PUCRS

ESCOLA POLITÉCNICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

BERNARDO SCAPINI CONSOLI

# HOLISTIC PATIENT REPRESENTATION LEARNING AND AUTOMATIC ANNOTATION OF ELECTRONIC HEALTH RECORDS

Porto Alegre
2025

PÓS-GRADUAÇÃO - *STRICTO SENSU*


Pontifícia Universidade Católica
do Rio Grande do Sul

**PONTIFICAL CATHOLIC UNIVERSITY OF RIO GRANDE DO SUL**
**SCHOOL OF TECHNOLOGY**
**COMPUTER SCIENCE GRADUATE PROGRAM**

# HOLISTIC PATIENT REPRESENTATION LEARNING AND AUTOMATIC ANNOTATION OF ELECTRONIC HEALTH RECORDS

**BERNARDO SCAPINI CONSOLI**

Doctoral Thesis submitted to the Pontifical Catholic University of Rio Grande do Sul in partial fulfillment of the requirements for the degree of Ph. D. in Computer Science.

Advisor: Prof. Isabel Harb Manssour

**Porto Alegre**
**2025**

# Ficha Catalográfica

C755h Consoli, Bernardo Scapini

Holistic patient representation learning and automatic annotation of electronic health records / Bernardo Scapini Consoli. – 2025.
97 f.
Tese (Doutorado) – Programa de Pós-Graduação em Ciência da Computação, PUCRS.

Orientadora: Profa. Dra. Isabel Harb Manssour.

1. aprendizado de representação de pacientes. 2. registros eletrônicos de saúde. 3. BRATECA. 4. fluxo de pacientes. 5. anotação automática. I. Manssour, Isabel Harb. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da PUCRS
com os dados fornecidos pelo(a) autor(a).
Bibliotecária responsável: Clarissa Jesinska Selbach CRB-10/2051

**BERNARDO SCAPINI CONSOLI**

# HOLISTIC PATIENT REPRESENTATION LEARNING AND AUTOMATIC ANNOTATION OF ELECTRONIC HEALTH RECORDS

This Doctoral Thesis has been submitted in partial fulfillment of the requirements for the degree of Ph. D. in Computer Science, of the Computer Science Graduate Program, School of Technology of the Pontifical Catholic University of Rio Grande do Sul

Sanctioned on March 26th, 2025.

## COMMITTEE MEMBERS:

Prof. Dr. Marcio Sarroglia Pinho (PPGCC/PUCRS)

Prof. Dr[a]. Mariana Recamonde Mendoza (INF/UFRGS)

Prof[a]. Dr[a]. Aline Marins Paes Carvalho (IC/UFF)

Prof. Isabel Harb Manssour (PPGCC/PUCRS - Advisor)

## ACKNOWLEDGMENTS

Agradeço o Dr. Rafael Heitor Bordini e a Dra. Renata Vieira pelo seu apoio contínuo durante toda esse jornada. Agradeço também a Dra. Ying Ding por todo seu apoio durante minha estadia na Universidade do Texas em Austin, e a Dra. Isabel Harb Manssour por seu apoio como orientadora durante o último ano do meu doutorado.

Esta pesquisa foi apoiada pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex.

Esta pesquisa também foi apoiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) Brasil, através da chamada 25/2020.

# APRENDIZADO DE REPRESENTAÇÕES HOLÍSTICAS DE PACIENTES E ANOTAÇÃO AUTOMÁTICA DE REGISTROS ELETRÔNICOS DE SAÚDE


## RESUMO

Aprendizado de representações de pacientes é o uso de inteligência artificial para reinterpretar dados conhecidos de pacientes, extraídos de Registros Eletrônicos de Saúde, para que modelos de aprendizado de máquinas consigam fazer previsões sobre pacientes que possam ajudar profissionais médicos no diagnóstico e na administração de cuidados adequados. É importante observar que dados médicos estão vinculados ao seu local de origem. Para lidar com este aspecto vital para o desenvolvimento das tecnologias nacionais de medicina computacional, desenvolvemos o BRATECA, uma coleção de dados de hospitais terciários brasileiros. Esta coleção for aberta para acesso credenciado e era o maior banco de dados hospitalares Brasileiros quando foi lançada. Utilizando ela em tarefas de fluxo de pacientes, atingimos resultados de até 0,88 de F1 para a tarefa de Predição de Admissão de pacientes e de até 0,84 de F1 para Predição de Estadia Longa de pacientes. Desenvolvemos também a arquitetura de anotação automática SDoH-GPT, a qual validamos nos banco de dados de UTI estadunidenses MIMIC-III e atingimos correlação medida em mais de 0,8 pontos no kappa de Cohen para todas categorias entre nossas anotações automáticas e anotações humanas.



**Palavras-Chave:** aprendizado de representações de pacientes, registros eletrônicos de saúde, BRATECA, fluxo de pacientes, anotação automática, aprendizado profundo, modelos preditivos, tempo de estadia, internet das coisas, dados médicos, histórico clínico de pacientes, dados heterogêneos.

# HOLISTIC PATIENT REPRESENTATION LEARNING AND AUTOMATIC ANNOTATION OF ELECTRONIC HEALTH RECORDS


## ABSTRACT

Patient representation learning is the use of artificial intelligence technologies to reinterpret known patient data, extracted from Electronic Health Records, in a way that allows machine learning models to predict data and outcomes that could help medical professionals in diagnosis and the administration of proper care. It is important to note that medical data is tied to its place of origin. To deal with such a vital aspect to the development of national computational medicine solutions, we developed BRATECA, a collection of Brazilian tertiary care hospital data. This collection is open for credentialed access and was the largest collection of Brazilian medical data at the time of its release. Utilizing this collection in patient flow tasks, we achieved results of up to 0.88 F1 in patient Admission Prediction and up to 0.84 F1 for patient Extended Stay Prediction. We also developed an architecture for automatic annotation of social determinants of health in Electronic Health Records, which was validated on the US intensive care data collection MIMIC-III, where we achieved correlations of more than 0.8 measured in Cohen's kappa for all annotation categories between our automatic annotation and human annotations.



**Keywords:** patient representation learning, electronic health records, BRATECA, patient flow, automate annotation, deep learning, predictive models, length-of-stay, internet of things, medical data, patient clinical history, heterogeneous data.

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ACRONYMS

AI – Artificial Intelligence
AUPRC – Area Under the Precision-Recall Curve
AUROC – Area Under the Receiver Operating Characteristic
BRATECA – Brazilian Tertiary Care Dataset
CBOW – Continuous-Bag-Of-Words
CNN – Convolutional Neural Networks
CSV – Comma Separated Values
CXR – Chest X-Ray
ED – Emergency Department
EHR – Electronic Health Record
FCNN – Fully-Connected Deep Neural Networks
GRU – Gated Recurrent Unit
ICU – Intensive Care Unit
KB – Knowledge Base
LLM – Large Language Model
LOS – Length-of-Stay
LSTM – Long-Short Term Memory
MIMIC – Medical Information Mart for Intensive Care
NHS – National Health Service
NLP – Natural Language Processing
NN – Neural Network
PRL – Patient Representation Learning
RNN – Recurrent Neural Networks
SBDH – Social and Behavioral Determinants of Health
SDOH – Social Determinants of Health
SUS – Sistema Único de Saúde

# CONTENTS

# 1. INTRODUCTION

Decision making in healthcare settings has been a topic of growing interest in the field of artificial intelligence [87, 88, 61]. Studies on methods for predicting disease [92, 101], mortality [101], length-of-stay [92, 101], admission [59], and interventions [96] have become more common with the adoption of Electronic Health Records (EHR) in hospitals around the world, which in turn has led to efforts to deidentify this information and make it available for use in related research [87].

These EHRs contain a variety of information used by medical professionals to form a holistic understanding of patients that can enable proper diagnosis and treatment, such as patient demographics, vital signs, laboratory data, medications, admission and discharge information, clinical notes on patient status and progress, and more. Some of these data categories are structured (e.g. diagnosis codes, laboratory and vital sign readings), but others are unstructured (e.g. clinical notes, medical images) [88]. In other words, the data show heterogeneous characteristics that make them difficult to use in a single learning model, since all these different inputs need to be homogenized for use with neural networks, which are central to predictive model architectures [55]. This heterogeneity has led many studies to use only part of the available data to train predictive models, usually only structured data or only data that present unstructured free text such as clinical notes [87, 88, 61].

Patient Representation Learning (PRL) is the name for the many techniques used to unify medical data into useful mathematical representations [87]. These patient representations are vector spaces derived from the data present in EHRs, transformed into a format more easily processed by machine learning algorithms. Such representations have been used in the literature in common tasks for computational medicine mentioned previously, such as length-of-stay prediction, medical intervention prediction, mortality prediction, and disease diagnosis prediction [87]. The most common tasks require classification and regression to perform outcome prediction from ML models, for which researchers have used architectures such as Random Forests or many Neural Networks (NN) frameworks [67].

Still, these studies tend to use only a limited subset of the data from what they have available. This is often not due to the fact that more data would harm the predictive models in question, but rather due to the heterogeneity of the data, which makes it difficult to create models that take into account all the information to represent the complete health status of a patient [88]. While only a few data points may be sufficient for simple diagnoses, it is reasonable to assume that the less complete the patient information used in the prediction, the more likely it is that the prediction will not be optimal.

This, among other factors already mentioned, has led researchers to consider that patient representation methods that only use some of the available data are not ideal and to

postulate that a more holistic view of patients, inclusive of all collected data, would support better predictions that on the surface do not seem to require certain data points [88]. This is due to human health's inherently complex and interconnected nature, where seemingly disconnected variables may have some obscure cause-effect relationship [88].

The field of patient representation is quite expansive [87], so we must set several boundaries, such as the data that will be used, the tasks that will be the focus of the work, the architecture that will be developed and used to train models, and the methods by which the results will be evaluated and examined.

Our main objective with this work was to develop solutions for Brazilian hospitals, and, as such, we could not rely on the MIMIC collection, the most commonly used dataset for these tasks, as it reflects the clinical realities of the United States of America rather than Brazil. Because of this and the dearth of national data available at the outset of this project, a major objective for this work was the creation and distribution of our own Brazilian clinical data collection. So, alongside the Institute for Artificial Intelligence in Healthcare [1], we helped create a new data collection for Brazilian tertiary hospitals, which we called BRATECA [20]. This new dataset was used throughout this work and from which tasks were created and models were trained.

With BRATECA created, we could then decide which tasks to focus on. As BRATECA has less information than MIMIC, and notably lacks the minute-by-minute vital sign data that is widely used to achieve the best results for most tasks, we had to explore which tasks could feasibly be put together using only the available information without further annotation. We found that the most relevant tasks we could tackle immediately were Patient Flow tasks, specifically admission prediction and extended stay prediction. These specific tasks are also lower-risk among the many medical tasks that are usually tackled and, as such, are a good entry point for actual implementation into hospitals. Because of our lack of annotations, we also identified another side-objective for our work: examining the possibility of automating data annotation pipelines through LLMs to increase the number of possible tasks datasets enable researchers to tackle at lowered costs.

For architectures, we tested combinations of classic machine learning, NNs, and Large Language Models (LLM). For the patient flow tasks, the main objective for the architectures was to attempt to use as much information as possible from what was available in the BRATECA Collection to reach the best results possible. For the automated data annotation task, the main objective for the models was to approximate human annotation performance while also being quicker and cheaper to accomplish on the MIMIC-III dataset, which has been manually annotated for several tasks over the years it has been available in the hopes such techniques could be helpful for the BRATECA dataset in future.

Evaluations for patient flow were performed using the tasks of admission prediction and extended stay prediction, as previously mentioned. These tasks were chosen be-

[1] https://noharm.ai/en/

cause training and testing sets for them can be created using the BRATECA Collection. The evaluations for automated data annotation were performed using MIMIC-III and one of the annotations data sets that exist for MIMIC-III: MIMIC-SBDH [2], an annotated dataset for several categories of Social Determinants of Health (SDoH). MIMIC-SBDH was used as our annotation standard, which we compared our automated annotations against.

Given this introduction, we present two hypotheses that our work tackles:

1. holistic patient representations (i.e., representations that can use more patient data) significantly outperform limited patient representations for patient flow prediction tasks using Brazilian clinical data;
2. it is possible to leverage LLMs to create silver annotated datasets that are reliable, cheaper and faster to develop than relying entirely on human labor;

In this work, we have advanced this field of research in several ways:

1. we developed the first large collection of Brazilian tertiary care hospital data, the BRATECA Collection, composed of deidentified information from over 70 thousand patients and including over 3 million words of clinical note text, as wells as several other kinds of patient information;
2. we performed numerous tests for patient flow tasks such as admission prediction and extended stay prediction with several kinds of machine learning algorithms and showed that holistic use of data outperforms data cherry-picking while achieving good results for these tasks;
3. we performed automated annotation over the MIMIC dataset in collaboration with the Health AI Lab of the University of Texas at Austin as proof of concept that such techniques can work and to encourage future research efforts into annotation process for Brazilian data;

The rest of this work is organized as follows: Chapter 2 introduces the background and related work; Chapter 3 introduces the BRATECA resource; Chapter 4 introduces the test sets for patient flow and automated data annotation used to assess our machine learning architectures; Chapter 5 introduces the architectures we proposed for both patient flow and automated data annotation; All results are examined in Chapter 6; and Chapter 7 is about the conclusions, limitations and future work.

# 2. BACKGROUND AND RELATED WORK

PRL research must start by defining four main aspects: the data used to learn the representations; the tasks in which the representations are to be used; the architectures used to perform the representation learning; and the evaluation protocols for the architectures. Furthermore, to advance the field, it is also important to acknowledge the current challenges found in the literature. Thus, it is important to address each of these topics when investigating PRL. This chapter expounds upon each of them in order.

## 2.1 Data

Using Google Scholar, we performed a review of available hospital data for our use. Use used the following terms during our search: Medical Information; Hospital Data; Health Dataset; Clinical Patient Database; Brazilian Hospital Database; Brazilian Clinical Patient Data; Hospital Admission Data; Hospital Length-of-Stay dataset; Brazilian Clinical Length-of-Stay Data; and several other similar terms using combinations of the above terms.

All of the relevant works we identified are detailed in this section. We divided our findings into two categories: Foreign Data and Brazilian Data. Table 2.1 summarizes all data found to be related to our work:

### 2.1.1 Foreign Data

Si et al. (2021) [88] performed a systematic review of PRL literature wherein they identified several databases used throughout the literature. The most used dataset found in their review was the third version of the Medical Information Mart for Intensive Care (MIMIC-III) [50]. MIMIC, in its many versions, is the largest critical care dataset and among the ones that researchers can access with the most ease.

Its newest version, MIMIC-IV [49], was released in 2020 and is separated into six modules: core, hosp, icu, ed, cxr, and note. The *core* module comprises patient demographics, hospitalization records, and ward stay records. The *hosp* module is composed of data recorded during the patient's hospital stay, such as lab measurements, medication administration and prescription, billing information, etc. The *icu* module is composed of data taken from patients in intensive care units (ICUs), and include intravenous and fluid inputs, patient outputs, procedures, date and time information, etc. The *ed* module is composed of data from emergency department (ED) patients and includes reason for admission, triage

| Dataset | Origin | Release | Brief Description | Access Link |
|---|---|---|---|---|
| MIMIC-III | United States | 2016 | The most used hospital dataset by researchers, This is an ICU-centric dataset and includes information such as demographics, lab measurements, medication administration and prescription information, and vital sign information. | https://physionet.org/content/mimiciii/1.4/ |
| MIMIC-IV | United States | 2020 | Base on MIMIC-III, MIMIC-IV includes all of its data, more patients, and some images, such as chest X-rays. | https://physionet.org/content/mimiciv/3.1/ |
| eICU | United States | 2018 | A multi-center intensive care unit (ICU)database with high granularity data for over 200,000 admissions to ICUs monitored by eICU Programs across the United States. The database is deidentified, and includes vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, treatment information, and more. | https://eicu-crd.mit.edu/ |
| NHS Data Collections | United Kingdom | Variable* | A vast collection of data from across the United Kingdom about health and social care. The data is collected from all across the NHS system and includes datasets such as Healthcrae Operation dataflows, Emergency Cara data, maternity data and mental health services, among many others. | https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets |
| HiRID | Switzerland | 2022 | A freely accessible critical care dataset containing data from more than 33,000 patient admissions to the Department of Intensive Care Medicine, the University Hospital of Bern, Switzerland (Inselspital) from January 2008 to June 2016. It contains de-identified demographic nformation and a total of 712 routinely collected physiological variables, diagnostic test results, and treatment parameters. | https://physionet.org/content/hirid/1.1.1/ |
| SemClinBR | Brazil | 2020 | A corpus that has 1,000 clinical notes, labeled with 65,117 entities and 11,263 relations, and can support a variety of clinical NLP tasks and boost the EHR's secondary use for the Portuguese language. | https://github.com/HAILab-PUCPR/SemClinBr |
| OpenDataSUS | Brazil | 2020 | Several datasets about respiratory syndromes like the flu and COVID-19, hospital bed availability, births, and mortality. None of these provide unstructured data like clinical notes, focusing instead on providing large amounts of structured data. | https://opendatasus.saude.gov.br/dataset/ |
| COVID-19 Data Sharing/BR | Brazil | 2020/2021 | Several collections which provide structured information about COVID-19 cases. None of the datasets contain unstructured data. | https://repositorio datasharingfapesp.uspdigital.usp.br/handle/item/2 |
| BRAX | Brazil | 2022 | The dataset contains 24,959 chest radiography studies from patients presenting to a large general Brazilian hospital. A total of 40,967 images are available in the BRAX dataset. | https://physionet.org/content/brax/1.1.0/ |
| COVID Twitter Collection | Brazil | 2020 | A collection of 3,925,366 posts from Twitter and 18,413 online news gathered from the UOL web site regarding the online discussion on COVID-19 in Brazil. | https://data.mendeley.com/datasets/vhxdgjfjnk/3 |
| BRATECA | Brazil | 2022 | A Portuguese-language tertiary care data collection that contains 73,040 admission records of 52,973 unique adults (18 years of age or older) extracted from 10 hospitals located in two Brazilian states. | https://physionet.org/content/brateca/1.0/ |

Table 2.1: A table of all most prominent datasets available for free or credentialed usage. *The NHS collections are released by the Government of the UK separately, and have many publications and dates of release associated with them.

assessment, vital signs, etc. The *cxr* module contains chest x-ray (CXR) images from ED patients from multiple viewpoints. Finally, the *note* module contains patient's deidentified free-text clinical notes for hospitalization, although this module is not yet available to the public. MIMIC-IV is expected to almost completely replace its predecessor as the main dataset in use by the literature in the coming years, as was the case when MIMIC-III was originally released.

