sets of patients (i.e., ensuring the same distribution between positive and negative classes) for training and validating the model, respectively, during the hyper-parameters optimization step. The remaining 10% set of patients is reserved for the k-fold cross-validation step.

10.3.2 Performance metrics

We used the same performances metrics like the ones defined for the task of diabetic retinopathy prediction (Section 4.4.5): The area under the ROC curve (AUROC) and the F1-score.

10.3.3 Setup + Hyper-parameter tuning

We ran a Bayesian hyper-parameter optimization with 100 trials to maximize the AUROC score and computed the final scores using nested 5-fold cross-validation. Appendix D lists the parameters search space for our proposed architecture and the baselines. We applied L2-weight decay regularization to all models to prevent over-fitting and used a weighted cross-entropy loss for training. The experiments were run on a V-100 GPU with 32GB of memory.

10.4 Results analysis

The predictive performance of Hierarchical Transformer and baselines are presented in Table 10.1. The results shown are based on predicting in-hospital mortality after 24h of the admission and using a window of 6 hours to construct the sub-sequences of each visit with a maximum of 5 visits history. According to Table 10.1, the Transformer-based models are generally capable of achieving higher prediction performance in terms of AUROC. Our proposed architecture, Hierarchical Transformer, achieves the best performance, improving the AUROC by 5 % over the best vanilla baseline. The second-best performing model is the Hierarchical BiLSTM that shares the same architecture but with a different sequence encoder. This shows that our design choices of hierarchical architecture, attention-based sequence aggregator, and time gaps representation enable our model to learn superior representations. Like retinopathy application results, ignoring the time gaps between visits hurts the model's predictive power. This validates the importance of considering time information when modeling medical time series. Lastly, the Transformer-based encoder leads to better results than the BiLSTM one, suggesting that the self-attention encoder is a strong candidate for modeling short sequences as well.

Table 10.1: Test results: Mean and std of AUROC using 5 cross-validation sets of 512 patients.

Model	**ROC - AUROC - mean**	**ROC - AUROC - std**
Vanilla BiLSTM	0.852	0.096
Vanilla Transfomer	0.863	0.022
Hierarchical BiLSTM	0.885	0.008
Hierarchical Transformer - No Time	0.881	0.013
Hierarchical Transformer	**0.906**	0.002

On the other hand, Table 10.2 reports the average AUROC score over the 5-fold nested cross-validation for the five best architecture of Multi-HiTT variants. Taking away the diagnoses recorded during patient visits hurts the model's performance the most. Conversely, ignoring the procedures embeddings slightly decreased the prediction score. Meanwhile, when considering all available information, including text embeddings, we reached the highest score of 0.953. We could conclude that multi-modal representation of the patient health trajectory helped to improve the performance of the IHM predictive task. Although we argue that the choice of information to include highly

depends on the classification task and the input dataset.

Table 10.2: Area under ROC curve (AUROC) scores of the five variants of HiTT architectures for in-hospital mortality prediction task

	HiTT	**HiTT - continuous**	**HiTT - diagnoses**	**HiTT - procedures**	**HiTT + text**
AUROC score	0.906	0.875	0.851	0.898	**0.953**

Discussion

Time irregularity as additional information

Modeling the irregular time gaps between consecutive medical visits is critical for representing the patient's health trajectory as real-world data always contains such patterns. Moreover, these patterns are often valuable for understanding the disease evolution. However, most conventional statistical and machine learning methods require the aggregation of the temporal indicators using a time window segmentation, leading to information loss.

Our work considered time irregularities as additional information, arguing that it is relevant to the medical prediction tasks. We also kept the original sequences oh historical observations without using time-based aggregation or data imputation techniques. To represent the time between consecutive events, we compared two approaches: learning to express the time as an additional input to the neural method and injecting time information as an additional parameter of the neural architecture. To our knowledge, this is the first work that compares models from the two approaches when applied to highly variable real-world medical time series. Our experiment results of the retinopathy prediction use-case (Fig. 8.2) suggest that representing time as an additional input feature leads to better performance. We argue that the second parametric approach is defining a more complex architecture with more parameters to optimize and therefore requires more data points to stabilize. To validate the latter point, we plan to collect more data in our future work to validate if such patterns are persistent when learning with a larger dataset.

On the other hand, the C-LSTM variants that include the time as part of their learning pa-

rameters show that the model needs patients with long-history profiles to reach higher performance scores. A result that consolidates the original work findings: *"DeepCare is more powerful with long trajectories of many episodes"* [53].

Besides, Fig. 8.3 plots the models variability with respect to the cross-validation folds. We observe that architectures not leveraging the information of time gaps are less stable than those using the same architecture but with the temporal representation technique. Indeed, both the attention-based and bidirectional-LSTM models are reaching the highest scores when combined with *time_concat_mlp* and *time_concat_soft* representation. Furthermore, injecting the time in the attention-based model's positional encoding layer leads to a higher score. On the other hand, the time-mask technique has a comparable score to the no-time models suggesting that learning to contextualize the event using a temporal mask is not enough to leverage the information contained in the temporal irregularity.

Lastly, Multi-HiTT architecture leverages the time irregularity in two levels: time-series recorded within the visit and time-gaps between consecutive visits. Decoupling the representation of these short and long-term temporal dependencies improved the model's performances by 5%, suggesting that time modeling should be considered at the visit-level and patient timeline-level. This hierarchical structure was not leverage in the different DL proposed methods for IMTS [72, 68, 22].

This analysis concludes that the temporal irregularity and the hierarchical attention-based temporal weights are crucial to accurately represent the highly variable sequence of medical observations and highlights its impact on the prediction of retinopathy complications and in-hospital mortality.

Trade off between training data size and patient's sequence length

Our proposed experiment of retinopathy prediction defined four datasets where we filtered patients based on the number of HbA1c measurements. Then, we executed the pipeline of hyperparameter tuning, training, and evaluation of all considered algorithms for each group. Our goal was to assess the trade-off between the size of training data and the length of the patient's history when comparing performance. Fig. 8.2 suggests that experiments conducted on $Group \geq 3$

(patients with at least three records) lead to the best-performing model as rich training data with different profiles help the model generalizability [150]. Whereas the prediction analysis of the performing model (Fig. 8.4b) showed that F1-score was higher for patients with a more significant number of records. We could conclude that the model benefits from training with extensive data containing variable patient profiles to represent the sequence inputs better. However, the model is more confident about the predicted risk score for patients with sufficient data records.

Therefore, the real-life application of such models should consider the observed trade-off. First, the medical research team needs to efficiently build large train datasets, including variable profiles of patients, to get a high-performance trained model. Nevertheless, integrating such a model into clinical decision support tools must compute those scores only for patients with sufficient records (the cut-off in our experiments was eight records).

Most efficient architecture for retinopathy prediction based on the conducted comparative study

Most models presented in the literature (presented in section 8.2) used conventional machine learning methods requiring a feature engineering step that derives input attributes from the original sequence of medical observations. Our work achieves higher scores (an improvement of 7.7% and 8.1% compared to the closest related works [136], and [151] respectively) by applying deep learning methods to the original sequence of HbA1c records, considering the temporal irregularity and three patient-level features. These promising results show how we can quickly test our clinical hypothesis using our developed deep learning framework. We could also achieve higher scores than conventional machine learning methods while using a smaller set of features, integrating temporal variables, and keeping the original data to avoid loss of information.

Figure. 8.1 shows that Transformer-based and Bidirectional-LSTM models combined with time feature representations lead to higher scores. However, selecting the best appropriate algorithm among these five models is challenging as they all lead to comparable performance scores. One could argue that the best model is *attention + time encode* as it has a minor variance, while others

would instead select the model *attention + time_concat_soft* as it leads to the highest score. These results raise the question of choosing the relevant algorithm for a specific clinical application and available dataset without the models' comparison done in this work.

A good practice for such clinical studies is to run the following pipeline: pre-process input data, build the models' architectures, find the best hyperparameters for each model, and evaluate using the cross-validation protocol. Our implemented framework integrates all those steps and includes the latest temporal representation techniques and sequential neural networks, aiming to facilitate the experiment's workflow for medical researchers. Furthermore, our framework supports the processing of contextual information such as patient-level features. The results of Fig. 8.5 show that the patient's side information helped the best performing neural network models to achieve higher prediction scores when we included patients with few numbers of HbA1c measures. Lastly, the faster execution time of attention-based models can be an argument to use those types of architectures when large datasets and access to GPU computing power are available.

Medical benefits of deep learning-based methods for complication predictions Our work aims to reduce this gap by studying a cohort of type 1 diabetes as a recent review [142] outlined the lack of studies predicting complications of patients with type 1 diabetes. To our knowledge, this is the first work to apply deep learning models to longitudinal clinical data for retinopathy prediction.

We used HbA1c routines tests combined with three demographic features to define our models. All these features are recorded during the following up of diabetic patients, making our trained model easily deployable in every monitoring system to assist doctors in managing higher-risk patients.