Another critical care dataset was also used in the reviewed literature, the eICU Collaborative Database [76]. While individual patient data is less extensive than what can be found in MIMIC, the eICU has more individual patient entries and represents the care given from several hospitals rather than just the one found in MIMIC.

The United Kingdom's National Health Service's (NHS) comprehensive dataset collection[1] offers more generalized data, less focused on critical care. The data is collected in order to support the analysis of specific policies of interest as well as the effects of particular policy initiatives, and it is separated into several different datasets, each with a different focus and different kinds of data.

Another freely available critical care database is the HiRID dataset [103], containing over 33,000 patient admission to the University Hospital of Bern in Switzerland from January 2008 to January 2016. It contains demographic information, diagnostic test results, treatment parameters, and 712 routinely collected physiological variables, many with records for every two minutes.

A more task-focused example of an English language clinical dataset can be found in the National NLP Clinical Challenges (n2c2) datasets. These challenges have been proposed since 2006, starting with the i2b2 project, n2c2's predecessor. These two series of challenges have presented datasets for a variety of tasks, such as deidentification, obesity prediction, coreference, temporal relations, heart disease, clinical semantic textual similarity, and family history extraction. The current edition, n2c2 2022[2], proposes three tracks: Contextualized Medication Event Extraction; Extracting Social Determinants of Health; and Progress Note Understanding: Assessment and Plan Reasoning. Task-specific datasets were released alongside each of these challenges, though some, such as the current challenge's third track, make use of already available resources (MIMIC-III in this case) when they are appropriate for the proposed task.

### 2.1.2 Brazilian Data

The previous data are English-language collections extracted from hospitals in certain anglophone countries and do not conform to the clinical realities of Brazil. It is thus important to gather national data for local research projects that may positively impact Brazilian public health. The development of national clinical resources has started in earnest in recent years, with work such as SemClinBR [35], a dataset with 1000 clinical notes annotated with over 65,000 entities and over 11,000 relations. The dataset was manually annotated and may be used for a variety of tasks, such as clinical named entity recognition

[1] https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets
[2] https://n2c2.dbmi.hms.harvard.edu/2022-challenge

and negation detection. It bears more resemblance to the n2c2 challenge datasets than to MIMIC.

OpenDataSUS [3] provides several datasets about respiratory syndromes like the flu and COVID-19, hospital bed availability, births, and mortality. None of these provide unstructured data like clinical notes, focusing instead on providing large amounts of structured data.

COVID-19 Data Sharing/BR [4] is a COVID-19 data sharing effort between Brazilian institutions. It contains quite a few datasets which provide structured information about COVID-19 cases. None of the datasets contain unstructured data.

BioBERTpt [86] is a fine-tuned BERT model trained on clinical EHR texts as well as texts from the biomedical literature. It has three versions, each trained with a different corpus. The first was trained with more than 2 million clinical notes from Brazilian hospitals collected between 2002 and 2018. The second with titles and abstracts from Portuguese biomedical scientific papers published in PubMed and Scielo. A third version combining both corpora into one was also trained. The clinical note corpus does not seem to have been made available after its use in training the models.

The literature also covers a Brazilian healthcare image dataset, the labeled chest X-ray dataset BRAX [81]. Although it is not a language resource, that dataset is nonetheless an example of a Brazilian healthcare dataset, and it is similar to MIMIC's CXR, except that the images are not complemented by text-based healthcare resources like MIMIC's.

Another example of a Portuguese-language health-related dataset was developed by de Melo et al. 2020 [30]. Their Twitter-based dataset comprises nearly 4 million tweets and about 18,000 news articles related to COVID-19 in Brazil. It has a different domain from the other datasets presented thus far and so has a different overall purpose, being more focused on public discourse and sentiment about public health issues rather than clinical information.

Both literature reviews identified that works that used public datasets represent a little under half of those examined, however. The rest used private datasets which are inaccessible to the rest of the community. This hinders proper methodology comparison efforts and makes result reproduction impossible.

Regardless of which database was used, data in PRL research are usually separated into two categories: structured data, which encompasses diagnosis codes, procedure codes, medications codes, etc; and unstructured, which encompasses free-text notes. Only about a third of the works reviewed used unstructured data at all, even when available, as is the case with the critical care databases. A fifth of all studies reviewed used both kinds of data, but even among those, not all available data was used, with only a subset of each modality being used at all (e.g., topic models used to represent unstructured data).

[3] https://opendatasus.saude.gov.br/

[4] https://repositoriodatasharingfapesp.uspdigital.usp.br/handle/item/2

## 2.2 State-of-the-Art Solutions

The literature points out several types of representation architectures that encompass both former and current state-of-the-art for the field. These are identified by Si et al. (2021) [88] in their literature review as follows: vector-based, sequence-based, graph-based, matrix-based, and tensor-based. Each of these represents a different way to process information and is used in many combinations throughout the literature.

Another systematic review, performed by Liu et al. (2022) [61], presents a classification system by technical paradigm, dividing them into the following categories: statistics learning-based methods; knowledge-based methods; and graph-based methods. All types of architectures presented by Si et al. (2021) are encompassed by one of these three categories.

Both systems of categorization will be explored in this section, as the former provides more granularity to the exact computational techniques used in PRL models, while the latter shows what kinds of representation types are the most mutually compatible.

### 2.2.1 Architectures by Representation Type

| Representation Type | Architecture | Papers |
|---|---|---|
| Vector Based | Fully-Connected DNN | [19] |
| | Convolutional NN | [18], [89], [101], [107] |
| | Autoencoder | [19],[109], [108], [67] |
| | CBOW and Skipgram | [89], [18], [95], [24] |
| Sequence Based | Recurrent NN | [83], [60], [106] |
| | Trasnformers | [92] |
| Graph Based | Graphs | [107], [63] |
| Matrix and Tensor Based | Tensors and Matrices | [110], [104] |

Table 2.2: Summary of the architectures studied.

**Vector-based** PRL architectures seek to represent patient data as mathematical vectors. This embedded information can then be used in clinical pattern recognition, risk assessment, and varied prediction tasks.

**Fully-Connected Deep Neural Networks** (FCNN) are the simplest of the vector-based architectures. These are usually considered to be baseline architectures and can only process structured data if used by themselves. However, they are rarely used alone and are often used at the end of mixed architectures to make the final predictions for the task in question. Che et al. (2015) [19] developed an architecture of fully-connected stacked denoising autoencoders ending in a Laplacian regularization layer which incorporates do-

main knowledge to process structured multivariate time-series, as presented in Figure 2.1. They used it to discover physiologic patterns associated with known clinical phenotypes and predictive of health outcomes.

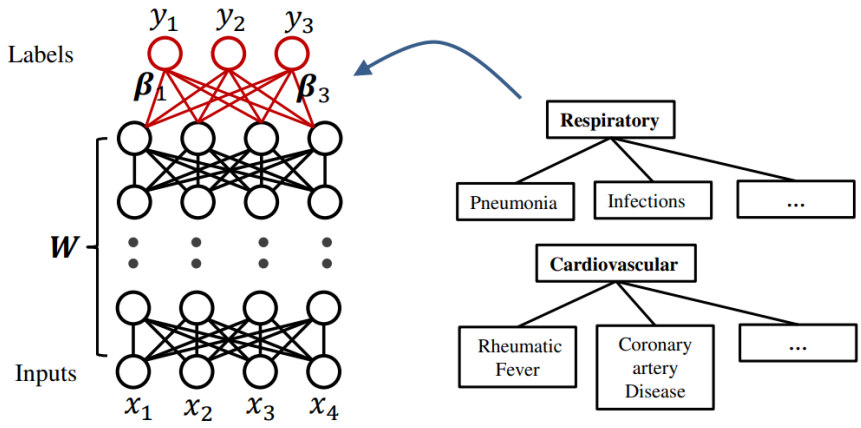



Figure 2.1: A miniature illustration of the deep network with the regularization on categorical structure. The regularization is applied to the output layer of the network. Image sampled from Che et al. (2015) [19].

**Convolutional Neural Networks** (CNN) were originally developed for image processing since they can identify desired features irrespective of location in a bitmap. The architecture has also been applied to text and waveforms successfully, however, and is used in this capacity for text-based PRL. Che et al. (2017) [18] developed an architecture which used word2vec, a word embedding architecture, to turn patient's medical events into vectors, which are then concatenated into a matrix where the X-axis represents the event vector's dimensions and the Y-axis represents each event, as presented in Figure 2.2. The matrix is then passed through a one-dimensional convolution layer over the temporal axis (Y-axis) in order to capture temporal dependency between medical events. This model was used in risk prediction tasks for diabetes and congestive heart failure.

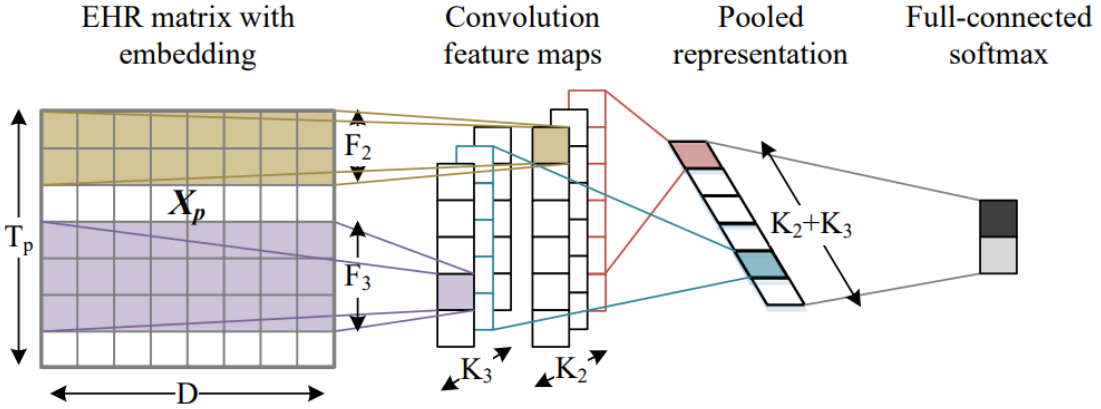



Figure 2.2: Convolutional neural network prediction model (with filters of size 2 and 3). Image sampled from Che et al. (2019) [18].

Si et al. (2019) [89] took another approach and developed a multi-task learning convolutional network that aggregates word embeddings derived from tokenized sentences of clinical notes into sentence representations through a convolutional layer, and subsequently uses another convolutional layer to create a patient vector representation from the sentence representations, as presented in Figure 2.3. This neural network was used in length-of-stay and mortality prediction tasks.

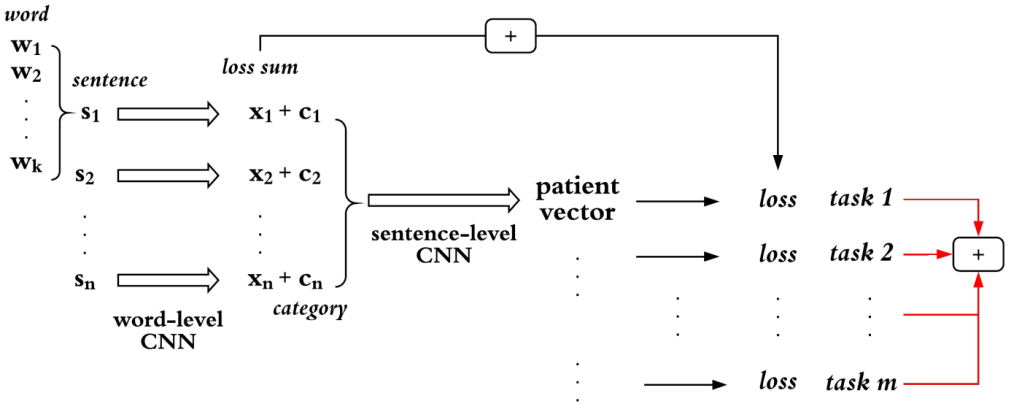



Figure 2.3: Deep patient representation model overview. Image sampled from Si et al. (2019) [89].

Xu et al. (2018) [101] use convolutional layers for a different purpose. Their RAIM model uses both dense and sparse structured continuous data as input. So that the dense data does not overshadow the sparse, irregular data, all channels pass through a convolutional neural network that outputs low-dimensional representations of the data, be it denser or sparser. Irregular events were also used to guide a multi-channel attention mechanism, as shown in Figure 2.4. This model was used to predict ICU 24 hour physiological decompensation and length-of-stay.

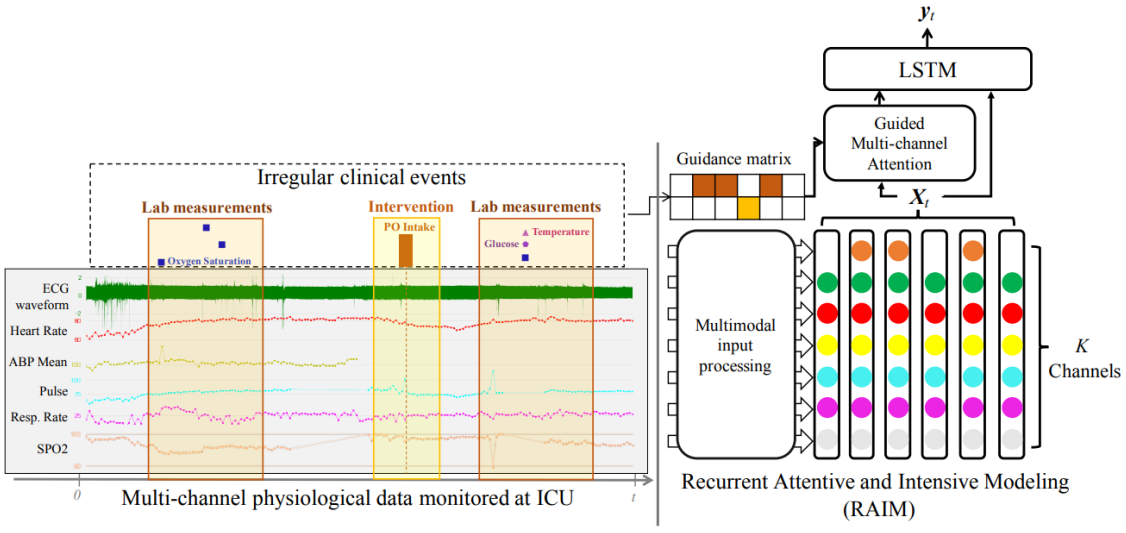



Figure 2.4: An overview of RAIM on multimodal continuous patient monitoring data. Image sampled from Xu et al. (2018) [101].

Zhang et al. (2017) [107] adapt a spatial CNN architecture to the irregular, heterogeneous domain of EHRs. Spatial CNNs were developed for use in graph spaces, and this architecture creates graph nodes from medical events and edges from the temporal relationships between them. It uses the nodes and edges to initialize node-specific parameters, which are then used in a heterogeneous convolution layer, concatenated with its nearest neighbors, pooled, and finally passed through a fully-connected neural network, as presented in Figure 2.5. This architecture was used for comorbidity risk prediction.

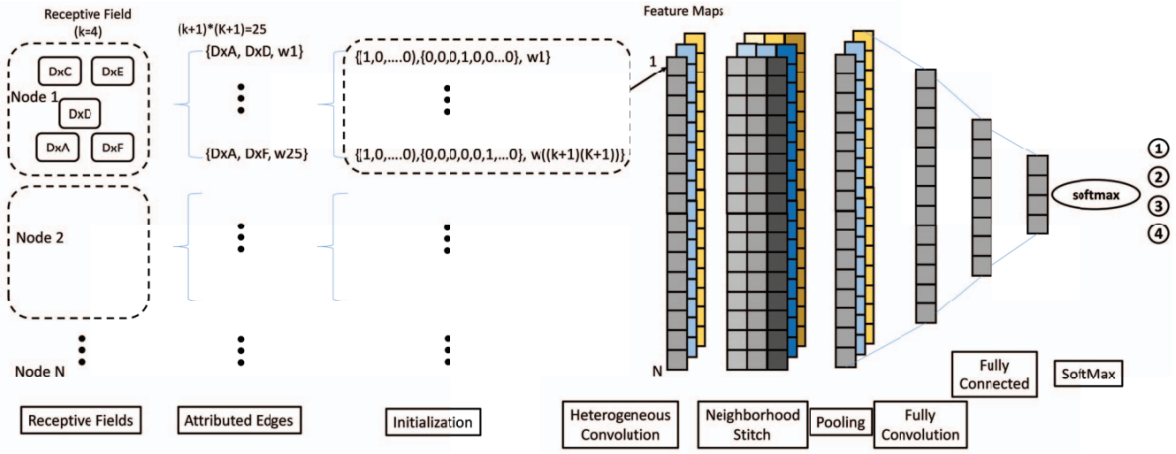



Figure 2.5: An overview of the deep learning architecture of the proposed model. The model accepts as input an attributed graph that represents a patient's EHR data. HCNN comprises the initialization of the attributed edges in the receptive fields, the heterogeneous layer, the pooling layer, fully convolutional layer, two fully connected layers, and the softmax output layer. The specified outputs are the probabilities of four types of chronic diseases. Image sampled from Zhang et al. (2017) [107].

**Autoencoders** are vector-based models that learn to compress high-dimensional, sometimes sparse data into lower-dimensional, denser data. It must be paired with other architectures to be used for clinical tasks, such as was seen in Che et al. (2015) [19] where denoising autoencoders were used in a fully-connected neural network. Another example of denoising autoencoders, a type of autoencoder often seen in the literature, used in clinical tasks is from Zhou et al. (2017) [109], who proposed using stacked denoising autoencoders to select the most useful features contained within interpolated irregular data so that classifiers would receive the best possible representation of a patient for the task of length-of-stay prediction, as seen in Figure 2.6.

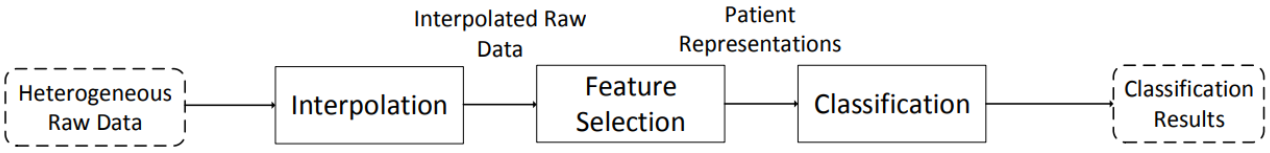



Figure 2.6: Overview of the predictive diagnosis framework. Image sampled from Zhou et al. (2017) [109].

Later, Zhou et al. (2019) [108] proposed a similar framework for other tasks. This framework, DFL, uses stacked denoising autoencoders to learn features from structured raw data that is built during a pre-processing step from heterogeneous data, as presented in Figure 2.7. The feature vectors learned from the autoencoders were then used in both a support vector machine and a fully-connected deep neural network trained for pneumonia prediction and alcoholism prediction.

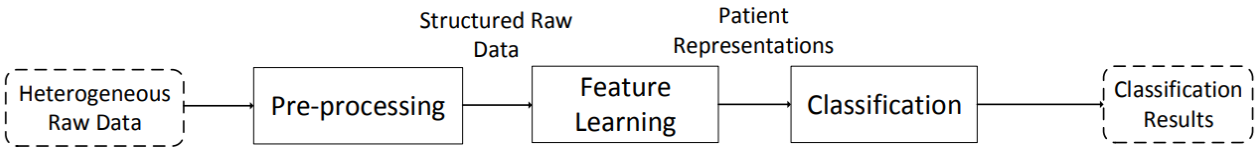



Figure 2.7: Overview of the DFL framework including the various data processing blocks. Image sampled from Zhou et al. (2019) [108].

Miotto et al. (2016) [67] developed another model using denoising autoencoders: Deep Patient. Rather than be aimed at a specific task, Deep Patient's focus is on creating deeply embedded feature vectors for patients, which would then be used as input in many tasks. Deep Patient was evaluated using disease prediction tasks centered around several diseases. These tasks were evaluated by disease (i.e., predict if a patient will develop a new disease within the time frame) and by patient (i.e., how many predictions were true for each patient).

**Continuous-bag-of-words (CBOW) and Skipgram embedding architectures** were developed to learn word embeddings from large-scale language resources. They can also learn embeddings of other kinds of sequences, such as clinical code sequences. Models like Si et al. (2019) [89] and Che et al. (2017) [18] used word embeddings created with the skip-gram and CBOW architectures, respectively, to embed free-text clinical notes. Stojanovic et al. (2017) [95], on the other hand, proposed using disease and procedure codes in timestamp order as the sequence to be embedded. As presented in Figure 2.8, sparse patient records of diseases and procedures are processed by their disease+procedure2vec method into dense vectors representing individual diseases and procedures, which are then summed together to represent a patient's visit. This patient representation can then be used in prediction models for several tasks, such as mortality prediction, length-of-stay, and medical charge regression.

Cui et al. (2018) [24] employed an approach similar to Stojanovic et al. (2017). They formed "medical sentences" from consecutive medical codes from which they trained their embeddings. The most important addition is the way their code vectors are constructed, being task oriented by following their custom process rather than being generalist in nature, as presented in Figure 2.9. The model was used in medical charge and length-of-stay regressions.

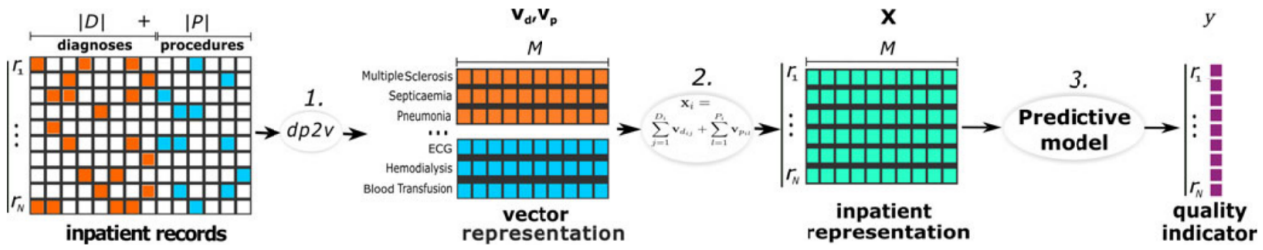



Figure 2.8: 1) Use the proposed embedding methodology to learn compact vector representation of diseases and procedures using raw EHR data. 2) Generate inpatient representation $X$ from the learned embeddings. 3) Train models to predict important indicators of healthcare quality $y$. Image sampled from Stojanovic et al. (2017) [95].

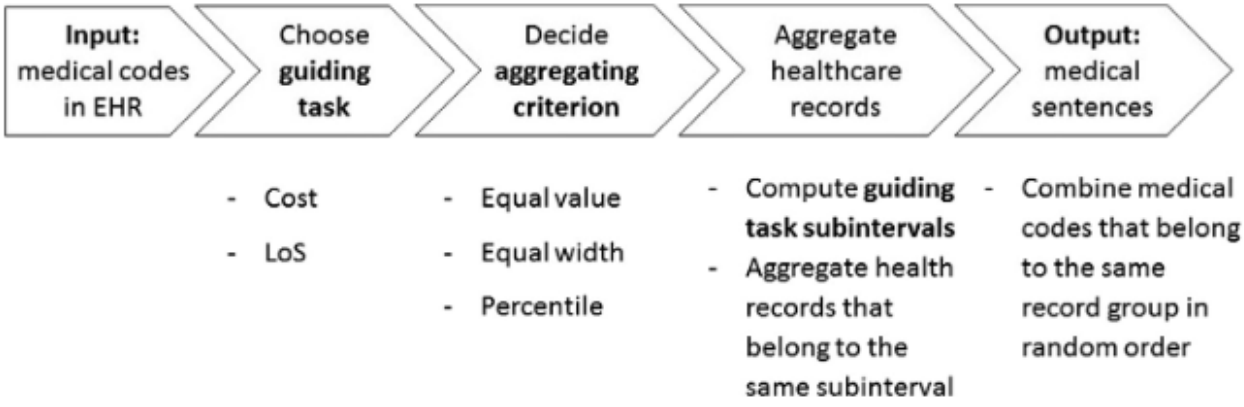


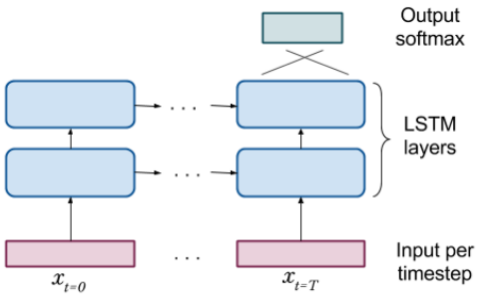

Figure 2.9: Cui et al. (2018)'s process. Image sampled from Cui et al. (2018) [24].

**Sequence-based** architectures are capable of processing sequential inputs such as language or time series. These architectures can be used to bring a temporal factor to the model and also open an alternative to word embeddings when attempting to parse free-text data.

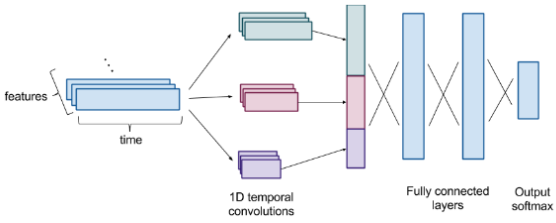
**Recurrent Neural Networks** (RNNs) process a sequence of inputs one at a time, transferring the hidden state information of a previous input into the next. Only variants of this original architecture are useful because the vanishing gradient problem means that only very short sequences can be considered by the original RNN architecture. The Long-Short Term Memory (LSTM) [83] and Gated Recurrent Unit (GRU) [60] variants of the RNN are widely used in PRL architectures.

Suresh et al. (2017) [96] compare a CNN-based architecture against an LSTM-based architecture for the intervention prediction task. The LSTM is fed with an hour of data at each timestep, predicting interventions at the final timestep, while the CNN performs temporal convolutions at 3, 4, and 5 hours granularities before passing the output through an FCNN to arrive at the prediction, as presented in Figure 2.10. This study found that RNNs either match or slightly outperform CNNs for the task in question.

Zhang et al. (2018) [106] developed a multi-input model which used a specific architecture for each kind of data. A CNN was used to process clinical notes and vital signs,




(a) The LSTM consists of two hidden layers with 512 nodes each. We sequentially feed in each hour's data. At the end of the example window, we use the final hidden state to predict the output.




(b) The CNN architecture performs temporal convolutions at 3 different granularities (3, 4, and 5 hours), max-pools and combines the outputs, and runs this through 2 fully connected layers to arrive at the prediction.

Figure 2.10: Schematics of the a) LSTM and b) CNN model architectures. Image sampled from Suresh et al. (2017) [96]

a combined CNN-LSTM was used to process prescription orders, and a FCNN was used to process clinical lab tests. The CNN-LSTM architecture processes averaged vectors of word embeddings extracted from the prescriptions. The inputs for each of these architectures were merged and used to train classifiers for the multi-label disease prediction task and the lab test order prediction task.

**Transformers** are equipped with self-attention mechanisms and positional embeddings to achieve better bidirectional representations. This architecture can encode time-stamped data as units and time series as sequences, upon which they employ attention and learn essential information. The attention mechanisms present in transformers are particularly useful for clinical tasks, as they allow the model to focus on small-albeit-important details found within large amounts of data.

Song et al. (2018) [92] developed the SAnD architecture to perform clinical time-series analysis using only attention mechanisms. It models the dependencies within a single sequence using self-attention and incorporates temporal order by using positional encoding and dense interpolation, as seen in Figure 2.11. It uses sequences of clinical measurements after performing a vector sum with the previously mentioned positional encoding to generate sequence-level predictions specific to the task being attempted. These sequences are then processed by a multi-head scalar dot-product attention module to create multiple attention graphs, which are then concatenated and linearly projected to a FCNN module, which obtains the final prediction. This model was used for mortality prediction, decompensation, length-of-stay prediction, and patient phenotyping.

**Graph-based** PRL architectures are characterized by the construction of graphs for each patient where nodes represent clinical events and edges represent relationships between events. As previously explored, Zhang et al. (2017) [107] exemplify this aspect of

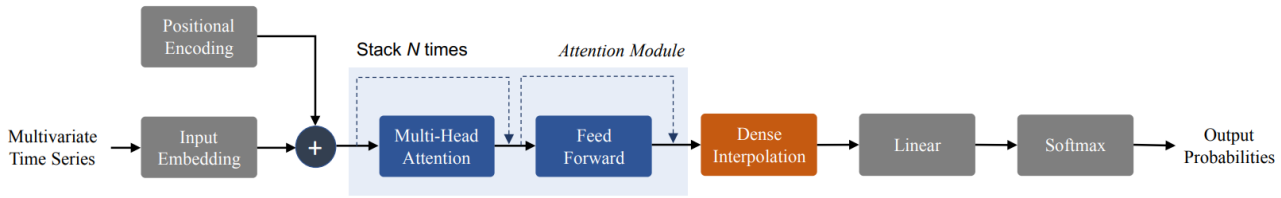



Figure 2.11: An overview of the SAnD architecture. This does not utilize any recurrence or convolutions for sequence modeling. Instead, it employs a simple self-attention mechanism coupled with a dense interpolation strategy to enable sequence modeling. The attention module is comprised of $N$ identical layers, which in turn contain the attention mechanism and a feed-forward sub-layer, along with residue connections. Image sampled from Song et al. (2018) [92].

graph-based architectures with the first three stages of their HCNN model, as seen in Figure 2.5 as well as the pre-processing of their data. This architecture is particularly useful when one wants to introduce domain knowledge into the architecture to enhance interpretability and performance.

Ma et al. (2018) [63] used directed acyclic graphs to add domain knowledge to their KAME model. This graph is used to create a knowledge based attention mechanism by identifying ancestors of the medical codes (which are leaves in the graph). The attention is used on the output of a recurrent neural network that sequentially processes medical code embeddings initiated from an input of medical codes pertaining to a hospital visit, as seen in Figure 2.12. The embeddings are created using the word2vec methods discussed previously. This model is used for disease prediction tasks.

**Matrix-based** and **Tensor-based** PRL architectures are based on the construction of multi-dimensional matrices to represent clinical events. The Matrix-based architecture uses two-dimensional matrices where one dimension is related to time and the other to clinical events. The Tensor-based architecture uses three or more dimensional tensors consisting of events such as diagnoses and treatments, time, etc.

Zhou et al. (2014) [110] used their Pacifier matrix-based framework to construct a longitudinal patient matrix on the medical feature $x$ time axes using EHR data. Their model then identifies latent medical concepts (reoccurring feature groups) and maps them onto a concept value evolution matrix, which is on a temporal axis, as presented in Figure 2.13. This model densifies sparse EHR data and uses the new data for phenotype prediction (for cardiovascular diseases, diabetes, and lung diseases).

Yin et al. (2019) [104] propose a model that constructs a third order tensor from patient data. The three axes are time, lab tests, and medication. They used it to learn dynamic patient-specific representations and phenotype definitions shared across all patients.

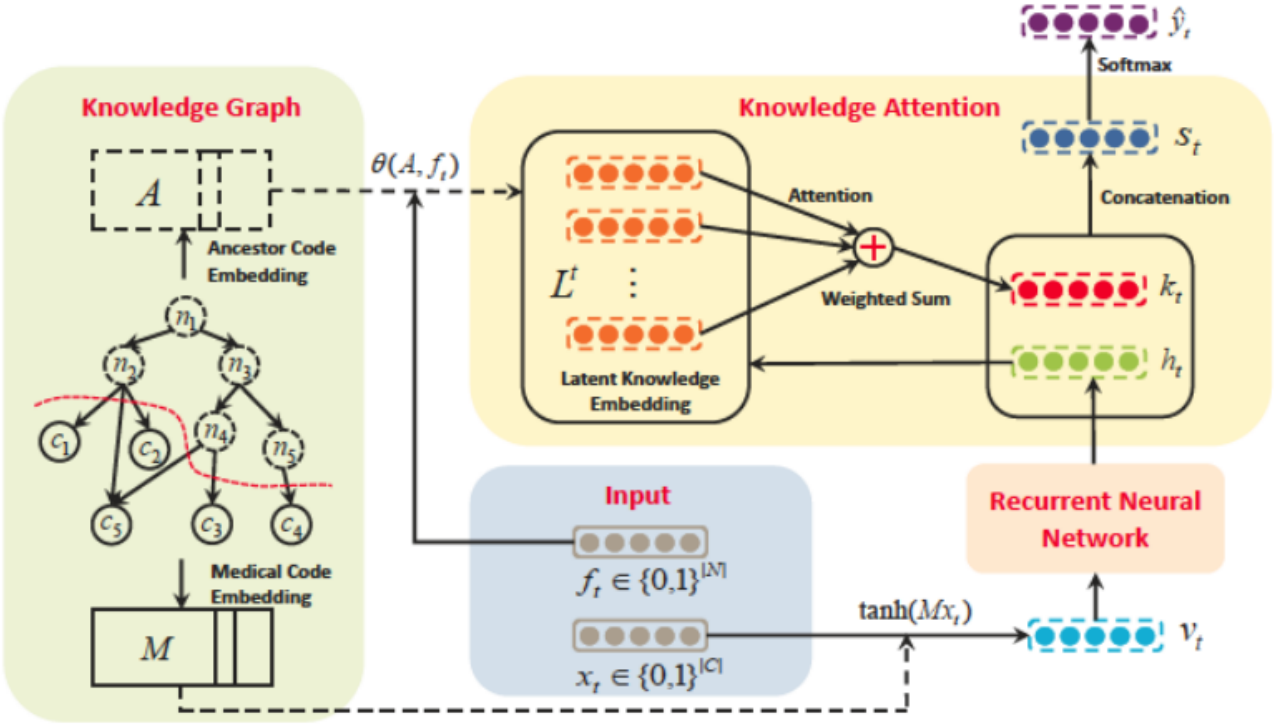



Figure 2.12: The KAME model. Image sampled from Ma et al. (2018) [63].