Collection of high-quality annotated clinical notes Our simulation study showed that the model achieved 80% F1-score with 70% fewer data than passive supervised learning when using Dynamic-HWUS active learning. Based on this result, we can show the validity of training Transformer-based NER classifiers using Active Learning. As well, it demonstrates a good poten-

tial for applying state-of-the-art NLP models that were published in the general domain to complex real-life medical data. An active-learning platform would allow doctors to annotate large amounts of clinical notes quickly and easily, and to prepare high-quality training data for downstream use. We believe the adoption of Active Learning should be at the core of collecting complex annotated medical data.

Chapter 11

Conclusion and Future Works

11.1 Conclusion

Unlike general domains such as Image Analysis and Natural Language Processing, real-world medical data is very challenging and requires additional processing to represent the patient's history and build solid predictive models. Considering these challenges when defining the deep learning-based representation model is crucial for accurate medical studies.

In our first work, we tackled the time-irregularity of episodic medical visits, tested various architecture to model the sequence of a patient's health history, and included functional representations of the irregular time gaps between consecutive observations. The developed generic framework allowed us to consolidate the importance of temporality management in two clinical settings: predicting diabetic retinopathy using a long-span chronic medical time series and estimating the risk of in-hospital mortality using short critical care time series.

Second, an overview of existing methods for medical concept extraction suggested that Transformer-based architecture is the most performant architecture reaching SOTA scores in all the i2b2 medical extraction challenges. These architectures contain a large number of parameters and thus require large annotated corpora to converge. In real-world clinical settings, annotations are often unavail-

able, limiting the application of such deep learning models. We proposed a novel active learning strategy that enables incremental training of Transformer-based architectures. The simulated study showed promising results as the model reached an 80% extraction score with 70% fewer data than passive supervised learning. Our results proved the validity of training a Transformer-based NER classifier using the Active Learning strategy. Such a technique would reduce the gap between research models' results and their application to supporting e-Health systems in practice.

Lastly, the key to building accurate and comprehensive patient representations is combining all forms of clinical information gathered during patient visits. The proposed multi-modal HiTT architecture offers a novel structure to combine the different levels of sequential information: visit-level and patient-level. Additionally, it supports any type of data integration, including continuous, categorical, and textual. The end-to-end training of such an architecture provided SOTA results for predicting in-hospital mortality.

11.2 Future directions

Open-source English databases such as MIMIC-III and i2b2 challenges have led to numerous research works that advanced the application of deep learning models to real-world medical data and allowed research teams to ensure their work's reproducibility and comparison against existing methods. Meanwhile, there are relatively fewer published studies that use french clinical databases. We argue that the main reason is the limited access to a large-scale medical dataset and the absence of publicly available research databases. We believe that developing such databases could be highly beneficial to the French research teams.

While datasets' availability enables efficient models, the research teams should consider their applicability in real-world clinical systems designed for doctors. To achieve that, those models should account for all the challenges of medical data stated in the introduction of this work. We believe that designing an end-to-end system that encapsulates optimized modules tackling the different challenges could bridge the gap between the definition of the state-of-the-art deep learning methods and their integration in routine monitoring systems used by doctors.

Our preliminary work presented in the first part showed the effectiveness of temporal-based deep learning-based methods in clinical predictions. Before deploying such algorithms in clinical settings, we need to ensure the following points:

1. **Model's explainability**: conduct a prediction analysis to evaluate the impact of each considered variable. To understand the predictability of such complex models, it is also necessary to visualize the attention scores learned by the models. Eventually, it will allow us to detect patterns in the time series that correlate with a higher risk.

2. **Higher generalizability power**: consider a more comprehensive set of training data in terms of the number of patients. Our goal is to verify whether the temporal patterns and important factors observed in this study can be generalized.

3. **Unit and integration tests**: extend the framework to include automatic testing of the model definition and deployments workflows to ensure the stability of the deployed model performances. The trained model should go through a rigorous validation process before integrating it into a routine monitoring system used by doctors.

4. **Dynamic training**: Design an efficient pipeline that integrates new records and dynamically updates the models' parameters using time-window-based training. Indeed, incremental training [152, 153] would allow for continuously learning new knowledge from new patient's observations while existing past knowledge is maintained.

For more clinically relevant application, we plan to extend and test our framework to temporal window-based predictions to estimate the clinical outcome in a more distant future step, allowing doctors to intervene ahead of time and better manage their patients. Transparency and explainability play a crucial role in adopting Artificial Intelligence (AI) models into clinical systems, as incorrect predictions could significantly impact patients' health [154, 155]. Therefore, another refinement we intend to study is the predictions analysis to explain such black-box models. Finally, once validated, these models could be integrated into a decision-support tool of numerous medical information systems as they only require routine data.

In the second part, we proposed Hybrid Weighted Uncertainty Sampling (HWUS) that considers the context embedding learned by the Transformer-based approach to measuring the representativeness of samples. This work has diverse notable limitations. First, our simulation study assumed that the annotation cost for each sentence was the same. In practice, this hypothesis does not hold [156, 113], which limits the generalization of the obtained results to real-time experiments [157] with medical experts. Second, we used as base architecture the ClinicalBert model, pre-trained on English clinical notes from the MIMIC-III database. Transposing these AL experiments to french data would require an additional step of pre-training the contextual embeddings using a large set of medical notes. Efforts [158, 159] have already been made to publish contextual embeddings trained on general french texts, reaching high-performance scores in various NLP tasks. These models can serve as a base to learn clinical embeddings using french medical notes. Lastly, the effectiveness of active learning for clinical NER needs to be evaluated with real-world experiments. As future work, we aim to conduct these user studies based on a cost-aware active learning-enabled annotation interface and to involve medical experts.

The study presented in the last chapter of the second part showed that it would be beneficial to integrate clinical narratives with the sequential patient representation since unstructured text provides highly valuable information. Nevertheless, our work has three main limitations that need to be addressed. First, we trusted the fine-tuned ClinicalBert extractor to gather the medical concepts. A post-processing step to evaluate the extraction performance of the model on MIMIC-III is required to validate the considered outputs. Second, we used fixed pre-trained embeddings as input to our HiTT architecture. A potential scenario to test in the future is the possibility of fine-tuning these embeddings during the IHM training step, as it would lead to task-aware embeddings and thus improve performance. Finally, we would like to run similar experiments using a french clinical database to validate the generalizability of our multimodal architecture. In particular, we are interested in building French ClinicalBert pre-trained embeddings using a large set of clinical notes and publicly available medical courses materials.

Specifically, a future direction we would like to explore is the development of an end-to-end

system that combines the active learning-based annotation tool **MAP** (The design of the interface is detailed in Appendix C) and the temporal model **HiTT**. We aim to conduct a real-life experiment using a French clinical database and evaluate the effectiveness of building such an end-to-end system.

Bibliography

[1] William R Hersh. Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. *Clin Pharmacol Ther*, 81:126–128, 2007.

[2] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13:55–75, 2018.

[3] Michela Assale, Linda Greta Dui, Andrea Cina, Andrea Seveso, and Federico Cabitza. The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Frontiers in Medicine*, 6, 2019.

[4] Xavier Amatriain. NLP & Healthcare: Understanding the Language of Medicine, November 2018.

[5] Suvajit Dutta, BC Manideep, Syed Muzamil Basha, Ronnie D Caytiles, and NCSN Iyengar. Classification of diabetic retinopathy images by using deep learning models. *International Journal of Grid and Distributed Computing*, 11:89–106, 2018.

[6] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[7] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. LSTM Fully Convolutional Networks for Time Series Classification. *IEEE Access*, 6:1662–1669, 2018.

[8] Nelly Elsayed, Anthony S. Maida, and Magdy Bayoumi. Deep Gated Recurrent and Convolutional Network Hybrid Model for Univariate Time Series Classification. *arXiv:1812.07683 [cs, stat]*, 2019.

[9] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33:917–963, 2019.

[10] Truyen Tran, Tu Dinh Nguyen, Dinh Phung, and Svetha Venkatesh. Learning vector representation of medical objects via emr-driven nonnegative restricted boltzmann machines (enrbm). *Journal of Biomedical Informatics*, 54, 2015.

[11] Riccardo Miotto, Li Li, Brian A. Kidd, and Joel T. Dudley. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, 6:1–10, 2016.

[12] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. pages 3504–3512. 2016.

[13] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 2016.

[14] Ozlem Uzuner, Peter Szolovits, and Isaac Kohane. i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association*, 2006.

[15] Marie-Helene Metzger Sara Rabhi, Jérémie Jakubowicz. Deep learning versus conventional machine learning for detection of healthcare-associated infections in french clinical narratives. *Methods of information in medicine*, 58:031—41, 2019.

[16] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. Risk Prediction with Electronic Health Records: A Deep Learning Approach. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 432–440, 2016.

[17] Edward Choi, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association: JAMIA*, 24:361–370, 2017.

[18] Xiaokang Zhou, Yue Li, and Wei Liang. Cnn-rnn based intelligent recommendation for online medical pre-diagnosis support. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[19] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *arXiv:1511.05942 [cs]*, 2016.