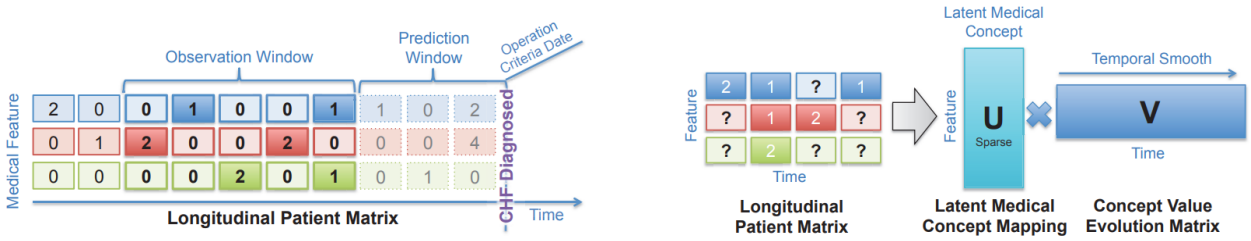



Figure 2.13: **Left:** Construction of the longitudinal patient matrix from EHRs. The goal is to predict a patient's disease status at the operation criteria date, given the past medical information before the prediction window. **Right:** Illustration of the Pacifier framework. A longitudinal patient matrix is treated as a partially observed matrix from a complete patient matrix. We assume the medical features can be mapped to some latent medical concepts with a much lower dimensionality such that each medical concept can be viewed as a combination of several observed medical features. Image sampled from Zhou et al. (2014) [110].

### 2.2.2 Architectures by Technical Paradigm

**Statistics learning-based** methods learn statistical representations from raw data, such as medication codes or procedure codes in the case of structured data or free-text clinical notes in the case of unstructured data. These statistical representations usually take the form of knowledge-embedded vectors located in continuous vector spaces. These methods must contend with the heterogeneous, temporally dependent, irregular, and sparse nature of EHRs directly to obtain good quality representations, but require the least amount of pre-

processing in order to use in downstream, task-focused neural networks. This paradigm accounts for all vector-based and sequence-based architectures discussed in the previous section, as well as graph-based representations specifically used for knowledge injection.

To better acquire relevant statistical knowledge from EHRs, researchers have taken to adopting strategies of knowledge preservation/injection and temporal dependency preservation to their models. Knowledge preservation/injection is the use of medical ontologies, knowledge graphs, and related texts to guide and enrich statistical EHR representations. Temporal dependency preservation focuses on preserving the information inherent to codified timestamps and time intervals between entries.

Though these two strategies have been implemented in the literature in several ways, the most effective seems to be by using attention mechanisms. Ma et al. (2018) [63] used knowledge graphs in the KAME model to direct attention mechanisms as an example of knowledge preservation. An example of temporal preservation can be found in Huang et al. [45], who used the attention mechanism to enhance clinical event vectors according to the time interval between temporally adjacent events.

**Graph-based** methods require the creation of a graph from the EHR data and use entity-relation-entity to calculate continuous vector spaces. These methods revolve around matrix and tensor decomposition, and graph neural networks. This paradigm encompasses all matrix-based, tensor-based, and graph-based architectures discussed in the previous section.

The matrix and tensor decomposition methods first reduce the dimensions of the adjacency matrix of the graph, obtaining low-dimensional representations of nodes. Graph Neural Networks aggregate neighbor node information recursively over the graph.

These methods can effectively learn structural information, temporal and semantic relations in EHRs, but graph construction remains a problem for research into these methods. Another challenge specific to this method is the creation of more expressive and interpretable graphs.

**Knowledge-based** methods are an offshoot of Graph-based models. The principal difference between graph-based methods and knowledge-based methods is that knowledge-based methods use knowledge graphs, which are condensed representations of knowledge association, while graph-based methods use object graphs, which are intuitive representations of object association.

These methods focus on representing knowledge in entity-relation-entity fact triples, which are then processed by linear, neural, or translation algorithms. Linear and translation algorithms were not mentioned by Si et al. (2021) [88], but neural algorithms are comprised of all techniques seen in vector-based and sequence-based architectures.

Linear algorithms use linear combinations between entities' and relations' representations to calculate probabilities. Neural algorithms use neural networks that take enti-

ties' and relation's representations as input to generate probability. Translation algorithms, the most successful and most used type of model in the literature, use relations as translations between two entities, having been inspired by translation invariance in word vector spaces.

While knowledge-based methods were shown to be of significant help in knowledge acquisition, fusion, and inference, the development of knowledge graphs from EHRs requires a lot of expert knowledge and human labor. The lack of public medical knowledge graphs fine-grained enough to use with specific clinical tasks is also an enormous hindrance. For these reasons, these methods see very little research, having been only 10.69% of all papers reviewed by Liu et al. (2022) [61].

## 2.3 PRL Testsets and Assessment

Per Liu et al. (2022) [61], PRL models can be assessed both intrinsically and extrinsically. Intrinsic evaluations measure the coherence between the model's ability to encode information and human judgment. Extrinsic evaluations are performed by judging their performance in downstream tasks.

The most used forms of *intrinsic evaluation* were similarity evaluations (used by 35% of the reviewed literature, which used intrinsic evaluations) and visualization evaluations (used by 71% of the reviewed literature, which used intrinsic evaluations).

Similarity evaluations compare the similarity of concepts in a vector space (words, patients, etc.) calculated by the model against human-perceived similarity as annotated by experts. The most common ways to calculate similarity between vectors are Euclidean distance and Cosine distance, while the two most common ways to measure the correlation between model similarity and human-annotated similarity are the Pearson and Spearman coefficients.

Visualization evaluations require the development of data visualization structures, with which human experts can analyze the data and subjectively discern its quality. Though visualizations are extremely varied, a common example is to reduce the dimensionality of vector spaces to two, and map concepts to a two-dimensional matrix in order to more easily analyze concept clusters.

The three most studied *extrinsic evaluation* types found in the review were Clinical Tasks (used by 72% of the reviewed literature), Named Entity Recognition (used by 6%), a task focused on finding and labeling named entities in free-text, and Relation Extraction (used by 2%). Clinical tasks dominate this category, as they represent real-world problems that might be solved with this technology, while all other tasks serve the main tasks.

Most of the literature used extrinsic evaluations based on clinical tasks, and we will follow this trend for several reasons. First, it is the best way to evaluate the representations

when there is a lack of dedicated expert help. Second, it is the most quickly understandable way to communicate the success of the machine learning models to outside experts. Third, extrinsic tasks provide the best way to compare different approaches to the same problem directly.

## 2.4 Patient Flow Prediction Subtasks

There are many tasks that could benefit from advances in PRL. Si et al. (2021) [88] identified several clinical task categories that are most explored within this field. These are, in order of most to least papers identified as tackling the category in their review, as follows: disease prediction, mortality prediction, length-of-stay forecasting, (re)admission prediction, patient subtyping, intervention prediction, and medical cost forecasting. The tasks that we have chosen to focus on are the Patient Flow tasks, specifically admission and discharge prediction.

### 2.4.1 Discharge Prediction

Discharge prediction, also referred to as the Length-of-Stay (LOS task), is the most studied patient flow task by far in Health AI [78, 102, 52]. It has been thoroughly studied using international datasets such as MIMIC-III.

Jaotombo et al. (2023) [47] studied 14-day length-of-stay at a French hospital, finding an AUROC of 0.8101 using a Gradient Boosting (GB) architecture. The GB architecture outperformed several other classic machine learning architectures, and even a multilayer perceptron architecture, by fractions of a percentage point. This study did not take into account outpatients, however, and excluded patients with LOS under 24 hours.

Kadri et al. (2023) [51] investigated LOS in a French pediatric emergency department, which included only non-adult outpatients of the ER. They found that a generative adversarial network (GAN) outperformed other networks, achieving an R2 of 0.871.

Alghatani et al. (2021) [4] used the MIMIC-III dataset [50] to predict ICU LOS. They set up a binary classification to predict whether a patient will stay longer than the median (2.64 days). XGBoost (XGB) achieved the best result with an AUROC of 0.70. As seen in those works, there are many ways to set up the LOS task.

Such methods have been shown to work in many international datasets, and, though data can be scarcer for Brazilian hospitals, there are several studies performed about the hospital LOS task in Brazilian settings. Kurtz et al. (2022) [53] studied mortality and LOS for stroke patients in 43 Brazilian hospitals. Their LOS goal was to predict whether a stroke patient's admission would exceed 14 days. This study found that the best performing machine

learning models for their data were GB and random forests (RF), both having achieved an AUROC of 0.73.

Silva et al. (2020) [36] investigated what factors most affect the LOS of cancer patients in Brazilian hospitals. They found that tumor location and stage are the most relevant, but whether it was an emergency hospitalization and patient age also played a significant role in LOS.

Medeiros and Tortorella (2023) [68] studied LOS for pediatric patients in a Brazilian university hospital. They attempted to predict the exact LOS, rather than binary classification, and their best performing model was an RF architecture that achieved 0.63 $R^2$.

Peres et al. (2022) [74] studied ICU LOS for several Brazilian hospitals. Their best performing model was a stacking RP and Linear Regression (LR) architecture, achieving 0.36 $R^2$. They also performed binary categorization tasks to predict whether a patient would stay for longer than 14 days. This task's best performing model, a stacked RF and LR architecture, which used GB as a meta learner, achieved 0.87 AUROC. None of the works we found focused on the early prediction of inpatient admission, and all of them used closed datasets with no availability for public use.

### 2.4.2 Admission Prediction

Few papers have investigated the task of inpatient admission prediction, and none have investigated this task for Brazilian hospitals. This task aims to discover which patients coming into the hospital will become inpatients in need of bed and extended care and which will remain outpatients. Emergency departments are the most common form of inpatient admission [44], so typically, works that attempt to predict inpatient flow use emergency department admissions [13, 25, 7] but they are not only source. Inpatients are often officially described as patients whose hospital stay exceeds 24 hours [28].

## 2.5 Automated Annotation Task

Social determinants of health (SDoH) were chosen for our automated annotation tasks because they contribute to an astonishing 80-90% of health outcomes [11, 64], with multiple factors significantly exacerbating health risks [6, 82]. Most importantly, the literature has found that critical SDoH information is only present in unstructured clinical narratives [99, 94].

Methods for SDoH extraction using NLP encompass rule-based (using keyword matching/counts or regular expression [43]), tool-based (specialized, task-specific system tools, such as Moonstone NLP [23] or cTAKES [85]), and supervised/unsupervised learn-

ing approaches (relying on annotated data and lexicons constructed manually or semi-automatically to train the learning models [72]). The manual procedure of training data annotation depends extensively on guidelines that steer the annotation process [100, 72], which are typically task-specific and, as such, have poor reusability.

Large Language Models (LLM), including pre-trained domain specific LLMs, have demonstrated promising potential across various healthcare applications [48, 69, 38, 12]. Large Language Models have the possibility of reducing the cost and improving the quality of data labeling [33, 1, 97]. Related studies have explored the application of LLMs to SDoH extraction with varying degrees of success in extracting SDoH.

Guevara et al. (2024) [41] employed a fine-tuned Flan-T5 model to classify SDoH categories, achieving mid-range F1 scores of 0.55 in Economics and 0.44 in Community. Similarly, the model used in Ramachandran et al. (2023) [79] demonstrated the ability to extract SDoH data from medical notes with moderate success but required extensive fine-tuning and significant computational resources.

## 2.6 Evaluation Metrics

Our two kinds of tasks required different kinds of metrics to evaluate our findings. Patient Flow tasks were evaluated using accuracy, precision, recall, and F1, while Automated Data Annotation used AUROC, Cohen's Kappa, and estimated time and money costs. We will use this section to explain each of the metrics used, especially the unusual cost metrics.

### 2.6.1 Accuracy, Precision, Recall, and F1

Accuracy is the proportion of all classifications that were correct, whether positive or negative. Precision is the proportion of all positive classifications that are actually positive. Recall is the proportion of all actual positives that were classified correctly as positives. Precision improves as false positives decrease, while recall improves when false negatives decrease.

F1 is the harmonic mean of a model's recall and precision scores. The closer to 1 the model's F1 score is, the better the model's performance in the categorization task. Weighted F1 is calculated by averaging the weighted score of each class (in our case, positive and negative). Weights for scores are calculated by dividing the number of occurrences for the class in question by the total number of samples. The Weighted F1 score is ideal for datasets where the classes are not expected to be balanced.

### 2.6.2 Cohen's kappa

To assess the level of concordance between SDoH-GPT and human annotation, each prompt was employed to annotate the respective testing dataset for each task. The inter-annotator agreement, quantified using Cohen's kappa metric, was computed by comparing the human annotations of the test dataset with the newly generated annotations by SDoH-GPT. Cohen's kappa accounts for chance agreement, making it a valuable measure for binary categorization, where it is likely that two annotators will achieve a high percent agreement due to guesswork alone. In cases such as our tests, Cohen's kappa provides a much more reliable agreement metric. Cohen's kappa (K) is calculated as follows:

$$K = \frac{P_{agree} - P_{chance}}{1 - P_{chance}} \tag{2.1}$$

Where $P_{agree}$ is the proportion of agreement, and $P_{chance}$ is the proportion of agreement which would be expected due to chance. In interpreting Cohen's kappa, 0 signifies no agreement and 1 signifies perfect agreement. McHugh (2021) [65] reports that for healthcare and clinical research, results falling between 0.6 and 0.79 can be considered as "moderate" agreement, those between 0.8 and 0.89 as "strong" agreement, and results exceeding 0.9 as "near perfect" agreement.

### 2.6.3 AUROC

The Receiver Operating Characteristics (ROC) Curve assesses a model's ability to discriminate between binary classes. It is drawn by plotting sensitivity (True Positive Rate) against 1-specificity (False Positive Rate) at several thresholds [10]. As the curve is based on probability thresholds, it cannot be plotted when given solely binary predictions.

The Area Under the ROC Curve (AUROC) is calculated by measuring the percentage of area below the ROC curve as it was plotted. If a model reaches an AUROC of 0.5, it is a random predictor and essentially useless. A model with an AUROC of 1 is considered a perfect discriminator. Defining what threshold determines a "good" AUROC score is difficult, but many works agree that an AUROC score above 0.9 is "excellent" [29, 17].

### 2.6.4 Estimating Time and Money Costs

To provide an estimation of annotation costs, we assume a perfectly balanced dataset, from which precisely 1024 negative samples and 1024 positive samples are extracted. Al-

though this scenario is seldom encountered in practice, the linearity of cost increase allows us to infer that both human annotation and SDoH-GPT annotation will scale at their respective rates the more data points we annotate. This calculation inherently favors human annotation, as a reduction in the number of notes to be annotated diminishes the relative advantage of SDoH-GPT. Despite this inherent advantage, the fact that human annotation is still less cost-effective underscores the substantial value of the SDoH-GPT system.

To estimate the costs associated with **human annotation**, several factors must be considered: the word count of the dataset, the reading speed of annotators, and the hourly rate for their services. To give a sensible estimate of the word count for all human-annotated training sets within a given task, we first calculate the average number of words per note (avg(#wpn)). This is achieved by determining the average number of notes from the complete 2048-sample training set. Once avg(#wpn) is calculated, we can estimate the word count (WC) for the simulated annotated sample scarcity datasets, ranging from 16 to 2048 samples. The WC is calculated by multiplying avg(#wpn) by the size of the dataset (size(dataset)), as seen in Equation 2.2.

$$WC = avg(\#wpn) * size(dataset) \tag{2.2}$$

Brysbaert (2019) [14] reports that the average reading speed for adults in non-fiction is 238 words per minute, with individuals at the upper range reaching up to 300 words per minute while maintaining full comprehension. We assume that professional medical annotators maintain the faster reading speed of 18,000 words per hour (300 words per minute times 60 minutes), the man-hours of labor (MHL) required for annotating the dataset can be estimated by dividing the WC by 18,000, as seen in Equation 2.3.

$$MHL = \frac{WC}{18000} \tag{2.3}$$

Carrel et al. (2016) [16] reported that the rate for annotators was $46.03 per hour (in 2014 dollars). Adjusting for inflation to 2023, this translates to $59.32. Furthermore, they also report that a 99% accurate annotation requires the efforts of two human annotators. This means that the human monetary cost in dollars (HMC$) for annotating the dataset is obtained by multiplying the MHL by 118.64 (59.32 * 2), as seen in Equation 2.4.

$$HMC\$ = MHL * 118.64 \tag{2.4}$$

In the calculation of HMC$, we assume that both annotators will take roughly the same amount of time to perform the annotation and that they can conduct the annotation simultaneously. This assumption entails that MHL for one annotator is the human time cost in hours ($HTC_h$), as seen in Equation 2.5. It is important to note that this calculation

disregards the subsequent analysis of annotator agreement and re-annotation in case of disagreement.

$$HTC_h = MHL \tag{2.5}$$

To estimate **SDoH-GPT annotation** costs, it is important to determine the token count in the dataset slated for annotation, the cost per token processed by the server, the time required for the server to process one sample, and the server's simultaneous processing capacity. An estimated token count is obtained using the Tiktoken Python module to calculate the total number of tokens in the 2048-sample dataset. Subsequently, the average token count per note (avg(#TK)) is computed. Once avg(#TK) is determined, the token count for the annotated sample scarcity datasets (ranging from 16 to 2048 samples) is estimated by multiplying avg(#wpn) by the size of the dataset (size(dataset)), as seen in Equation 2.6.

$$TKC = avg(\#TK) * size(dataset) \tag{2.6}$$

The cost of using GPT-3.5 in Azure is $0.002 per thousand tokens. The SDoH-GPT monetary cost in dollars (AMC$) for annotating the dataset is calculated by dividing TKC by 1000 and multiplying the result by 0.002. To account for the use of human annotation in the SDoH-GPT system, the cost of human annotation for 100 samples by two annotators is added to the monetary cost. This addition simulates the necessity of having a small human-annotated dataset for prompt refinement and example retrieval. The equation for AC discussed in the Human Annotation Cost Estimation section is employed for human annotation costs, considering 100 as the size (dataset) for the human-annotated samples. This can be seen in Equation 2.7.

$$AMC\$ = (\frac{TKC}{1000} * 0.002) + ((\frac{avg(\#wpn) * 100}{18000}) * 59.32 * 2 \tag{2.7}$$

To estimate SDoH-GPT time cost in hours ($ATC_h$), it was observed that the server answered a request in an average of 1 second and could answer four requests simultaneously. This implies our server could annotate roughly 240 samples per minute or 14,400 samples per hour. The time for human annotation of 100 samples is added to the calculated GPT time to simulate the need for a small human-annotated dataset for prompt refinement and example retrieval, akin to how AMC$ was calculated. The equation for MHL discussed above is used to calculate human annotation costs, considering 100 as the size(dataset) for the human-annotated samples. This can be seen in Equation 2.8.

$$ATC_h = (\frac{size(dataset)}{14400}) + (\frac{avg(\#wpn) * 100}{18000}) \tag{2.8}$$

## 2.7 Open Challenges

Both the review performed by Si et al. (2021) [88] and the one performed by Liu et al. (2022) [61] showcase several challenges that must be tackled by researchers in the PRL field. These challenges pertain to data processing, information preservation, and result explainability.

### 2.7.1 Data Acquisition and Processing

As previously explored in this chapter, freely available medical data is very scarce, and most of it does not come from Brazilian hospitals. This is because medical data acquisition is a very high-risk data since collecting and distributing such data presents many ethical challenges [98, 46], and as such, it is incredibly difficult find free-to-use data even if access must be credentialed.

The data must be de-identified before being put into open-access, and this is a difficult and expensive process. To achieve this, computer scientists must work closely with medical professionals and hospital administrations. This cooperation is currently uncommon but vital to the advancement of computational medicine [34].

Though recent advances have encouraged more and more digitalization of health data into Electronic Health Records, these data pose challenges in its use [56]. These challenges include the heterogeneity of medical data, which often includes structured data, such as categorical demographic information, numerical laboratory results, codes for procedures, diagnosis, and medications, as well as unstructured data, such as free-text clinical narratives that document a patient's stay in the hospital and images from radiology examinations [42].

The challenge of unifying the heterogeneous data into a coherent representation is still present in all areas of computational medicine. The task of mapping data in different forms that come from diverse sources into a unified space can be performed in two main ways: to use a single model to process all the data; or to use many models, each of which processes one or more types of data, and then performing a fusion over the resulting representation spaces [88, 61].

This is a complex task, however, and the reviews have shown that few studies take full advantage of the data they acquire [88, 61]. Instead, many use subsets of data and outright ignore others. Unstructured data is often cut from studies because of how much more difficult it is to work with when compared with structured data, for example.

### 2.7.2 Communication and Explainability

While many strides have been made in the field of medical computation, the conclusions reached by machine learning models are very abstruse and opaque for humans [80]. This is problematic in the incredibly high-risk field of clinical medicine, which prevents the use of the fruits of this research. Explainability and interpretability are thus paramount for building expert's trust in these models [32].

To build trust, it is necessary for experts to know the reasons behind decisions and predictions from machine learning clinical models [32]. However, the literature has been shown to be somewhat uninterested in tackling this particular issue and few efforts dive into the challenges of making EHR representations more human-friendly [88]. Still, some progress has been made. Visualization techniques often used to intrinsically evaluate models have been used to make representation spaces more intuitive, for example [88, 61]. Additionally, attention mechanisms help researchers discover which parts of the data are most important in prediction tasks, and this information can be used to add reasoning to a decision support system's suggestions [71].

## 2.8 Conclusion

After studying the state-of-the-art in the field of PRL, as presented in this chapter, we devised two ways to advance the field at both a general and national level.

First, we propose a holistic representation as an attempt to resolve two of the main challenges in the field: data processing and information preservation. To do this, our holistic representations were designed to use all available information about the patient whilst preserving temporal relativity between data points.

These points are not often discussed within the literature, as most works tend to use only subsets of the available data or ignore the temporal aspect altogether. They are, however, often cited as good avenues of research in literature reviews, however, given the inherently interconnected nature of human health.

In agreement with these conclusions, we seek to remedy these gaps to see whether our hypothesis that such holistic representation outperforms limited representations by a significant margin truly bears results. This is an important question, as it would impact how we gather clinical data, an already contentious topic by itself. If much of the available data is unimportant, perhaps the community needs to focus its efforts on discovering what can be ignored; if most or all of the data is important, then perhaps focusing on the deidentification of real world data and creation of synthetic data would best supplement research efforts.

Furthermore, clinical data is not widely available on a national level. As our goals are not only to advance the field of PRL research as a whole but also to advance its usability within a Brazilian context, we must expand national data availability for Brazilian-focused clinical AI research. Our efforts in data collection and test set development are expounded upon in the following chapters.

We endeavored, through this thesis to create of national medical data collection with BRATECA and used it, alongside the MIMIC dataset, to further the use of clinical AI in real-world national scenarios.

# 3. THE BRATECA TERTIARY CARE DATA COLLECTION

Because of the lack of data collections with the appropriate information and size to achieve our goals, we developed a new Brazilian tertiary hospital data collection alongside the Institute for Artificial Intelligence in Healthcare (IAIH)[1]. The data was collected and deidentified by IAIH's NoHarm.ai systems, organized by us, and checked and distributed as an organized effort with the help of Physionet [2]. This collaborative effort resulted in the BRATECA Collection [20].

BRATECA is a Portuguese-language tertiary care data collection that contains 73,040 admission records of 52,973 unique adults (18 years of age or older) extracted from 10 hospitals located in two Brazilian states. Amongst those admissions, several are associated with specialty treatment wards, as follows: publicly funded wards (12,096 admissions total); intensive care wards (4,666 admissions total); obstetrics wards (5,550 admissions total); COVID-19 wards (1,714 admissions total); surgical wards (25,004 admissions total); emergency wards (37,392 admissions total); and ambulatory wards (3,107 admissions total). The remaining 8,674 admissions are associated with any specialty wards.

The median patient age is 54 (Q1 = 38, Q3 = 68), 41.3% of the patients are male, 70.7% are identified as white, 3.8% are identified as mixed, 3.8% are identified as black and 0.2% are identified as yellow, and the mortality rate of patients is 6.5%. Each admission is paired with laboratory exam results (2,374,807 total), prescriptions and their itemized contents (519,318 total), and clinical notes (2,849,572 total). An interactive dashboard has been created to present some details of BRATECA and is linked in the project's GitHub page[3]. Table 3.1 presents statistics for each admission type.

## 3.1 Classes of Data

BRATECA is composed of descriptive data, laboratory data, medication data, intervention data, and clinical notes. Descriptive data includes patient specific information such as dates of birth, admission and discharge, skin color, height, weight, and reasons for discharge. Laboratory data include data on various laboratory exam results for patients. Medication data includes prescription items, as well as dosage, frequency, and other such administration details specific to each patient and prescription. Intervention data includes notes on whether there were pharmacist interventions on specific prescriptions that may have been mistakenly administered, as identified by the Institute for Artificial Intelligence

[1] https://noharm.ai/en/
[2] https://www.physionet.org/
[3] https://github.com/noharm-ai/brateca

| Admission Type | Publicly Funded | Intensive Care | Obstetrics | COVID-19 | Surgical | Emergency | Ambulatory | Normal |
|---|---|---|---|---|---|---|---|---|
| Median Age (Q1-Q3) | 58 (38-69) | 64 (52-73) | 30 (25-36) | 61 (49-73) | 56 (40-68) | 54 (38-69) | 44 (34-59) | 56 (40-70) |
| Median Laboratory Results (Q1-Q3) | 25 (0-53) | 119 (48-263) | 0 (0-17) | 117 (54-290) | 0 (0-10) | 17 (0-29) | 0 (0-0) | 0 (0-19) |
| Median Prescriptions (Q1-Q3) | 3 (1-10) | 28 (15-52) | 3 (1-6) | 15 (8-36) | 2 (1-6) | 1 (1-4) | 1 (1-2) | 3 (1-7) |
| Median Clinical Notes (Q1-Q3) | 12 (3-57) | 140 (77-291) | 11 (3-42) | 106 (54-231) | 5 (2-26) | 4 (2-14) | 2 (2-5) | 19 (9-32) |
| Male Percentage | 42.2% | 55.48% | 0.14% | 55.54% | 41.92% | 43.44% | 31.48% | 41.71% |
| Mortality Percentage | 5.13% | 24.09% | 0.07% | 17.68% | 2.44% | 10.62% | 0.19% | 1.44% |
| Skin Color Percentages | W: 66.32%<br>B: 8.59%<br>M: 9.22%<br>Y: 0.21%<br>NI: 15.67% | W: 67.10%<br>B: 2.48%<br>M: 3.36%<br>Y: 0.06%<br>NI: 27.00% | W: 68.81%<br>B: 11.68%<br>M: 9.44%<br>Y: 0.13%<br>NI: 9.95% | W: 78.65%<br>B: 4.03%<br>M: 3.79%<br>Y: 0.23%<br>NI: 13.30% | W: 83.25%<br>B: 3.26%<br>M: 2.90%<br>Y: 0.14%<br>NI: 10.46% | W: 60.50%<br>B: 4.25%<br>M: 4.65%<br>Y: 0.11%<br>NI: 30.48% | W: 83.71%<br>B: 3.41%<br>M: 2.22%<br>Y: 0.29%<br>NI: 10.46% | W: 78.20%<br>B: 2.02%<br>M: 1.59%<br>Y: 0.16%<br>NI: 18.03% |

Table 3.1: Rows present each the following information, from top to bottom: Median age (Q1 through Q3) per admission type; Median number of laboratory results per patient per admission type; Median number of prescriptions per patient per admission type; Median number of clinical notes per patient per admission type; Percentage of male patients per admission type; mortality percentage per admission type; percentages for patient skin color identification (W is white, B is black, M is mixed, Y is yellow and NI is no information). Columns each present one type of ward. Wards deemed "normal" are those that do not fall into any of the other categories. Note that a single admission may have a patient move wards one or more times, and a single ward may belong to more than one category.

in Healthcare's NoHarm.ai clinical pharmacy AI system [26]. Notes are free-text clinical notes describing a patient's evolving hospital admission.

## 3.2 Development Methods

BRATECA is an edited and reorganized version of the Institute for Artificial Intelligence in Healthcare's own internal Brazilian tertiary care information database and is intended to be a public[4] edition for use in machine learning research. For this purpose, certain data tables deemed most useful at the time of extraction were reorganized into the 5 datasets of BRATECA. This section presents the process of extraction and deidentification of the database's information into the format presented in Section 3.3.

### 3.2.1 Dataset Organization

The Institute for Artificial Intelligence in Healthcare's database is centered around its prescription tables. This resulted in only admissions with prescriptions being extracted,

[4] Note that BRATECA is property of the Institute for AI in Healthcare and only credentialed access is allowed, but it is freely available for research use.

as the prescription tables contained ward information and were the best way to ascertain that only adult patients from the desired wards were extracted from the database.

Beyond those requirements, only admissions that both began and ended during a delimited time period of nine months were extracted. This time period was set to sometime between 2020 and 2021, but this will not be specified so as to further enhance patient privacy. All admissions that fit within the presented parameters had their IDs extracted and used to gather related data from the database and create the five separate but interconnected datasets: Admission, Exam, Clinical Note, Prescription, and Prescription Item. These datasets are further described in Section 3.3. The SQL scripts used to extract the data are available on the project's GitHub page[3].

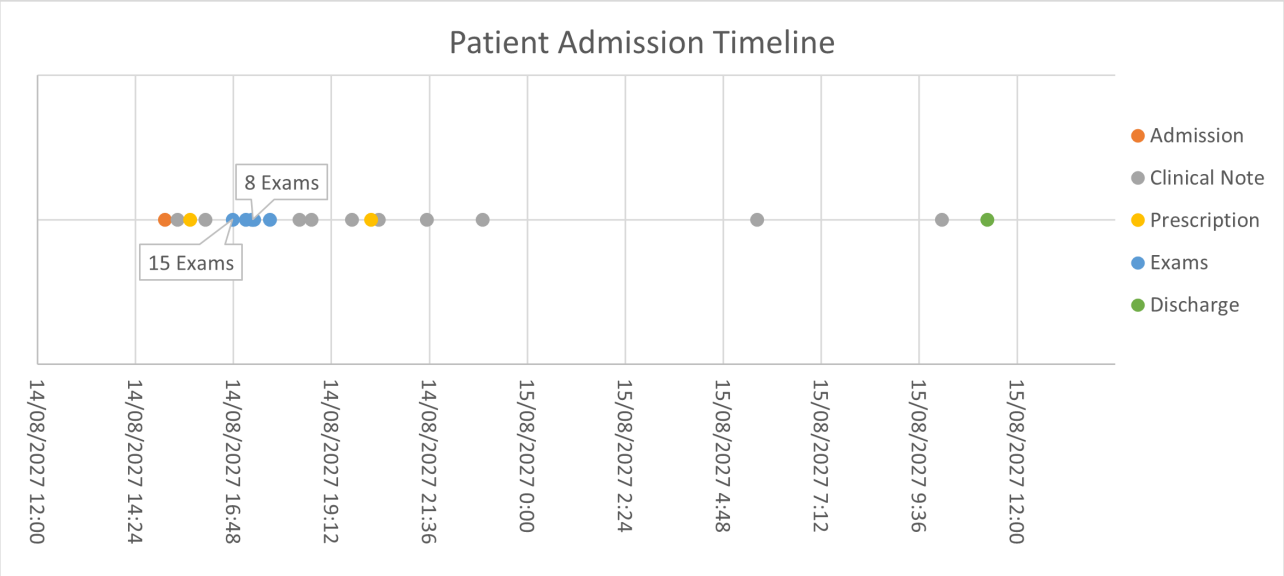



Figure 3.1: A simple example timeline of an admission, including recorded time of admission, laboratory examination, prescription administration, clinical note writing, and discharge. The two labels represent two instances where events were logged at the same time. In these cases, 15 and 8 exam results were logged simultaneously at two separate points in the timeline.

### 3.2.2 Deidentification

Though most columns in the datasets provide the exact information present in the original database, some had to be modified to further protect patients' sensitive information and attempt to prevent reverse engineering of identities from the provided data.

All names in BRATECA's free text notes were deidentified using state-of-the-art deep learning methods (Bi-LSTM-CRF) [3]. Two corpora and three language models were evaluated on a Named Entity Recognition (NER) task focused on person names to evaluate which combination delivered the best performance. The experiments revealed that using

domain-specific corpora (focused on deidentification of clinical notes) and a contextualized embedding stacked with word embeddings achieved the best results: an F-measure of 0.94 and Recall of 0.95 [84]. Dates present in the free text notes were also removed, though not using NER but rather regular expressions. The date removal script is available on the project's GitHub page[3].