[20] Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9:1–12, 2019.

[21] Xi Zhang, Jingyuan Chou, Jian Liang, Cao Xiao, Yize Zhao, Harini Sarva, Claire Henchcliffe, and Fei Wang. Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. *Scientific reports*, 9:1–12, 2019.

[22] Sajad Darabi, Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. TAPER: Time-Aware Patient EHR Representation. *arXiv:1908.03971 [cs, stat]*, 2020.

[23] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[24] Zhi Qiao, Xian Wu, Shen Ge, and Wei Fan. Mnn: multimodal attentional neural networks for diagnosis prediction. *Extraction*, 1:A1, 2019.

[25] Alexander Pate, Richard Emsley, Matthew Sperrin, Glen P Martin, and Tjeerd van Staa. Impact of sample size on the stability of risk scores from clinical prediction models: a case study in cardiovascular disease. *Diagnostic and prognostic research*, 4(1):1–12, 2020.

[26] World Health Organization et al. International classification of diseases—ninth revision (icd-9). *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 63(45):343–344, 1988.

[27] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[28] J Andrew, Shaun Shibu Mathew, and Batra Mohit. A comprehensive analysis of privacy-preserving techniques in deep learning based disease prediction systems. In *Journal of Physics: Conference Series*, volume 1362, page 012070. IOP Publishing, 2019.

[29] Xi Yang, Tianchen Lyu, Qian Li, Chih-Yin Lee, Jiang Bian, William R Hogan, and Yonghui Wu. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(5):1–9, 2019.

[30] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[31] Priyanka Gupta, Pankaj Malhotra, Jyoti Narwariya, Lovekesh Vig, and Gautam Shroff. Transfer learning for clinical time series analysis using deep neural networks. *Journal of Healthcare Informatics Research*, 4(2):112–137, 2020.

[32] Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. Transfer learning for named-entity recognition with neural networks. *arXiv preprint arXiv:1705.06273*, 2017.

[33] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

[34] Jiayu Zhou, Lei Yuan, Jun Liu, and Jieping Ye. A multi-task learning formulation for predicting disease progression. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 814–822, 2011.

[35] Jing Zhao, Panagiotis Papapetrou, Lars Asker, and Henrik Boström. Learning from heterogeneous temporal data in electronic health records. *Journal of Biomedical Informatics*, 65:105–119, 2017.

[36] Melanie Villani, Arul Earnest, Natalie Nanayakkara, Karen Smith, Barbora de Courten, and Sophia Zoungas. Time series modelling to forecast prehospital EMS demand for diabetic emergencies. *BMC Health Services Research*, 17:332, 2017.

[37] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

[38] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019.

[39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, June 2019.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[41] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52(2):571–601, 2018.

[42] William F Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. Temporal an-

notation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154, 2014.

[43] Pinal Patel, Disha Davey, Vishal Panchal, and Parth Pathak. Annotation of a large clinical entity corpus. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2042, 2018.

[44] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397, 2010.

[45] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[46] Ines Rehbein, Josef Ruppenhofer, and Alexis Palmer. Bringing active learning to life. In *Proceedings of the 23rd international conference on computational linguistics (COLING 2010)*, pages 949–957, 2010.

[47] Jingbo Zhu, Huizhen Wang, Benjamin K. Tsou, and Matthew Ma. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1323–1331, 2010.

[48] Yukun Chen, Thomas A Lask, Qiaozhu Mei, Qingxia Chen, Sungrim Moon, Jingqi Wang, Ky Nguyen, Tolulola Dawodu, Trevor Cohen, Joshua C Denny, et al. An active learning-enabled annotation system for clinical named entity recognition. *BMC medical informatics and decision making*, 17(2):35–44, 2017.

[49] Omolola Ogunyemi and Dulcie Kermah. Machine learning approaches for detecting diabetic retinopathy from clinical and public health records. In *AMIA Annual Symposium Proceedings*, volume 2015, page 983. American Medical Informatics Association, 2015.

[50] Jeong Min Lee and Milos Hauskrecht. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artificial Intelligence in Medicine*, page 102021, 2021.

[51] Phuoc Nguyen, Truyen Tran, Nilmini Wickramasinghe, and Svetha Venkatesh. $\mathtt Deepr$: A Convolutional Net for Medical Records. *IEEE Journal of Biomedical and Health Informatics*, 21:22–30, 2017.

[52] AM Diallo, JL Novella, C Lukas, PF Souchon, M Dramé, M François, B Decoudier, S Barraud, AS Salmon, D Ancelle, C Arndt, and B Delemer. Early predictors of diabetic retinopathy in type 1 diabetes: The Retinopathy Champagne Ardenne Diabète (ReCAD) study. *J Diabetes Complications*, 32:753–758, 2018.

[53] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69:218–229, 2017.

[54] Manxia Liu, Fabio Stella, Arjen Hommersom, Peter J. Lucas, Lonneke Boer, and Erik Bischoff. A comparison between discrete and continuous time Bayesian networks in learning from clinical time series data with irregularity. *Artificial Intelligence in Medicine*, 95, 2019.

[55] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In *Advances in Neural Information Processing Systems 32*, pages 5243–5253. Curran Associates, Inc., 2019.

[56] David Salinas, Valentin Flunkert, and Jan Gasthaus. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *arXiv:1704.04110 [cs, stat]*, 2019.

[57] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[58] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[59] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[60] Junyoung Chung et al. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[61] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[62] Xiaobing Sun and Wei Lu. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, 2020.

[63] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning–based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021.

[64] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[65] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[66] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, 2019.

[67] Chenxi Sun, Shenda Hong, Moxian Song, and Hongyan Li. A Review of Deep Learning Methods for Irregularly Sampled Medical Time Series Data. *arXiv:2010.12493 [cs, stat]*, 2020. arXiv: 2010.12493.

[68] Luntian Mou, Pengfei Zhao, Haitao Xie, and Yanyan Chen. T-lstm: A long short-term memory neural network enhanced by temporal information for traffic flow prediction. *IEEE Access*, pages 98053–98060, 2019.

[69] Yang Li, Nan Du, and Samy Bengio. Time-Dependent Representation for Neural Event Sequence Prediction. *arXiv:1708.00065 [cs]*, 2018.

[70] Da Xu, Chuanwei Ruan, Sushant Kumar, Evren Korpeoglu, and Kannan Achan. Self-attention with Functional Time Representation Learning. *arXiv:1911.12864 [cs, stat]*, 2019.

[71] Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to Do Next: Modeling User Behaviors by Time-LSTM. page 7, 2017.

[72] Qingxiong Tan, Mang Ye, Baoyao Yang, Siqi Liu, Andy Ma, Terry Yip, Grace Wong, and PongChi Yuen. DATA-GRU: Dual-Attention Time-Aware Gated Recurrent Unit for Irregular Multivariate Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:930–937, 2020.

[73] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8:1–12, 2018.

[74] Inci M. Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K. Jain, and Jiayu Zhou. Patient Subtyping via Time-Aware LSTM Networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 65–74, Halifax NS Canada, 2017.

[75] 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*, pages 261–269. SIAM, 2018.

[76] Luchen Liu, Haoran Li, Zhiting Hu, Haoran Shi, Zichang Wang, Jian Tang, and Ming Zhang. Learning hierarchical representations of electronic health records for clinical outcome predic-

tion. In *AMIA Annual Symposium Proceedings*, volume 2019, page 597. American Medical Informatics Association, 2019.

[77] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 787–795, 2017.

[78] Ke Yu, Mingda Zhang, Tianyi Cui, and Milos Hauskrecht. Monitoring icu mortality risk with a long short-term memory recurrent neural network. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 103–114. World Scientific, 2019.

[79] Sindhu Tipirneni and Chandan K Reddy. Self-supervised transformer for multivariate clinical time-series with missing values. *arXiv preprint arXiv:2107.14293*, 2021.

[80] P Nguyen, T Tran, and S Venkatesh. Finding algebraic structure of care in time: A deep learning approach. arxiv [cs. lg] 2017.

[81] Jinghe Zhang, Kamran Kowsari, James H Harrison, Jennifer M Lobo, and Laura E Barnes. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access*, 6:65333–65346, 2018.

[82] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911, 2017.

[83] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. Predicting clinical events by combining static and dynamic information using recurrent neural networks. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 93–101. IEEE, 2016.

[84] Satya Narayan Shukla and Benjamin M Marlin. Interpolation-prediction networks for irregularly sampled time series. *arXiv preprint arXiv:1909.07782*, 2019.

[85] Jeong Min Lee and Milos Hauskrecht. Recent-context-aware lstm-based clinical time-series prediction. *In Proceedings of AI in Medicine Europe (AIME)*, 2019.

[86] Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 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*, pages 1315–1324, 2017.

[87] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[88] Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion. Deep transformer models for time series forecasting: The influenza prevalence case. *arXiv preprint arXiv:2001.08317*, 2020.

[89] Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. Towards generating a patient's timeline: extracting temporal relationships from clinical notes. *Journal of biomedical informatics*, 46:S40–S47, 2013.

[90] Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61, 2016.

[91] Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Hua Xu. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, 17(2):53–61, 2017.

[92] Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304, 2019.

[93] Naomi Sager, Carol Friedman, and Margaret S Lyman. *Medical language processing: computer management of narrative data*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[94] Lois C Childs, Robert Enelow, Lone Simonsen, Norris H Heintzelman, Kimberly M Kowalski, and Robert J Taylor. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. *Journal of the American Medical Informatics Association*, 16(4):571–575, 2009.

[95] Timothy Miller, Dmitriy Dligach, Steven Bethard, Chen Lin, and Guergana Savova. Towards generalizable entity-centric clinical coreference resolution. *Journal of biomedical informatics*, 69:251–258, 2017.

[96] Cody C Wyles, Meagan E Tibbo, Sunyang Fu, Yanshan Wang, Sunghwan Sohn, Walter K Kremers, Daniel J Berry, David G Lewallen, and Hilal Maradit-Kremers. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *The Journal of bone and joint surgery. American volume*, 101(21):1931, 2019.

[97] Lance De Vine, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon, and Anthony Nguyen. Analysis of word embeddings and sequence features for clinical information extraction. In *Australasian Language Technology Association Workshop 2015: Proceedings of the Workshop*, pages 21–30. Australasian Language Technology Association (ALTA), 2015.

[98] Yanshan Wang, Sijia Liu, Naveed Afzal, Majid Rastegar-Mojarad, Liwei Wang, Feichen Shen, Paul Kingsbury, and Hongfang Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20, 2018.

[99] Sebastien Dubois, Nathanael Romano, David C Kale, Nigam Shah, and Kenneth Jung. Effective representations of clinical notes. *arXiv preprint arXiv:1705.07025*, 2017.

[100] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[101] Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360, 2018.

[102] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th workshop on biomedical natural language processing*, pages 166–174, 2016.

[103] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018.

[104] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[105] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[106] Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*, 2018.

[107] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[108] Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

[109] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[110] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[111] Yukun Chen, Subramani Mani, and Hua Xu. Applying active learning to assertion classification of concepts in clinical text. *Journal of biomedical informatics*, 45(2):265–272, 2012.

[112] Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18, 2015.

[113] Qiang Wei, Yukun Chen, Mandana Salimi, Joshua C Denny, Qiaozhu Mei, Thomas A Lasko, Qingxia Chen, Stephen Wu, Amy Franklin, Trevor Cohen, et al. Cost-aware active learning for named entity recognition in clinical text. *Journal of the American Medical Informatics Association*, 26(11):1314–1322, 2019.

[114] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *arXiv preprint arXiv:2009.00236*, 2020.

[115] Burr Settles. Active learning literature survey. 2009.

[116] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[117] Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

[118] Murray Aitkin and Rob Foxall. Statistical modelling of artificial neural networks using the multi-layer perceptron. *Statistics and Computing*, 2003.

[119] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings*

of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 207–212. Association for Computational Linguistics, 2016.

[120] Christopher M Bishop et al. *Neural networks for pattern recognition*. Oxford university press, 1995.

[121] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

[122] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[123] Jeremy Howard et al. fastai. `https://github.com/fastai/fastai`, 2018.

[124] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10), 2013.

[125] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[126] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[127] Pascale Massin and Sylvie Feldman-Billard. Référentiel pour le dépistage et la surveillance des complications oculaires du patient diabétique – 2016. Validé par la Société Francophone du Diabète (SFD) et par la Société Française d'Ophtalmologie (SFO). *Médecine des Maladies Métaboliques*, pages 774–784, 2016.

[128] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[129] Arianna Dagliati, Simone Marini, Lucia Sacchi, Giulia Cogni, Marsida Teliti, Valentina Tibollo, Pasquale De Cata, Luca Chiovato, and Riccardo Bellazzi. Machine learning methods to predict diabetes complications. *Journal of diabetes science and technology*, 12(2):295–302, 2018.

[130] Ahmed M. Alaa, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, and Mihaela van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS One*, 14:e0213653, 2019.

[131] Sanjay Basu, Karl T. Johnson, and Seth A. Berkowitz. Use of Machine Learning Approaches in Clinical Epidemiological Research of Diabetes. *Current Diabetes Reports*, 20:80, December 2020.

[132] Branimir Ljubic, Ameen Abdel Hai, Marija Stanojevic, Wilson Diaz, Daniel Polimac, Martin Pavlovski, and Zoran Obradovic. Predicting complications of diabetes mellitus using advanced machine learning algorithms. *Journal of the American Medical Informatics Association: JAMIA*, 27:1343–1351, 2020.

[133] Oleg Metsker, Kirill Magoev, Alexey Yakovlev, Stanislav Yanishevskiy, Georgy Kopanitsa, Sergey Kovalchuk, and Valeria V. Krzhizhanovskaya. Identification of risk factors for patients with diabetes: diabetic polyneuropathy case study. *BMC medical informatics and decision making*, 20:201, 2020.

[134] Omolola I Ogunyemi, Meghal Gandhi, Martin Lee, Senait Teklehaimanot, Lauren Patty Daskivich, David Hindman, Kevin Lopez, and Ricky K Taira. Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system. *JAMIA open*, 4(3):ooab066, 2021.

[135] K. M. Alabdulwahhab, W. Sami, T. Mehmood, S. A. Meo, T. A. Alasbali, and F. A. Alwadani. Automated detection of diabetic retinopathy using machine learning classifiers. *European Review for Medical and Pharmacological Sciences*, 25:583–590, 2021.

[136] Liang Zhou, Xiaoyuan Zheng, Di Yang, Ying Wang, Xuesong Bai, and Xinhua Ye. Application of multi-label classification models for the diagnosis of diabetic complications. *BMC medical informatics and decision making*, 21:182, 2021.

[137] Bin Cao, Ning Zhang, Yuanyuan Zhang, Ying Fu, and Dong Zhao. Plasma cytokines for predicting diabetic retinopathy among type 2 diabetic patients via machine learning algorithms. *Aging*, 13(2):1972–1988, December 2020.

[138] Omolola I. Ogunyemi, Meghal Gandhi, and Chandler Tayek. Predictive Models for Diabetic Retinopathy from Non-Image Teleretinal Screening Data. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2019:472–477, 2019.

[139] Marcus Lind, Aldina Pivodic, Ann-Marie Svensson, Arndis F. Ólafsdóttir, Hans Wedel, and Johnny Ludvigsson. HbA1c level as a risk factor for retinopathy and nephropathy in children and adults with type 1 diabetes: Swedish population based cohort study. *BMJ (Clinical research ed.)*, 366:l4894, 2019.

[140] Aki Kato, Keiichiro Fujishima, Kazuhisa Takami, Naomi Inoue, Noriaki Takase, Norihiro Suzuki, Katsuya Suzuki, Soichiro Kuwayama, Akiko Yamada, Katsuo Sakai, et al. Remote screening of diabetic retinopathy using ultra-widefield retinal imaging. *Diabetes Research and Clinical Practice*, page 108902, 2021.

[141] Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. Deep learning for diabetes: a systematic review. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[142] Qingqing Xu, Liye Wang, and Sujit S Sansgiry. A systematic literature review of predicting diabetic retinopathy, nephropathy and neuropathy in patients with type 1 diabetes using machine learning. *J. Med. Artif. Intell*, 3:1–13, 2020.

[143] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

[144] Romer Rosales, Praveen Krishnamurthy, and R. Bharat Rao. Semi-supervised active learning for modeling medical concepts from free text. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 530–536, 2007.

[145] Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[146] Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer, 1998.

[147] Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, and Anthony Nguyen. External knowledge and query strategies in active learning: a study in clinical information extraction. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 143–152, 2015.

[148] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[149] Mahnoosh Kholghi. *Active learning for concept extraction from clinical free text.* PhD thesis, Queensland University of Technology, 2017.

[150] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[151] Wei-Chun Lin, Jimmy S Chen, Michael F Chiang, and Michelle R Hribar. Applications of artificial intelligence to electronic health record data in ophthalmology. *Translational vision science & technology*, 9:13–13, 2020.

[152] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

[153] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[154] Federico Cabitza, Raffaele Rasoini, and Gian Franco Gensini. Unintended consequences of machine learning in medicine. *Jama*, 318:517–518, 2017.

[155] Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32:18069–18083, 2020.

[156] Katrin Tomanek and Udo Hahn. A comparison of models for cost-sensitive active learning. In *Coling 2010: Posters*, pages 1247–1255, 2010.

[157] Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1. Vancouver, CA:, 2008.

[158] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[159] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*, 2019.

[160] Dongyu Zhang, Jidapa Thadajarassiri, Cansu Sen, and Elke Rundensteiner. Time-aware transformer-based network for clinical notes series prediction. In *Machine Learning for Healthcare Conference*, pages 566–588. PMLR, 2020.

[161] Lei Lin, Beilei Xu, Wencheng Wu, Trevor W Richardson, and Edgar A Bernal. Medical time series classification with hierarchical attention-based temporal convolutional networks: A case study of myotonic dystrophy diagnosis. In *CVPR Workshops*, pages 83–86, 2019.