Furthermore, all dates not part of free-text notes were shifted randomly 5 to 10 years forward. Dates referring to the same admission were shifted the same amount of days forward (i.e., if admission "1" of the Admission dataset was shifted 100 days forward, all dates of all entries in the other 4 datasets which refer to admission "1" in their Admission ID field were shifted 100 days forward as well). This was done in order to maintain timeline coherence within the same admission. Note that multiple admissions of the same patient may not be in chronological order and do not maintain any sort of temporal relation in order to more thoroughly deidentify such patients.

All internal database IDs, such as those for Patient ID or Admission ID, were also deidentified. Each was assigned a random numerical ID, congruent between datasets (i.e., if Admission ID "123456" is assigned the new ID "789" in the Admission dataset, the Admission ID "123456" was also assigned "789" whenever it appeared in the other 4 datasets).

Finally, ward information[5] was generated using the actual names of the wards of the hospitals from which the information was collected. Ward names were replaced with the aforementioned labels to better prevent hospital identification while maintaining some of the more relevant information. The generation was performed with the help of an active healthcare professional.

## 3.3 Data Records

BRATECA is composed of five datasets in the CSV (Comma Separated Values) format. These are as follows: Admission, a dataset of every individual admission, which includes patient demographic data; Exam, a dataset of exams and their respective results performed for each admission; Prescription, a dataset of prescription headers, which includes information such as patient/admission ID for the patient/admission which received the prescription, pharmacy assessments, prescription date, expiration date, ward information, whether the prescription includes special medication such as controlled substances, intravenously administered drugs (IV drugs), and antibiotics; Prescription Item, a dataset of prescribed medications which includes details of each prescribed medication, including name, dosage, and information on how the medication is to be administered, with each entry of this dataset being directly related to a prescription header in the Prescription dataset;

[5] Public, IC, Obstetrics, COVID-19, Surgical, Emergency, and Ambulatory. See the Prescription row of Table 3.2 for further details.

| Dataset | *Column* | Description | *Column* | Description |
|---|---|---|---|---|
| Admission | *Hospital_ID* | The identification code for the hospital from which the data originated. | *Patient_ID* | The identification code for the patient for whom the admission was registered. |
| | *Admission_ID* | The identification code for the admission to which the information belongs. | *Date_of_Birth* | Patient's date of birth. |
| | *Gender* | Patient's gender. | *Admission_Date* | Date patient was admitted to hospital. |
| | *Skin_Color* | Patient's skin color. | *Height* | Patient's height. |
| | *Weight* | Patient's weight. | *Height_Date* | Date patient's height was measured. |
| | *Weight_Date* | Date the patient was weighted. | | |
| Exam | *Hospital_ID* | The identification code for the hospital from which the data originated. | *Patient_ID* | The identification code for the patient for whom the admission was registered. |
| | *Admission_ID* | The identification code for the admission to which the information belongs. | *Exam_Name* | Name of the exam that was performed. |
| | *Exam_Date* | Date the exam was performed | *Value* | Numerical value of the result of the exam. |
| | *Unit* | Unit of measurement the exam's Value is in. | | |
| Clinical Note | *Hospital_ID* | The identification code for the hospital from which the data originated. | *Patient_ID* | The identification code for the patient for whom the admission was registered. |
| | *Admission_ID* | The identification code for the admission to which the information belongs. | *Note_Date* | Date the note was written. |
| | *Note Text* | The contents of the note. | *Notetaker_Position* | Notetaker's job title. |
| Prescription | *Hospital_ID* | The identification code for the hospital from which the data originated. | *Patient_ID* | The identification code for the patient for whom the admission was registered. |
| | *Admission_ID* | The identification code for the admission to which the information belongs. | *Prescription_ID* | The identification code for the prescription to which prescription items are associated. |
| | *Prescription_Date* | Date the prescription note was written. | *Pharmacy_Assessment* | Whether the prescription was revised by a pharmacist. |
| | *Expiration_Date* | Prescription expiration date. | *Assessment_Date* | Date the pharmacy assessment was performed. |
| | *Allergy* | Whether patient is allergic to one or more of the prescribed medications. | *Prescription_Score* | Score generated by artificial intelligence (the higher the score, the more unusual the prescription). |
| | *Alerts* | Prescription alerts. A complete list of alerts is shared in the documentation. | *Score_One* | The quantity of prescription items given a "1" score by the AI. |
| | *Antibiotics* | Number of antibiotics prescribed. | *Score_Two* | The quantity of prescription items given a "2" score by the AI. |
| | *High_Alert* | Number of high alert medication prescribed. | *Score_Three* | The quantity of prescription items given a "3" score by the AI. |
| | *Controlled* | Number of controlled medication prescribed. | *Tube* | Number of IV drugs prescribed. |
| | *Not_Default* | Number of non-standard medications prescribed. | *Different_Drugs* | Number of prescribed medications not previously reviewed by a pharmacist. |
| | *Alert_Exams* | Alerts related to exams. Examples can be found in the documentation. | *Interventions* | Number of interventions related to the prescription. |
| | *Complication* | Number of complications detected in clinical notes related to the prescription. | *Public* | Whether or not the prescription is for a publicly funded ward. |
| | *IC* | Whether or not the prescription is for an Intensive Care ward. | *Obstetrics* | Whether or not the prescription is for a obstetrics ward. |
| | *COVID-19* | Whether or not the prescription is for COVID-19 ward. | *Surgical* | Whether or not the prescription is for a surgical recovery ward. |
| | *Emergency* | Whether or not the prescription is for an emergency ward. | *Ambulatory* | Whether or not the prescription is for an ambulatory ward. |
| Prescription Item | *Hospital_ID* | The identification code for the hospital from which the data originated. | *Patient_ID* | The identification code for the patient for whom the admission was registered. |
| | *Admission_ID* | The identification code for the admission to which the information belongs. | *Prescription_ID* | The identification code for the prescription to which prescription items are associated. |
| | *Drug_Name* | Name of the drug. | *Dosage* | Dosage of each administration. |
| | *Daily_Frequency* | Number of times a drug is administered per day. | *Administration_Route* | Route of drug administration. |
| | *Note* | Medical observations related to the prescription. | *Normalized_Dosage* | Dose converted to a single numerical unit. |
| | *Time* | Time each dose is to be administered. | *Source* | Whether it is nutrition, a drug, a procedure drug or a solution. |
| | *Suspension_Date* | Date the medication is to be suspended. | *(Solution)_Group* | Group to which the solution belongs. |
| | *(Solution)_at_Medical _Discretion* | Medical observations related to the prescribed solution. | *(Solution)_Steps* | Frequency of solution adminitration. |
| | *(Solution)_Hour* | Time each solution dose is to be administered. | *(Solution)_App_Time* | How long a solution is to be administered for. |
| | *(Solution)_Dosage* | The dosage of the solution. | *(Solution)_Unit* | The unit of measurement of the dosage. |
| | *Administration_Period* | The period during which the item is to be administered. | *Allergy* | Whether the patient is allergic to the prescription item. |
| | *Tube* | Whether the prescription item is administered intravenously. | *(Intervention)_Date* | Date of the intervention |
| | *(Intervention)_Note* | Medical observations related to the intervention. | *(Intervention)_Status* | Resolution of the intervention request. |
| | *(Intervention)_Update* | Date of the final intervention update. | *(Intervention)_Motive* | Motive of the intervention. |
| | *(Intervention)_Error* | Intervention considered a prescription error. | *(Intervention)_Cost* | Intervention that generated a reduction of costs. |

Table 3.2: Columns and descriptions of columns for each of the five datasets.

and Clinical Note, a dataset of free-text clinical notes on details of the patient's stay and

treatment. A simple example of a patient timeline that shows details from all datasets in conjunction can be seen in Figure 3.1.

All datasets have IDs that are used for the identification of relations between entries in each file. These are: Hospital ID, the identification for the hospital from which the raw data was collected; Patient ID, the ID for a given patient in the database; Admission ID, the ID for the patient's admissions, of which a single patient might have many; and Prescription ID, specific to the Prescription and Prescription Item datasets, which identifies prescription items as belonging to specific prescriptions.

The datasets were developed in the way described above so that they can be used separately as well as in conjunction. Each is composed of several columns from tables in the original database, organized for ease of use. The information in each of the five datasets is presented in Table 3.2.

## 3.4 Usage Notes

### 3.4.1 Data Access

As mentioned previously, BRATECA is distributed by the Institute for Artificial Intelligence in Healthcare through Physionet credentialed access. In order to receive access, the researcher must complete the following steps:

1. Sign in to and confirm your identity in the Physionet platform;
2. Complete a course on protecting human research participants;
3. If the requester is a student, their supervisor must also agree to the terms of confidentiality;
4. Access the BRATECA page in Physionet[6] and request access to the dataset;
5. Wait for approval by the Institute for Artificial Intelligence in Healthcare.

Once the process is complete, and if the request is accepted, the researcher will be granted access to the dataset files.

[6] https://doi.org/10.13026/v8a6-mr20

### 3.4.2 Example Usage

There are many tasks that could benefit or even require datasets, such as BRATECA. Prediction tasks can use these datasets for training purposes. Mortality prediction can use discharge information as mortality annotation, for example.

Researchers with access to the original database have already published several papers with the information that is to be released in BRATECA. Santos et al. (2021)[84], for example, used state-of-the-art methods to identify and remove names from clinical texts. These were the methods used to deidentify all free-text notes made available as part of BRATECA, as mentioned in Section 3.2.2. Other examples of use of the data are listed below:

- Evaluation of a Prescription Outlier Detection System in Hospital's Pharmacy Services [26];
- Analysis of Pharmaceutical Interventions Performed with Decision Support Using Artificial Intelligence in Brazilian Hospitals [27].
- PsyBERTpt: A Clinical Entity Recognition Model for Psychiatric Narratives [70]
- Benchmarking the BRATECA Clinical Data Collection for Prediction Tasks [22]
- Predicting Inpatient Admissions in Brazilian Hospitals [21]
- Semantic Textual Similarity for Abridging Clinical Notes in Brazilian Electronic Health Records [5]

Besides published papers, much research work making use of BRATECA is well underway. Some examples are listed below:

- A machine learning-based clinical decision support system to identify possible drug intervention [54];
- Detection of Drug-Induced Liver Injury (to be published);
- Trends in the use of corticosteroids during the Pandemic (to be published).

Finally, several other usages of the dataset are being investigated or set to be explored in the near future. The large amount of free text notes, for example, permits the training of domain-specific language models with word embedding architectures such as Word2Vec [66] and fastText [40], and also contextual embedding models such as ELMO [75] and BERT [31]. Embeddings like these can be even more specific, using only certain parts of the data, such as limiting training to texts about elderly patients or intensive care patients.

Another avenue of research being explored is the use of the information to create real-time digital twins of patients by utilizing representation learning technology. These

digital twins could be used to predict patient developments and aid medical workers in keeping track of the most important information for each of their patients via alerts, data organization, and information retrieval [87].

## 3.5 Ethical Concerns

BRATECA has been deidentified according to the Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. The NoHarm.ai system, developed by the Institute for Artificial Intelligence in Healthcare, gathers no identifiable information from patients.

The data used for the BRATECA Collection came from a research project developed with several hospitals in Brazil. Also, all data sharing was approved by each hospital participating in that research. Ethical approval to use the hospitals' datasets in this research was granted by the National Research Ethics Committee under the CAAE number 46652521.9.0000.5530.

# 4. DATA TREATMENT AND TEST SET CREATION

This chapter explores the development of clinical prediction tasks and resulting test sets from information present in the BRATECA collection. It also explores automatic augmentation and enhancement of this information for more effective use in tasks. Moreover, tests were performed on automatic annotation from information collected from the MIMIC dataset in the MIMIC-SDOH collection [2].

The clinical prediction test sets presented here were the first to be developed for the BRATECA dataset. At the time of their development, these were the largest publicly available test sets for admission prediction and length-of-stay prediction for Brazilian hospitals. They were used to assess the effectiveness of our PRL architectures, and to serve as proof-of-concept for future use in real settings. Since part of our objective is to offer a real, usable solution for patient flow tasks in real scenarios, the ability to build trust with scientific results is important.

Clinical prediction often involves challenging, high-risk scenarios and high quality validation is a good first step when attempting to show the reliability of PRL to those for whom computation is not their area of expertise, such as medical professionals. This makes good test set availability and quality some of the main priorities when choosing an appropriate data collection for training prediction models.

Test set creation is difficult and expensive, however, as it often requires further annotation than a raw dataset such as provided by the BRATECA collection. Patient Flow was specifically chosen as an area to focus on because its tasks of LOS and Admission Prediction did not need further annotation, given the data available in BRATECA.

As an effort to explore cost-saving options for the future creation and curation of annotated datasets, we have worked with the AI Health Lab at the University of Texas at Austin [1] to perform research into automated dataset annotation with the use of Large Language Models in the MIMIC dataset with the objective to provide a proof of concept for future developments in this field which might enable the use of this technology with Brazilian data, which we were unable to successfully perform for this thesis.

## 4.1 Heterogeneous Data Treatment

In an attempt to rearrange heterogeneous information in ways that are best usable by the models we planned for the Patient Flow tasks, we explored methods of data transformation and vectorization. Our two kinds of sources of data were structured data

[1] https://aihealth.ischool.utexas.edu/

points (demographic admission data, types and quantities of drugs prescribed, exams performed), and large amounts of unstructured text (clinical notes).

In reviewing the data available and the state-of-the-art solutions found during our literature review, we found that the most appropriate solutions would mainly be built around vector-based representations to extract meaning from our structured data and sequence based representations to extract meaning from our unstructured text data (as seen in Section 2.2.1) and a statistics leaning-based technical paradigm (as seen in Section 2.2.2).

We also noted, however, that our data was incredibly temporally sparse. Most patients, especially for the admission tasks, which have shorter input deadlines (1 hour and 8 hours, as seen in Section 4.2), don't have enough data on exams and prescriptions to make for anything but an extremely sparse and time-series. The exam dataset at least had fewer columns, at 117 possible exams a patient might take, though in most cases, they would only take a few exams a single time, and very rarely any exam more than twice. The Prescription dataset, however, if simply vectorized, would have had over 1000 even sparser columns. This led us to decide to vectorize only the Exams data as it was, but to change our approach to the Prescription data.

So we transformed the Prescription Items dataset, which includes details on all medication prescribed to the patient, including name, timeliness, and dosage, into text with Table-to-Text Generation. This would allow us to more easily vectorize this information into a denser (if not always a dense) information vector, alongside our preexisting Clinical Note dataset, which is comprised of an unstructured free text data.

As for the text data, three methods of text processing were tested: Term-Frequency Inverse-Document-Frequency (TF-IDF); BERT encoding; and LLM encoding. These methods were able to recover meaningful information from the text and encode it in a more homogeneous manner with the rest of the vectorized data.

The details of table-to-text generation and the text processing methods are explained below.

### 4.1.1 Table-to-Text Generation

A large quantity of information in BRATECA is in the shape of tables. In Table 3.2, Chapter 3, we see that Admission, Exam, Prescription, and Prescription Item are structured data. Many of these structured data are very sparse, especially those belonging to the Prescription and Prescription Data categories.

To mitigate this problem, we sought to perform a Table-to-Text transformation so that we might be able to more effectively transformer-based architectures such as BERT and most LLMs. This would bypass the need to learn directly from sparse data and allow

the use of new techniques for this problem. For the purposes of time and cost saving, we only used Template-based Table-to-Text generation.

Template-based generation involves creating a standard text template with specific places to insert the appropriate structured information. Since our tables were sparse, these templates were developed to elegantly handle empty data strings by programmatically omitting irrelevant parts of the template. This method is quick and computationally cheap and, as such, preferable when taking into account real-world usage.

### 4.1.2 TF-IDF Vectorization

TF-IDF [93] is a classic method of determining which words are most meaningful in a given collection of documents. It can determine how meaningful each of these most meaningful words are to each individual document, while also accounting for the fact that some words are likely to appear many times. Words which appear many times in many documents are understood to be less meaningful individually.

In our architecture, we first concatenated all text we had for a patient (both Clinical Notes and Table-to-Text Prescriptions) before the input cut-off time (as explained in Section 4.2) in the order in which each individual note was taken or prescription given. Once concatenated into a single document, the notes and associated admissions were divided into training and testing sets, and then used to construct a TF-IDF vector of the 2500 word vector of the most meaningful and important words in the training corpus. Finally, we mapped the concatenated text of each document to this 2500-number vector, vectorizing the unstructured text data.

### 4.1.3 BERT Vectorization

Bidirectional Encoder Representations from Transformers, or BERT [31], is an open-source machine learning framework that uses deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers to interpret and meaningfully encode natural language text. Several BERT models exist that have been trained on Portuguese texts, but the most relevant for our use is BioBERTpt [86], a model fine-tuned from BERT-multilingual-cased with clinical narratives from Brazilian hospitals and abstracts of scientific papers from Pubmed and Scielo. As in our TF-IDF vectorization, the available text is concatenated into a single document in order of data creation up until the input cut-off time.

BioBERTpt can only accept 512 tokens in its input at once, necessitating that most concatenated text be truncated so they might fit this maximum size. The model outputs

a 768-float vector representation of the input text information, which can then be used as input into our architectures.

### 4.1.4 LLM Vectorization

The use of BERT-style transformers is limited by a maximum size of input as well as the diversity of the data used to train it Since the Portuguese language has much less data from which to train models than the English language, its models are more limited as well. Since the BRATECA corpus' text is full of specialist language, often oddly formatted, and has quite long text entries, it plays into the weaknesses of the available transformers. Term-Frequency vectorizers, on the other hand, may not suffer from these weaknesses but lose much context when they transform text data.

In order to mitigate these weaknesses, we proposed the use of the much more powerful encoder-decoder large language model LLaMa-3.1 8B, with its 8 billion parameters and maximum input of over 128,000 tokens, to process this information and output a 4,096-float representational vector of the text. The LLM was prompted to summarize the entries into more readable and understandable forms while maintaining all the most important information, and the encoded vector one layer prior to the decoder's language generation was extracted to use as a meaningful dense knowledge vector. This method would use zero-shot LLM summarization to parse and regenerate the clinical note text so the encoded vectors can be extracted. As the text is sensitive, only local LLMs could be used, rather than API-based solutions such as ChatGPT.

## 4.2 Patient Flow Test Sets

As previously discussed, test sets for two tasks were prepared from BRATECA data: Admission prediction and Extended Stay prediction. These test sets were built using as much data as available in BRATECA, with inputs and outputs adjusted to their respective focus tasks. This section explains what each test set contains and the specifics of the tasks they were created to tackle.

We devised six datasets for our tests: two for Admission prediction and four for Extended Stay prediction. All of these datasets use the same data extracted from the BRATECA collection. The data encompass broad personal characteristics such as sex, skin color, and age; exam and prescription data; and clinical notes written about the patient's stay. There is no historical data beyond what the clinical notes may relay textually.

The full set of data transformations used to create usable datasets from the raw BRATECA data was extensive and spanned across all five of the BRATECA collections, as

they each intersect with the others to show a fuller picture of the patient's stay. Furthermore, each of the six task datasets draws from this same data pool but considers different levels of information. Each task has a cut-off time for collecting training data and a decision time. The cut-off time delineates that all data that comes before it may be used for training, e.g., for the Admission 8-to-24 task, the cut-off time is 8 hours, so all information collected from the patient in their first eight hours in the hospital may be used for training. The decision time delineates the moment where the prediction must be made for whether a patient will be discharged before or after, e.g., for the Admission 8-to-24 task, the decision time is 24 hours, so the model must predict whether a patient will be discharged before or after 24 hours in the hospital. These data were arranged as follows:

| Task Name | Input Cut-Off Time | Output Prediction Time | Data Size | Testing Data Size | Positive to Negative Ratio |
|---|---|---|---|---|---|
| Admission Prediction at 1 Hour | 1 Hour | 24 Hours | 69725 | 17431 | 47% |
| Admission Prediction at 8 Hours | 8 Hours | 24 Hours | 53271 | 13318 | 61% |
| Extended Stay Prediction at 24 Hours | 24 Hours | 7 Days | 32580 | 8145 | 32% |
| Extended Stay Prediction at 72 Hours | 72 Hours | 7 Days | 19126 | 4782 | 54% |
| Extended Stay Prediction at 24 Hours | 24 Hours | 14 Days | 32580 | 8145 | 15% |
| Extended Stay Prediction at 72 Hours | 72 Hours | 14 Days | 19126 | 4782 | 25% |

Table 4.1: Summary of all tasks and their test sets.

From the Admission dataset, we extracted the following variables: **sex**, a categorical variable that describes the identified sex of the patient; **skin color**, a categorical variable that describes the identified skin color of the patient; **age**, a continuous variable that describes the patient's age and is calculated using birth date and admission date; and **hours admitted**, a continuous variable that describes the length of a patient's stay in the hospital and is calculated using the patient's admission date and discharge date. The sex and skin color variables are one-hot encoded and added to the input. The age variables are added as numerical values to the input. The hours admitted variable is not used as part of the input, but rather it is used to determine the true label for the LOS task.

From the Exam dataset, we extracted the following variables: **exam name/unit**, a categorical variable that describes the name of the exam that was taken by the patient alongside the unit used to measure it and is created by concatenating exam name and unit; **value**, a continuous variable which describes the value of the exam that the patient took; and **hours exam**, a continuous variable that describes the length of time between the patient's admission and the time the exam was taken and is calculated using admission date and exam date. The exam name/unit variable is used as a column name for a linear vector wherein each exam name/unit column is filled with the continuous variable. Only exam name/unit categories that appear in more than 1000 admissions are considered. This makes

for a total of 117 exam measurements for each patient. For patients who did not take an exam, its corresponding columns are filled with zero. The hours exam continuous variable is not used as part of the input, but rather to decide if a specific exam will be part of the input by comparing this variable with the cut-off time. Only exams with an "hours exam" variable less than or equal to the cut-off time are considered. The exam vector with only considered values is then added to the input.

From the Prescription dataset we extracted the following variables:

- **public**, a categorical variable that describes whether a patient is receiving public healthcare benefits;
- **surgical**, a categorical value that describes whether a patient has received surgical care;
- **intensive care**, a categorical value that describes whether a patient has been admitted to an intensive care unit;
- **obstetrics**, a categorical value that describes whether a patient has been admitted to an obstetrics unit; **emergency**, a categorical value that describes whether a patient has been admitted through emergency care;
- **ambulatory**, a categorical value that describes whether a patient is receiving pre scheduled care;
- **COVID-19**, a categorical value that describes whether a patient is admitted to a COVID-19 ward;
- **allergy**, a continuous value that describes how many of the medications the patient is receiving may cause them allergic reactions;
- **antibiotics**, a continuous value that describes how many of the medications the patient is receiving are antibiotics;
- **high alert**, a continuous value that describes how many of the medications the patient is receiving are high alert medications;
- **controlled**, a continuous value that describes how many of the medications the patient is receiving are controlled substances;
- **not default**, a continuous value that describes how many of the medications the patient is receiving are special kinds of medication;
- **tube**, a continuous value that describes how many of the medications the patient is receiving are administered intravenously;

- **different drugs**, a continuous value that describes how many of the medications the patient is receiving are new when compared to his last prescription; and
- **hours prescription**, a continuous variable that describes the length of time between the patient's admission and the time the prescription was made and is calculated using admission date and prescription date.

The public, surgical, intensive care, obstetrics, emergency, ambulatory, and COVID-19 variables are one-hot encoded and added to the input. The allergy, antibiotics, high alert, controlled, not default, tube, and different drugs variables are added as numerical values in the input. The hours admitted variable is added to the input, but rather, it is used to determine which prescription is the oldest prescription still under the cut-off time. Only the information of the oldest prescription is considered.

From the PrescriptionItem dataset, we extracted the following variables: **drug name**, the designation of the drug being administered, often very detailed, i.e. "CLORETO DE SODIO SOLUCAO INJETAVEL 0,9% 500ML (BOLSA/FRASCO)"; **daily frequency**, the frequency in which a patient takes the drug; normalized dosage, the dosage of the drug, normalized to ml rather than a full dose of the given medicine; and **drug notes**, a free-text string variable of varying size that encompasses all rows pertinent to its parent prescription. The table rows were turned into text using template-based Table-to-Text transformations. Each PrescriptionItem is associated with an entry in the Prescription dataset, so whether they are added to an input is decided by the associated prescription's **hours prescription** value.

From the Clinical Note dataset, we extracted the following variables: **note text**, a free-text string variable of varying size that encompasses a note written by hospital staff about the patient; and **hours note**, a continuous variable that describes the length of time between the patient's admission and the time the note was written and is calculated using admission date and note date. All notes created before the input cut-off were concatenated in the order in which they were created. The text data created from the PrescriptionItem dataset through Table-to-Text Generation is concatenated with the notes at the same time in the same way.

All of these transformations are summarized in Table 4.2.

## 4.3 Automated Dataset Annotation

For this task, we used the English-language MIMIC-III dataset. MIMIC-III is a freely available database of deidentified EHR data associated with 46,520 Intensive Care Unit (ICU) patients from the Beth Israel Deaconess Medical Center collected between 2001 and 2012 [50]. It is available through the PhysioNet platform [39]. This database contains

| Data | Raw Input Columns | Transformation | Processed Input Examples |
|---|---|---|---|
| Admission | sex, skin color | one-hot | [0 1 0 0 1 0 0 0] |
| | birth date, admission date | age calculation | [30.58] |
| Exam | exam name, exam metric, exam value | vectorization into values for each exam name/exam metric | [0.29 0 0 0.62 0 0 0 ...] (117 columns) |
| Clinical Note | note text | concatenation | Anotação tomada 0.55 hora(s) após início de internamento. Conteúdo da anotação: [note 1 content];\n\n Anotação tomada 2.10 hora(s) após início de internamento. Conteúdo da anotação:[note 2 content]; ... |
| Prescription | Public, Surgical, Intensive Care, Obstetrics, Emergency, Ambulatory, COVID-19 | binary | [0 1 1 0 0 1 0] |
| | Allergy, Antibiotics, High Alert, Controlled, Not Default, Tube, Different Drugs | None | [0 1 4 0 2 2 8] |
| PrescriptionItem | Drug Name, Daily Frequency, Dosage, Additional Notes | Table-to-Text | Nome da Droga: [drug name] — Frequência Diária: [daily frequency] — Dose: [normalized dosage] — Notas Adicionais: [drug notes] |

Table 4.2: Data processing from BRATECA columns for each of the five BRATECA datasets. Text is further processed into vectors using one of the three alternative methods discussed in Section 4.1.

55,452 discharge summary notes associated with 37,444 non-neonatal patients. The "Social History" topic in the dataset was identified by Ahsan et al. (2021) [2] as containing all relevant information related to Social Determinants of Health (SDoH) for patients.

We utilized Regular Expressions (RegEX) for identifying the Social History subsection from discharge summaries of MIMIC-III, excluding neonates and those with missing values. These extracted 44,566 social histories can be divided into two distinct datasets: a dataset consisting of 7,008 social histories annotated with SDoH classifications by human annotators from MIMIC-SBDH [2]; and an additional dataset comprising the remaining 37,558 unannotated social histories.

MIMIC-SBDH provides annotations for Social and Behavioral Determinants of Health (SBDH) to 7,025 randomly selected discharge summary notes from MIMIC-III. MIMIC-SBDH encompasses annotations for four categories of SDoH, namely: Community (further divided into Community-Present and Community-Absent), Education, Economics, and Environment. It also provides annotations for three categories of Behavioral Determinants of Health (BDoH): Alcohol Use, Tobacco Use, and Drug Use. For clarity, we consider all seven determinants to be SDoH and do not divide them into SDoH and SBoH as MIMIC-SBDH does. Using our RegEX-based method for extracting the "Social History" sections, we finally identified 7,008 social history entries. Additionally, MIMIC-SBDH includes associated keywords for each annotation. All annotated keywords are present in the "Social History" section of the discharge summary notes.

The SDoH-GPT system was applied to the three categories found by MIMIC-SBDH to have the highest lexical complexity [2]: Community, Economics, and Tobacco Use. Lexi-

cal complexity refers to the variety and richness of vocabulary within a category. Categories with higher lexical complexity feature a broader range of unique terms appeared in this category, which reflects higher diversity in the concepts being discussed. Furthermore, as a source of validation for these methods, we used two other datasets which were reasonably different in presentation from the text available in MIMIC-III — Suicide Reports [58] and Sleep Notes [91]. These datasets were selected to assess SDoH-GPT's performance across diverse clinical narratives, and in replicating our results with MIMIC-III using these datasets, we hoped to prove the effectiveness of our methods.

### 4.3.1 MIMIC SDoH

As mentioned, the SDoH-GPT system was applied to the three categories: Community, Economics, and Tobacco Use. Figure 4.1 demonstrates the data labeling changes we performed for our experiments with MIMIC-SBDH.

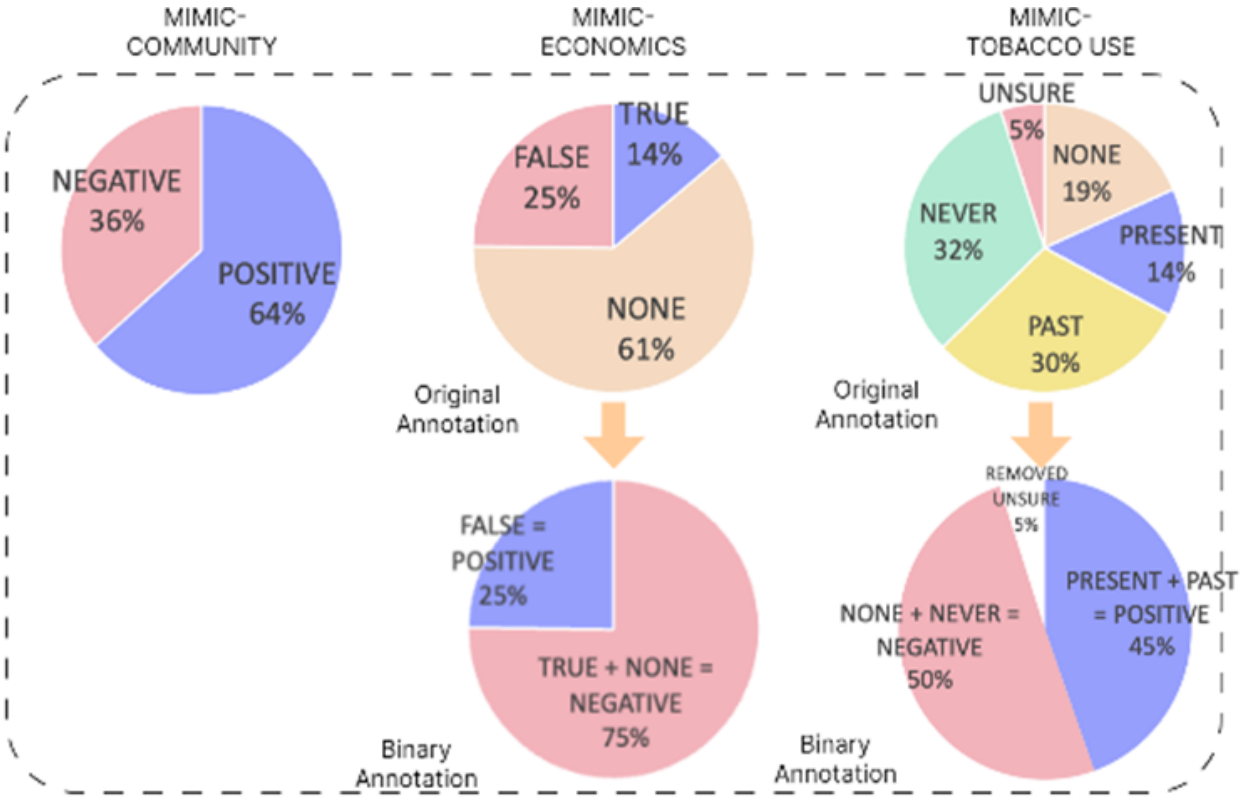



Figure 4.1: An overview of our relabeling of the MIMIC-SBDH corpus into a binary classification corpus for our tests with SDoH-GPT.

The **Community** category comprises two SDoH subcategories: Community-Present and Community-Absent. A discharge summary note received a True annotation for Community-Present if passages related to active social support, such as mentions of family or friends, were present in the social history and a False annotation if such passages were not present.

The dataset comprises 4,463 True-labeled data points and 2,562 False-labeled data points for Community-Present. Conversely, for Community-Absent, a True annotation was assigned if passages in the social history indicated a loss of social support, and a False annotation was assigned if none of the passages in the social history indicated a loss of social support. The data comprises 784 True-labeled data points and 6,241 False-labeled data points for Community-Absent. Notably, a discharge summary note can be simultaneously True for both categories, signifying instances where, for example, the social history references both the presence of a spouse (Community-Present True) and the recent loss of another family member (Community-Absent True).

Our investigation exclusively focuses on MIMIC-SBDH's Community-Present category and excludes considerations for Community-Absent annotations. Employing identical annotation criteria as specified by MIMIC-SBDH for Community-Present, we classified a discharge summary note as "Positive" if passages relating to active social support were present in the social history and "Negative" if such passages were absent. Consequently, our binary Community dataset comprises 4,463 positive data points and 2,562 negative data points.

Given the absence of a specific definition for "social support" in the MIMIC-SBDH paper, beyond referencing "a family member or friend," we adopted the definition from the American Psychological Association's Dictionary of Psychology. According to this source, social support is defined as "the provision of assistance or comfort to others, typically to help them cope with biological, psychological, and social stressors. Support may arise from any interpersonal relationship in an individual's social network involving family members, friends, neighbors, religious institutions, colleagues, caregivers, or support groups. It may take the form of practical help (e.g., doing chores, offering advice), tangible support that involves giving money or other direct material assistance, and emotional support that allows the individual to feel valued, accepted, and understood."

Following are two examples, one with a sentence positive for Community, and one with a sentence Negative for Community:

- Positive-indicative passage example: "The patient is married."
- Negative-indicative passage example: "The patient lives alone."