[162] Bryan Lim and Mihaela van der Schaar. Disease-atlas: Navigating disease trajectories using deep learning. In *Machine Learning for Healthcare Conference*, pages 137–160. PMLR, 2018.

[163] Jacek M Bajor and Thomas A Lasko. Predicting medications from diagnostic codes with recurrent neural networks. 2016.

Appendix A

YAML configuration of the temporal Framework

```yaml
name: 'time_concat_soft'
# Specify paths to  data and results directory
data_directory:  ./modelisation_hba1c_retino_v3/dictionaries/
validation_file_name:  "121_patients_min_3_seq_infos_Valid.sav"
train_file_name:  "965_patients_min_3_seq_infos_Train.sav"
test_file_name:  "121_patients_min_3_seq_infos_Test.sav"
min_measurement: 3
use_time_info: True
use_patient_info: True
result_dir: /home/dataset/temporal_hba1c/results
side_info: ['duree_non_suivi_norm']
# Training info
cycle_len: 30
batch_size: 16
```

```yaml
val_bs: 121
device: cuda
n_workers: 4
class_weight: [0.6, 0.4]
max_len: 151
#optim info
optimizer: radam
weight_decay: 0.000007751
max_lr: 0.001261
monitor: auc_score
mode: max
# patient module
patient_config:
  representation_type: mlp
  num_inputs: 1
  hidden_dims: [32]
  activation: 'gelu'
  output_dim: 56
# time module config
time_config:
  representation_type: soft-one-hot
  hidden_dims: None
  projection_size: 12
  activation: 'gelu'
  output_dim: 56
  embeddings_init_std: 0.2
# event module config
```

```yaml
event_config:
  categoricals: []
  continuous: ['seq_hba1c']
  categorical_representation: mlp
  continuous_representation: mlp
  categorical_embeddings: []
  continuous_hidden_dims: [16]
  continuous_output_dim: 40
  tf_activation: 'relu'
#model info
aggregation_mode: concat
model_type: attention
temporal_model:
  input_size : None
  hidden_size : None
  dropout : None
  bidirectional : False
  model_time : None
  timedecay_size : None
  num_layers : 4
  pad_value : -0.2
  attn_heads : None
  hidden_dropout_prob : 0.1
  feed_forward_hidden : 56
  hidden_act : 'gelu'
  attn_dropout_prob : 0
  num_input :  1
```

```
  output_self_attention : True
  use_position_embedding: True
# classification head
classifier:
  input_dim: None
  hidden_dim: 24
  output_dropout: 0.5
  num_classes: 2
  use_patient_info: False
  patient_embedding_dim: None
```

Appendix B

Diabetic retinopathy prediction: Additional materials

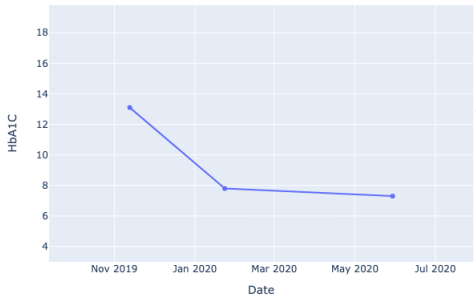
B.1 Rule-based algorithm for retinopathy target definition

We define the presence or absence of retinopathy based on different features related to records filed either by the ophthalmologist during visits or diabetologist during hospitalizations. Particularly, we label the presence of retinopathy based on five checks in the following order :

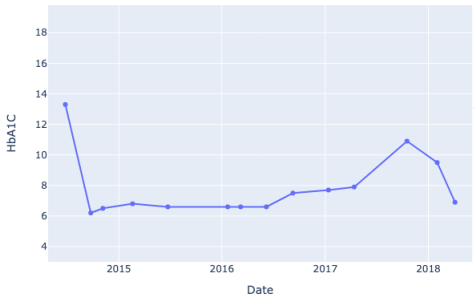
1. The fundus result shows retinopathy.
2. The patient has been administered one of the following retinopathy treatments: laser, pan photocoagulation, or vitrectomy.
3. The patient had a complication of retinopathy (intravitreal hemorrhage, neovessels).
4. The presence of fundus abnormalities.

Generally, the ophthalmologist fills in the date of onset of retinopathy in the corresponding reports. In other cases, the diabetologist manually fills in the information when it is available.

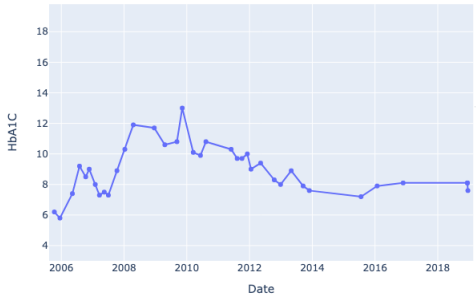
B.2 HbA1c profiles in CaRéDIAB



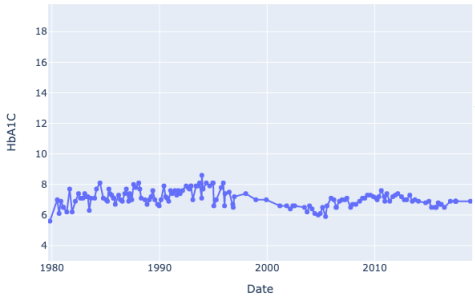
(a) Profile with 3 records



(b) Profile with 14 records



(c) Profile with 35 records



(d) Profile with 151 records

Figure B.1: The history profile of HbA1c records for four diabetic patients: we note that the number of records, the HbA1c levels, and the duration of follow-up are highly variable between these patients. Patient (a) have a short duration of follow-up. On the other hand, the patient (d) has good glycemic control throughout 40 years of follow-up. We note that good glycemic control is defined by HbA1c level $\leq 7\%$.

B.3 Hyper parameter fine-tuning results for Diabetic retinopathy prediction

This appendix presents the search space of the hyper-parameters required for the model training, the patient representation module, the time representation module, the event representation module, the sequential temporal module, and the classification module. The results of the Bayesian optimization for the 12 models tested on patients of "$group \geq 3$" are reported in our online appendix: https://github.com/sararb/Temporal-Deep-Learning-for-Medical-time series/blob/main/Appendix-A.md

Table B.1: The search space of hyper-parameters for the Bayesian optimization

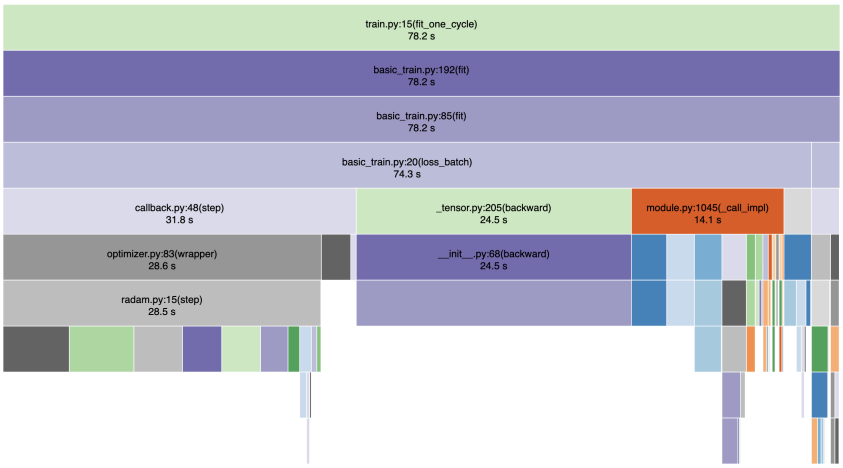
Module	**Hyper-parameter**	**Search space**	**distribution**
Training	cycle_len	[5, 40]	discrete with step size of 5
Training	batch_size	[4, 32]	discrete with step size of 4
Training	class_weight	[0.1, 0.7]	discrete uniform with step size of 0.05
Training	optimizer	[radam, adam, sgd]	categorical choice
Training	max_lr	[1e-4, 1e-2]	logarithmic uniform
Training	weight_decay	[1e-6, 1e-4]	logarithmic uniform
patient_model	hidden_dims	[4, 32]	discrete with step size of 4
patient_model	activation	[tanh, relu, gelu]	categorical choice
patient_model	output_dim	[8, 64]	discrete with step size of 8
time_model	hidden_dims	[4, 32]	discrete with step size of 4
time_model	projection_size	[4, 32]	discrete with step size of 4
time_model	output_dim	[8, 64]	discrete with step size of 8
event_model	continuous_hidden_dims	[4, 32]	discrete with step size of 4
event_model	tf_activation	[tanh, relu, gelu]	categorical choice
event_model	continuous_output_dim	[8, 64]	discrete with step size of 8
temporal_model	hidden_size	[8, 64]	discrete with step size of 8
temporal_model	num_layers	[1, 10]	discrete with step size of 1
temporal_model	dropout	[0, 0.5]	discrete uniform with step size of 0.1
temporal_model	attn_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
temporal_model	hidden_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
temporal_model	feed_forward_hidden	[8, 64]	discrete with step size of 8
temporal_model	hidden_act	[tanh, relu, gelu]	categorical choice
temporal_model	timedecay_size	[1, 5]	discrete with step size of 1
classifier_model	hidden_dim	[8, 64]	discrete with step size of 8
classifier_model	output_dropout	[0, 0.5]	discrete uniform with step size of 0.1

B.4 Table of retinopathy prediction results for group of patients with at least three records

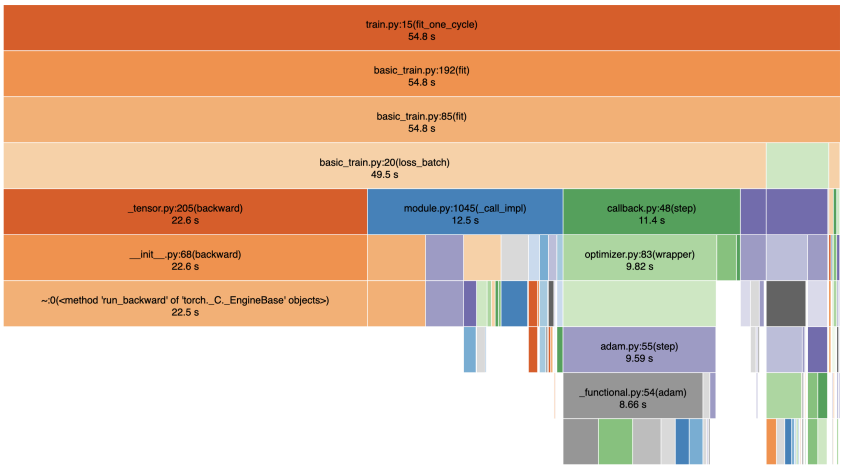
Table B.2: Test results: Mean and std of AUROC using 5 cross-validation sets of 121 patients.

Model	**Time representation**	**AUROC - mean**	**AUROC - std**
attention	no_time	0.7105	0.0969
attention	time_concat_mlp	0.8358	0.02286
attention	time_concat_soft	**0.8399**	0.0353
attention	time_mask	0.7679	0.0715
attention	time_encode	0.8336	0.0143
BiLSTM	no_time	0.8187	0.0254
BiLSTM	time_concat_mlp	0.8363	0.0199
BiLSTM	time_concat_soft	0.8350	0.03152
BiLSTM	time_mask	0.8097	0.0385
clstm	forget	0.7413	0.0852
clstm	output	0.6362	0.01682
clstm	forget_output	0.6308	0.02146

B.5 Profiles of the execution of the model training loop including data-loading, and optimization steps



(a) Execution time of BiLSTM model



(b) Execution time of Self-Attention model

Appendix C

MAP: AL-empowered medical annotator interface

C.1 User Interface standards

We aimed to design a user-interface Web-App for efficiently collecting manual annotations from clinical text. To achieve that, our interface should meet the following requirements :

1. Including multi-project and multi-user management for efficiently gathering annotations from different annotators and taking advantage of previous projects.
2. Optimizing the visualisation of text information to help the medical expert on focusing on relevant information related to the annotation sub-task objective.
3. Querying and selecting the set of samples to annotate based on active learning strategy through a Python API.
4. Optimizing the visualization of automatic annotations for medical expert validation.
5. Including metrics for evaluating the annotation cost of each annotator.

6. Visualizing the evolution of performances and saving trained supervised methods in the system.

C.2 Data model

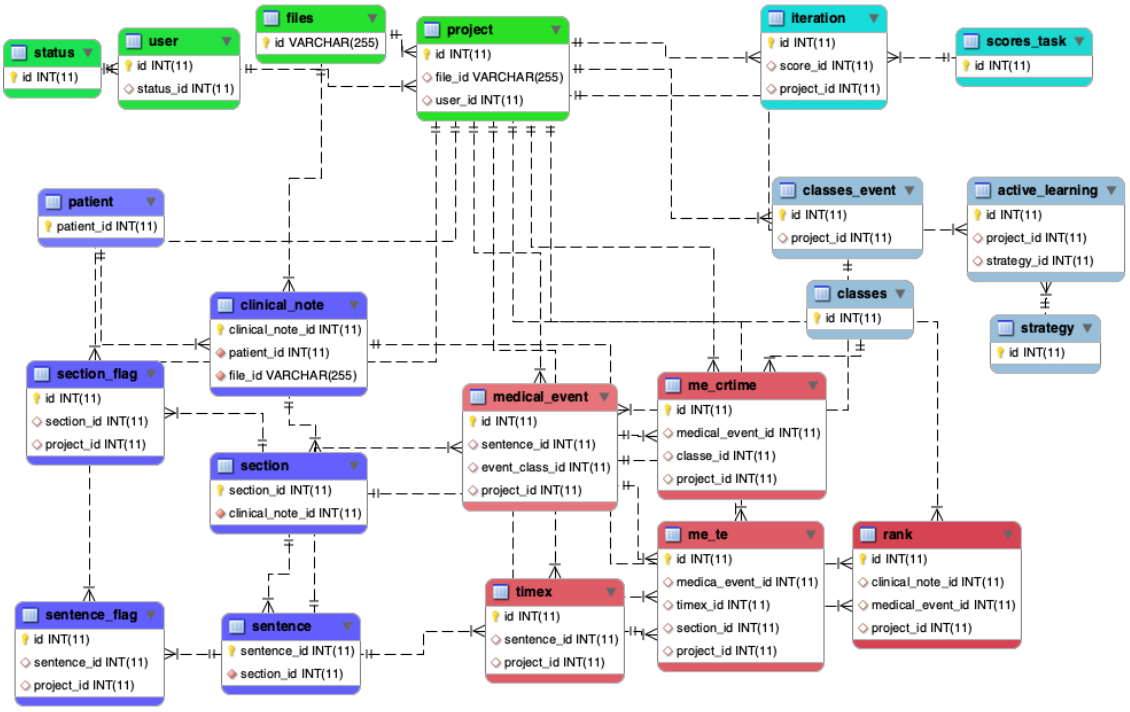


Figure C.1: EER diagram of MAP tool

We defined the data model, detailed in figure C.1, to ensure all the standards detailed in section C.1. The resulting database model includes five main classes:

1. **System tables**: user, status, project, files. These tables handle the creation and management of different users and projects and the storage of data files loaded by the medical expert.

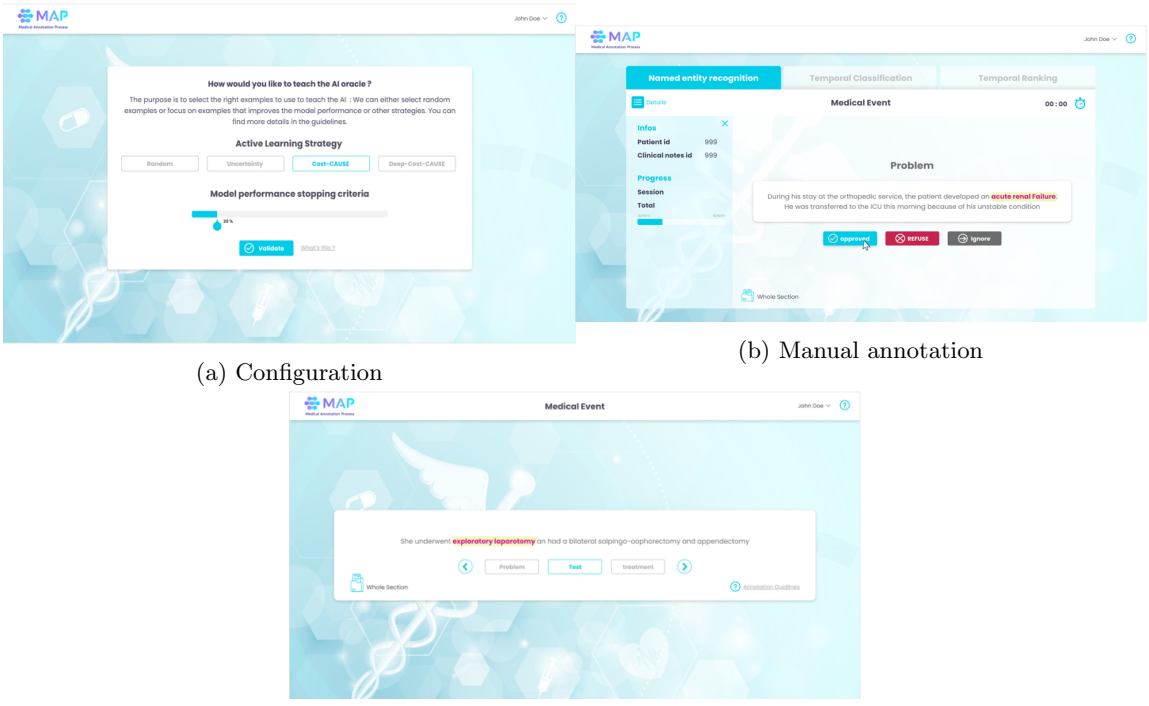
2. **Data tables**: patient, clinical_note, section and sentence. These tables store the raw text data of medical reports and patient ids from the file loaded by the expert. It also stores the medical sections and sentences automatically derived from the clinical note text.

3. **Annotation tables**: medical_event, timex, me_crtime, me_te and rank. These tables store manual and automatic annotations extracted from the text for the 5 sub-tasks.

4. **Automatic annotation configurations**: classes_event, classes, active_learning, strategy. These tables store the configuration of entity classes and the active learning stategy to use for selecting the samples to be annotated.

5. **Performance tables**: iteration, scores_task. The two tables store the performance of the supervised model at each AL-iteration.

C.3 Interface design



(a) Configuration

(b) Manual annotation

(c) Automatic annotation validation

Figure C.2: Active annotation process of the sub-module "Medical Event Extraction"

- We divided the annotation process into three steps:

1. **Configuration**: The expert select the sampling strategy and the performance threshold for exporting annotated data and trained model.

2. **Manual annotation**: The expert manually annotates the visualized text with respect to the task objective. For example, in Figure C.2b, the expert highlights the part of text corresponding to a medical event and selects its class.

3. **Automatic annotation**: the annotations are automatically computed from the supervised model and are visualized for expert validation.

Appendix D

Hyperparameters for variants of HiTT architecture

D.1 The search space of HiTT architecture's parameters

Table D.1: The search space of hyper-parameters for the Bayesian optimization of HiTT architectures

Module	Hyper-parameter	Search space	distribution
Training	cycle_len	[5, 40]	discrete with step size of 5
Training	batch_size	[4, 32]	discrete with step size of 4
Training	class_weight	[0.1, 0.7]	discrete uniform with step size of 0.05
Training	optimizer	[radam, adam, sgd]	categorical choice
Training	max_lr	[1e-4, 1e-2]	logarithmic uniform
Training	weight_decay	[1e-6, 1e-4]	logarithmic uniform
`VisitEncoder`	procedures_embed_dim	[4, 48]	discrete with step size of 4
`VisitEncoder`	procedures_attention_hidden	[8, 64]	discrete with step size of 8
`VisitEncoder`	diagnoses_embed_dim	[4, 48]	discrete with step size of 4
`VisitEncoder`	diagnoses_attention_hidden	[8, 64]	discrete with step size of 8
`VisitEncoder`	soft_projection_size	[4, 32]	discrete with step size of 4
`VisitEncoder`	continuous_embed_dim	[4, 48]	discrete with step size of 4
`VisitEncoder`	activation	[tanh, relu, gelu]	categorical choice
`VisitEncoder`	transformer_hidden_size	[8, 64]	discrete with step size of 8
`VisitEncoder`	transformer_n_layers	[1, 10]	discrete with step size of 1
`VisitEncoder`	attn_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
`VisitEncoder`	hidden_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
`Time-aware Transformer`	transformer_hidden_size	[8, 64]	discrete with step size of 8
`Time-aware Transformer`	transformer_n_layers	[1, 10]	discrete with step size of 1
`Time-aware Transformer`	attn_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
`Time-aware Transformer`	hidden_dropout_prob	[0, 0.5]	discrete uniform with step size of 0.1
classifier_model	hidden_dim	[8, 64]	discrete with step size of 8
classifier_model	output_dropout	[0, 0.5]	discrete uniform with step size of 0.1

D.2 The best hyper-parameters of HiTT architectures

Table D.2: Best hyper-parameters results of HiTT architectures

Module	Hyper-parameter	HiTT	HiTT + text	HiTT - continuous	HiTT - diagnoses	HiTT - procedures	HiTT - Time
Training	cycle_len	30	40	20	25	15	25
Training	batch_size	8	4	32	8	16	32
Training	class_weight	0.65	0.55	0.55	0.6	0.5	0.45
Training	optimizer	adam	adam	adam	radam	adam	sgd
Training	max_lr	0.000323	0.000135	0.000523	0.000897	0.000214	0.000534
Training	weight_decay	1.70E-06	3.45E-06	2.37E-06	1.23E-06	1.93E-05	2.23E-05
`VisitEncoder`	procedures_embed_dim	16	12	24	16	-	24
`VisitEncoder`	procedures_attention_hidden	56	56	24	16	-	48
`VisitEncoder`	diagnoses_embed_dim	32	24	48	-	16	32
`VisitEncoder`	diagnoses_attention_hidden	32	48	24	-	16	32
`VisitEncoder`	soft_projection_size	8	12	-	4	8	4
`VisitEncoder`	continuous_embed_dim	20	32	-	16	20	24
`VisitEncoder`	activation	gelu	gelu	-	gelu	tanh	relu
`VisitEncoder`	transformer_hidden_size	56	56	-	32	48	64
`VisitEncoder`	transformer_n_layers	3	4	-	3	2	2
`VisitEncoder`	attn_dropout_prob	0.2	0.2	-	0.4	0.2	0.1
`VisitEncoder`	hidden_dropout_prob	0.1	0.3	-	0	0.1	0
`Time-aware Transformer`	transformer_hidden_size	56	56	32	48	56	32
`Time-aware Transformer`	transformer_n_layers	3	5	3	2	4	4
`Time-aware Transformer`	attn_dropout_prob	0.2	0	0.2	0.3	0.1	0.3
`Time-aware Transformer`	hidden_dropout_prob	0.1	0.2	0.2	0.3	0.1	0.2
classifier_model	hidden_dim	16	24	16	24	24	32
classifier_model	output_dropout	0.4	0.5	0.2	0.4	0.3	0.2

Appendix E

Review of deep learning methods for medical time series modeling

Table E.1: An overview of the DL-based models used for medical time series applications

Year	Core-architecture	Task	ML task	Datasets	Reference
2021	Transformer + Attention	in-hospital mortality	Next event prediction	MIMIC III PhysioNet	[79]
2021	RNN + Time	clinical event time series	next-event prediction	MIMIC-III	[50]
2020	RNN + Attention	in- hospital mortality length of hospital stay readmission	Binary Classification + regression	MIMIC-III	[22]
2020	RNN + Time	in-hospital mortality	Binary Classification	MIMIC- III eICU	[72]
2020	RNNs + Time	in-hospital mortality	Binary Classification	MIMIC-III CINC2012 CINC2019 COVID-19	[67]
2020	Transformer + LSTM + Time	Unplanned Re-admission In-Hopital Mortality	Binary Classification	MIMIC	[160]
2019	RNN	predict future medication procedures/lab tests physiological signals	Next event prediction	MIMIC	[85]
2019	CNN + Hierarchial attention	myotonic dystrohpy diagnosis	Binary Classification	Private data	[161]
2019	RNN + Hierarchial attention	Mortality prediction ICU admission prediction	Binary Classification	MIMIC-III	[76]
2019	RNN + Time	in-hospital mortality length of hospital stay	Binary Classification + regression	MIMIC-III	[84]
2019	RNN	ICU mortality risk	Binary Classification	MIMIC III	[78]
2018	RNN + Attention	In-Hospitality Mortality Readmission Rate Length of Stay	Binary Classification + regression	Private EHR datamarts	[23]
2018	RNN + Time	In Hospital Mortality	Binary Classification	MIMIC III + PhysioNet	[73]
2018	Transformer + Time	In Hospital Mortality	Binary Classification	MIMIC-III	[33]
2018	RNN + Hierarchial attention	Future Hospitalization	Binary Classification	Private EHR data	[81]

Year	Core-architecture	Task	ML task	Datasets	Reference
2018	RNN + Attention + Time	Heart Failure onset	Binary Classification	SNOW + EMRbots	[75]
2018	RNN	Prediction of Cystic Fibrosis	Binary Classification	Private EHR data	[162]
2018	RNN + Time	In Hospital Mortality	Binary Classification	MIMIC III + PhysioNet	[73]
2018	Transformer + Time	In Hospital Mortality	Binary Classification	MIMIC-III	[33]
2017	LSTM	Unplanned Re-admission Disease Progression	Binary Classification	Private EHR data	[53]
2017	RNN + Attention	Treatment Recommendation	seq2seq prediction	MIMIC + Sutter	[86]
2017	Graph + Attention	Heart Failure onset	Binary Classification	MIMIC + Sutter	[77]
2017	SkipGram + RNN	Heart Failure onset	Binary Classification	Sutter	[17]
2017	RNN + Time	Diabetes Mellitus prediction	Binary Classification	Synthetic data + PPMI	[74]
2017	RNN + Hierarchial attention	Category of next visit diagnoses	multi-label classification	Medicaid + Diabetes claim	[82]
2017	LSTM + T-SNE	Patient Subtyping	Clustering	PPMI	[21]
2017	RNN	Unplanned Re-admission	Binary Classification	Private EHR data	[80]
2016	RNN + Hierarchial attention	Next diagnoses	Next event prediction	MIMIC-III	[12]
2016	CNN	Congestive Heart Failure (CHF) Chronic Obstructive Pulmonary Disease (COPD)	Binary Classification	Private EHR data	. [16]
2016	RNN	Medication Prescriptions	multi-label classification	Sutter	[19]
2016	RNN	Medication Prescriptions	multi-label classification	Private EHR data	[163]
2016	Denoising Auto Encoder	Disease classification	multi-label classification		[11]
2016	RNN	3 clinical outcomes of kidney transplantation	multi-label classification	Private EHR data	[83]

Abbreviations

AL Active Learning

ALC Area under Learning Curve

ANN Artificial Neural Networks

AUROC Area under ROC curve

BERT Bidirectional Encoder Representations from Transformers

BiLSTM Bidirectional Long Short Term Memory

C-LSTM Contextual Long Short Term Memory

CAUSE Clustering And Uncertainty Sampling Engine

CNIL Clustering And Uncertainty Sampling Engine

CNN Convolutional Neural Network

CRF Conditional Random Fields

CV Cross Validation

DL Deep Learning

DR Diabetic Retinopathy

EER Enhanced entity-relationship

EHR Electronic Health Records

ETDRS Early Treatment Diabetic Retinopathy Study

GRU Gated Recurrent Unit

GRU-D Gated Recurrent Unit with Decay

HAI Health Acquired Infection

HBA1c Hemoglobin A1C

HiTT Hierarchical Time-aware Transformer

HWUS Hierarchical Time-aware Transformer

ICD-9 International Classification of Diseases, Ninth Revision

ICU Intensive Care Unit

IHM In Hospital Mortality

IMTS Irregular Medical Time Series

LDA Latent Dirichlet Allocation

LOS Length Of Stay

LR Logistic Regression

LSTM Long Short Term Memory

MAP Medical Annotation Process

MIMIC-III Medical Information Mart for Intensive Care

MLM Masked Language Modeling

MLP Multi Layer Preceptron

NER Named Entity Recognition

NLP Natural Language Processing

RNN Recurrent Neural Network

RS Random Sampling

SOTA State Of The Art

SVM Suport Vector Machine

T-GRU Time-aware Gated Recurrent Unit

T1D Type-1 Diabetes

ECOLE DOCTORALE

Titre : Méthode multimodale optimisée basée sur l'apprentissage profond pour les données médicales temporelles irrégulières

Mots clés : Analyses statistiques, Données médicales, Réseaux de Neurones, Apprentissage statistique, Traitement de langage naturel, Apprentissage actif

Résumé : L'adoption des dossiers médicaux électroniques (DME) dans les systèmes d'information des hôpitaux a conduit à la définition de bases de données Big Data regroupant divers types de données telles que des notes cliniques textuelles, des événements médicaux longitudinaux et des informations tabulaires sur les patients. Toutefois, les données ne sont renseignées que lors des consultations médicales ou des séjours hospitaliers. La fréquence de ces visites varie selon l'état de santé du patient et des habitudes locales. Ainsi, un système capable d'exploiter les différents types de données collectées à différentes échelles de temps est essentiel pour reconstruire la trajectoire de soin du patient, analyser son historique et, par conséquent, délivrer des soins plus adaptés. Ce travail de thèse aborde deux défis principaux du traitement des données médicales : (1) Représenter la séquence des observations médicales à échantillonnage irrégulier et (2) optimiser l'extraction des événements médicaux à partir des textes de notes cliniques. Notre objectif principal est de concevoir une représentation multimodale de la trajectoire de soin du patient afin de résoudre les problèmes de prédiction clinique. Notre premier travail porte sur la modélisation des séries temporelles médicales irrégulières afin d'évaluer l'importance de considérer les écarts de temps entre les visites médicales dans la représentation de la trajectoire de soin d'un patient donné. À cette fin, nous avons mené une étude comparative entre les réseaux de neurones récurrents, les modèles basés sur l'architecture "Transformer" et les techniques de représentation du temps. De plus, l'objectif clinique était de prédire les complications de la rétinopathie chez les patients diabétiques de type 1 de la base de données française CaRéDIAB (Champagne Ardenne Réseau Diabète) en utilisant leur historique de mesures HbA1c. Les résultats de l'étude ont montré que le modèle "Transformer", combiné à la représentation 'Soft-One-Hot' des écarts temporels a conduit à un score AUROC de 88,65% (spécificité de 85,56%, sensibilité de 83,33%), soit une amélioration de 4,3% par rapport au modèle basé sur l'architecture "LSTM". Motivés par ces résultats, nous avons étendu notre étude à des séries temporelles multivariées plus courtes et avons prédit le risque de mortalité à l'hôpital pour les patients admis en soins intensifs présents dans la base de données MIMIC-III. L'architecture proposée, Hierarchical Time-aware Transformer (HiTT), a amélioré le score AUC de 5% par rapport à l'architecture de base "Transformer". Dans la deuxième étape, nous nous sommes intéressés à l'extraction d'informations médicales pertinentes à partir des comptes rendus médicaux afin d'enrichir la trajectoire de soin du patient. En particulier, les réseaux de neurones basés sur le module "Transformer" ont montré des résultats encourageants dans l'application d'extraction d'informations médicales. Cependant, ces modèles complexes nécessitent souvent un grand corpus annoté. Cette exigence est difficile à atteindre dans le domaine médical car elle nécessite l'accès à des données privées de patients et des annotateurs experts. Pour réduire les coûts d'annotation, nous avons exploré les stratégies d'apprentissage actif qui se sont avérées efficaces dans de nombreuses tâches, notamment la classification de textes, l'analyse d'image et la reconnaissance vocale. En plus des méthodes existantes, nous avons défini une stratégie d'apprentissage actif, nommée Hybrid Weighted Uncertainty Sampling (HWUS), qui utilise la représentation cachée du texte donnée par le modèle "Transformer" pour mesurer la représentativité des échantillons. Une simulation utilisant l'ensemble de données du challenge i2b2-2010 a montré que la métrique proposée réduit le coût d'annotation de 70% pour atteindre le même score de performance que l'apprentissage supervisé passif. Enfin, nous avons combiné des séries temporelles médicales multivariées et des concepts médicaux extraits des notes cliniques de la base de données MIMIC-III pour entraîner une architecture multimodale. Les résultats du test ont montré une amélioration de 5,3% en considérant des informations textuelles supplémentaires.

Title : Optimized Deep Learning-Based Multimodal method for Irregular Medical Timestamped data

Keywords : Statistical analysis, Deep learning, Medical data, Active learning, Time series, Natural language processing

Abstract : The wide adoption of Electronic Health Records in hospitals' information systems has led to the definition of large databases grouping various types of data such as textual notes, longitudinal medical events, and tabular patient information. However, the records are only filled during consultations or hospital stays that depend on the patient's state, and local habits. A system that can leverage the different types of data collected at different time scales is critical for reconstructing the patient's health trajectory, analyzing his history, and consequently delivering more adapted care. This thesis work addresses two main challenges of medical data processing: learning to represent the sequence of medical observations with irregular elapsed time between consecutive visits and optimizing the extraction of medical events from clinical notes. Our main goal is to design a multimodal representation of the patient's health trajectory to solve clinical prediction problems. Our first work built a framework for modeling irregular medical time series to evaluate the importance of considering the time gaps between medical episodes when representing a patient's health trajectory. To that end, we conducted a comparative study of sequential neural networks and irregular time representation techniques. The clinical objective was to predict retinopathy complications for type 1 diabetes patients in the french database CaRéDIAB (Champagne Ardenne Réseau Diabetes) using their history of HbA1c measurements. The study results showed that the attention-based model combined with the soft one-hot representation of time gaps led to AUROC score of 88.65% (specificity of 85.56%, sensitivity of 83.33%), an improvement of 4.3% when compared to the LSTM-based model. Motivated by these results, we extended our framework to shorter multivariate time series and predicted in-hospital mortality for critical care patients of the MIMIC-III dataset. The proposed architecture, HiTT, improved the AUC score by 5% over the Transformer baseline. In the second step, we focused on extracting relevant medical information from clinical notes to enrich the patient's health trajectories. Particularly, Transformer-based architectures showed encouraging results in medical information extraction tasks. However, these complex models require a large, annotated corpus. This requirement is hard to achieve in the medical field as it necessitates access to private patient data and high expert annotators. To reduce annotation cost, we explored active learning strategies that have been shown to be effective in tasks such as text classification, information extraction, and speech recognition. In addition to existing methods, we defined a Hybrid Weighted Uncertainty Sampling active learning strategy that takes advantage of the contextual embeddings learned by the Transformer-based approach to measuring the representativeness of samples. A simulated study using the i2b2-2010 challenge dataset showed that our proposed metric reduces the annotation cost by 70% to achieve the same score as passive learning. Lastly, we combined multivariate medical time series and medical concepts extracted from clinical notes of the MIMIC-III database to train a multimodal transformer-based architecture. The test results of the in-hospital mortality task showed an improvement of 5.3% when considering additional text data. This thesis contributes to patient health trajectory representation by alleviating the burden of episodic medical records and the manual annotation of free-text notes.

Institut Polytechnique de Paris
91120 Palaiseau, France