In the **Economics** category, a discharge summary note was labeled as True if the social history confirms the patient's current employment, as False if the patient is confirmed as unemployed or retired, and as None when there is no passage relating to employment status in the social history. This category includes 988 True-labeled data points, 1,742 False-labeled data points, and 4,295 None-labeled data points.

To facilitate binary categorization for our tests, we have restructured the MIMIC-SBDH Economics annotations into a new binary dataset. In this reorganized dataset, a

discharge summary note is marked as "Positive" if the social history indicates that the patient is unemployed or retired and "Negative" if the social history indicates employment or lacks any passage concerning the patient's employment status. Our reorganized binary Economics dataset comprises 1,742 positive data points and 5,283 negative data points.

Using the original False annotation for our Positive class may be unintuitive, but there are too few True-labeled data in the original dataset as we require at least 1,536 positive data points and 1,536 negative data points to create the testing and human-annotation training data sets, which are used later in the Method. Since the MIMIC-SBDH None-label must be part of the negative class, as it cannot be used to positively affirm anything, we had to use the more populated False-label as our positive class.

Following are two examples, one with a sentence positive for Economics, and one with a sentence Negative for Economics:

- Positive-indicative passage example: "The patient is a retired engineer."
- Negative-indicative passage example: "The patient is a secretary."

In the **Tobacco Use** category, a discharge summary note is labeled as "Present" if the social history confirms the patient's current tobacco use, "Past" if it indicates that the patient was previously a tobacco user but has since ceased, "Never" if it affirms that the patient has never used tobacco, "Unsure" if the passage is ambiguous and cannot be categorized as Present, Past, or Never, and "None" if the social history lacks any reference to tobacco use. The Tobacco Use dataset comprises 1,006 Present-labeled data points, 2,121 Past-labeled data points, 2,252 Never-labeled data points, 1,291 None-labeled data points, and 355 Unsure-labeled data points.

To facilitate binary categorization for our analyses, we have restructured the MIMIC-SBDH Tobacco Use annotations into a new binary dataset. In our reorganized dataset, a discharge summary note is labeled as "Positive" if the social history confirms the patient's current tobacco use or if the patient was previously a tobacco user but has since quit. Conversely, it is labeled as "Negative" if the social history indicates that the patient has never used tobacco or does not contain any passage regarding tobacco use. Our reorganized binary Tobacco Use dataset comprises 3,127 positive data points and 3,543 negative data points.

Following are two examples, one with a sentence positive for Tobacco Use, and one with a sentence Negative for Tobacco Use:

- Positive-indicative passage example: "The patient quit smoking 20 years ago."
- Negative-indicative passage example: "The patient has no smoking history."

### 4.3.2 Suicide Reports

The Suicide Reports dataset uses the National Violent Death Reporting System (NVDRS) dataset, which covers 500,072 incidents of suicide deaths across all 50 U.S. states, Puerto Rico, and the District of Columbia from 2003 to 2020. The research has been approved by the NVDRS Restricted Access Database proposal. Each incident in the NVDRS is documented with two death investigation notes, one from the Coroner or Medical Examiner (CME) perspective and the other from the Law Enforcement (LE) perspective.

For this study, we chose a **Job Problem** as a typical crisis. The Job Problem label is deemed true if, "at the time of the incident the victim was either experiencing a problem at work (such as tensions with a co-worker, poor performance reviews, increased pressure, feared layoff) or was having a problem with joblessness (e.g., recently laid off, having difficulty finding a job), and this appears to have contributed to the death". Within the NVDRS dataset, 30,525 incidents are labeled as positive for the Job Problem, and 469,547 are labeled as negative. For this study, 1,024 positive data points and 1,024 negative data points were randomly selected for training. Additionally, the testing dataset was also randomly sampled, comprising 512 positive data points and 512 negative data points, which is used to calculate all Cohen's kappa and AUROC measures presented in the Method. This standardized approach ensures consistency in the assessment of model performance. The data points used in the test datasets are not used in training or for any other purpose.

Following are two examples, one with a sentence positive for Job Problem, and one with a sentence Negative for Job Problem:

- Positive-indicative passage example: "The patient showed increasing concern about his job situation."
- Negative-indicative passage example: "The patient did not appear to be worried about his joblessness."

### 4.3.3 Sleep Notes

The Sleep Notes dataset uses an Alzheimer's Disease dataset (AD) collected by the University of Pittsburgh Medical Center (UPMC) between January 2016 and December 2020. The data was collected through the data service provided by the University of Pittsburgh Health Record Research Request (R3). The University of Pittsburgh's Institutional Review Board (IRB) reviewed and approved this study's protocol.

The dataset has a cohort of 7,266 patients associated with 379,120 clinical documents, 193,351 of which contained keywords related to sleep. A gold standard of 320 doc-

uments was annotated by clinicians, from which seven categories were identified: sleep apnea, napping, sleep problems, bad sleep quality, daytime sleepiness, night wakings, and sleep duration.

For this study, we chose **Sleep Apnea** as it contains the most positive labels, which is the minority class, with a total of 118 positive data points. We randomly selected 118 negative data points for 236 total data points, which we divided into a human-annotated training set for XGBoost and a testing set. The human-annotated training set comprises 37 positive data points and 37 negative data points, for a total of 74 data points. The testing set comprises 81 positive data points and 81 negative data points, for a total of 162 data points, which is used to calculate all Cohen's kappa and AUROC measures presented in the Method. This standardized approach ensures consistency in the assessment of model performance. The data points used in the test datasets are not used in training or for any other purpose.

Following are two examples, one with a sentence positive for Sleep Apnea, and one with a sentence Negative for Sleep Apnea:

- Positive-indicative passage example: "The patient showed signs of snoring during sleep."
- Negative-indicative passage example: "The patient has shown no sign of sleep apnea."

# 5. PATIENT FLOW

This chapter will detail our architectures and the results we achieved using them for the patient flow tasks detailed in Chapter 4. The Patient Flow tasks chosen were Admission Prediction and Extended Stay Prediction. The architectures chosen were one based on XGBoost and one based of DNNs.

## 5.1 Architectures

Our Patient Flow tasks were performed using architectures can be divided into those with an XGB base and those with an FCDNN base. Furthermore, both bases were tested using the three different ways to process text into homogeneous vectors described in Sections 4.1.2, 4.1.3, and 4.1.4. We also performed ablation studies, removing structured and text data from the input to observe how such changes affected the architectures' learning.

### 5.1.1 XGBoost Architectures

Our XGBoost architecture is straightforward, but proved very effective. We fed the data, which was vectorized as explained in Chapter 4, to the XGBoost model as implemented by Sci-Kit learn [15]. The model was set to predict binary classification, and the seed is set to 20 to ensure the model always initializes with the same parameters, but no other parameters are tuned. We performed ablation studies removing either the structured data or the text data to see how these new parameters affected the model's learning. The data pipeline and architectures are explained in Figure 5.1.

### 5.1.2 Neural Network Architecture

Though we made attempts to use many of the techniques explored in Section 2.2.1, we did not find that our data was able to take advantage of CNNs or attention mechanisms at all, only ever lowering the efficacy of our architectures. We believe that this is because, unlike MIMIC, BRATECA's data is too scarce and, as such, difficult to fit with these models. Such architectures, which were leveraged for their ability to parse temporal data more effectively in the literature, proved not at all effective in an environment with a comparative dearth of sequential information outside of textual data, even in the Admission tasks, which have the most training data.

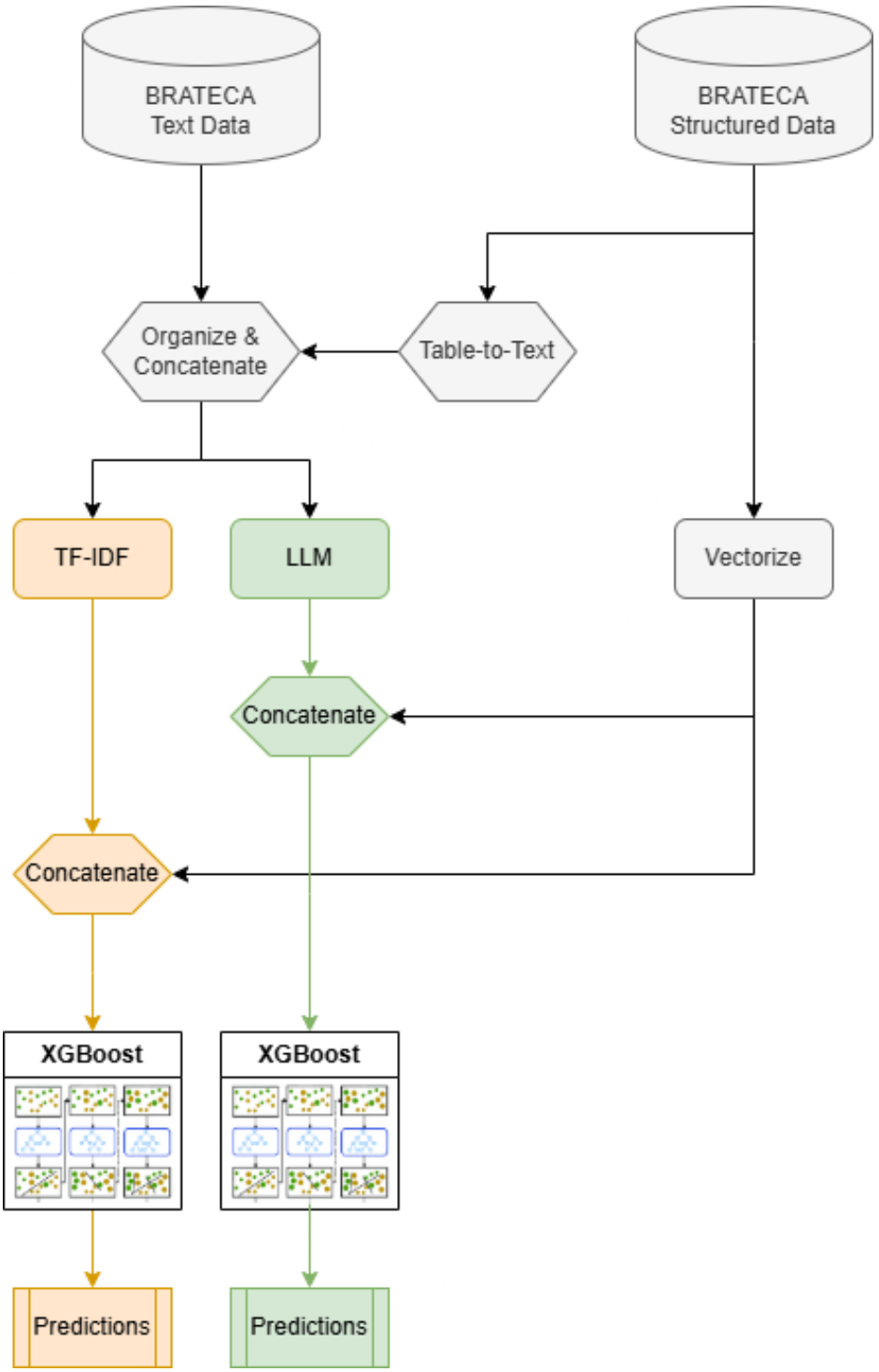



Figure 5.1: An overview of the XGBoost Patient Flow architectures.

Because of this, our architecture leverages mostly the predictive learning power of FCDNN, which ends with a sigmoid predictive output, as this model proved most effective when parsing through the vectorized structured data alongside the different ways in which we explored text data processing. We performed ablation studies removing either the structured data or the text data to see how these new parameters affected the model's learning. The data pipeline and architectures are explained in Figure 5.2.

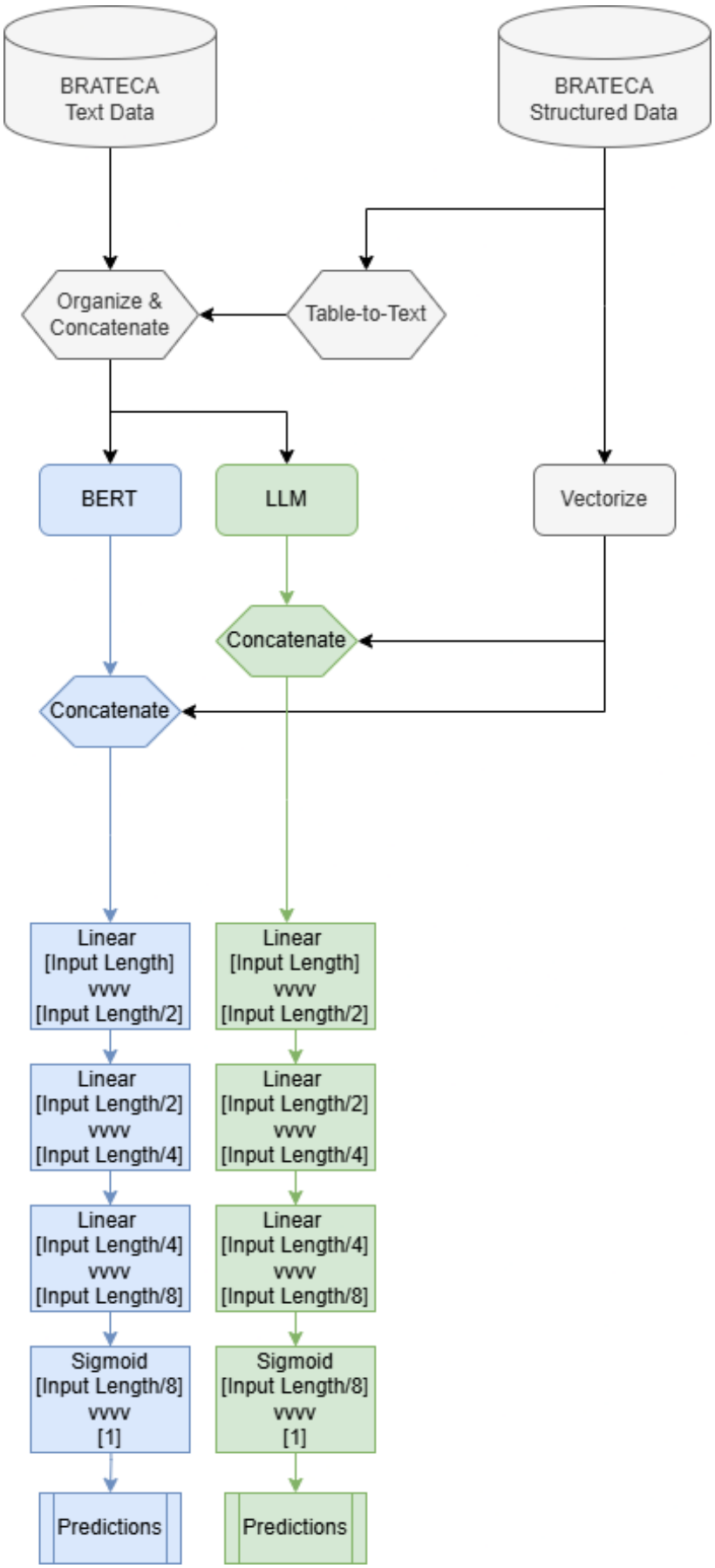



Figure 5.2: An overview of the FCDNN Patient Flow architectures.

## 5.2 Patient Flow Results

The results for all six patient flow tasks, including both Admission prediction and the four Extended Stay prediction tasks are presented in this section. The section ends with a discussion of the results and an analysis of what can be gleaned from them.

We used two core architectures for our models in the following manner: XGB, the classic machine learning architecture using XGBoost as its core; and NN, the neural network architecture based around FCDNN layers. Each of the core architectures had to use one of the three following modules to encode textual data into numerical information vectors for training: TF, a Term Frequency-Inverse Document Frequency vectorizer; BERT, a BERT transformer followed by an LSTM layer for text embedding and vectorizing; and LLM, a LLaMa-3 8B language model used to encode the text data into a numerical information vector. We will present the following combinations of core architectures and text encoding modules: XGB-TF, XGB-LLM, NN-BERT, and NN-LLM.

We will also show an ablation study for the best performing combination for each core architecture: XGB-TF and NN-BERT. The ablation studies include: STRUCT, which only uses structured table data; and TEXT, which uses encoded text data. The main results are presented in Table 5.1, and the ablation results are presented in Table 5.2.

Our tables will present a weighted average between positive and negative Precision, Recall, and F1. We will also present the confusion matrices for each of the results so we may further reason over the results we achieved.

### 5.2.1 Main Architecture Results

As we can see in Table 5.1, the best results were achieved by the XGBoost model with TF-IDF text encoding. This simpler model achieved convincingly better results than the other model/text encoder combinations except in the 7-Day Extended Stay Prediction at 72 Hours task. As seen in Figures 5.3 and 5.4, however, the confusion matrices show that none of the architectures were able to learn the 14-day tasks very well, regardless of model.

We conjecture that TF achieved the best results despite being so simple because it does not need previous training to encode the language, relying instead on lexical features rather than attempting to map semantic meaning, as the BERT and Llama-3.1 language models do. This means that the highly specialized jargon written in a sort of personalized shorthand by medical professionals for their own later use that was often seen in the BRATECA clinical notes did not affect TF-IDF in the same way it affected the ostensibly more powerful models. In more effectively mapping the lexical features of words that the

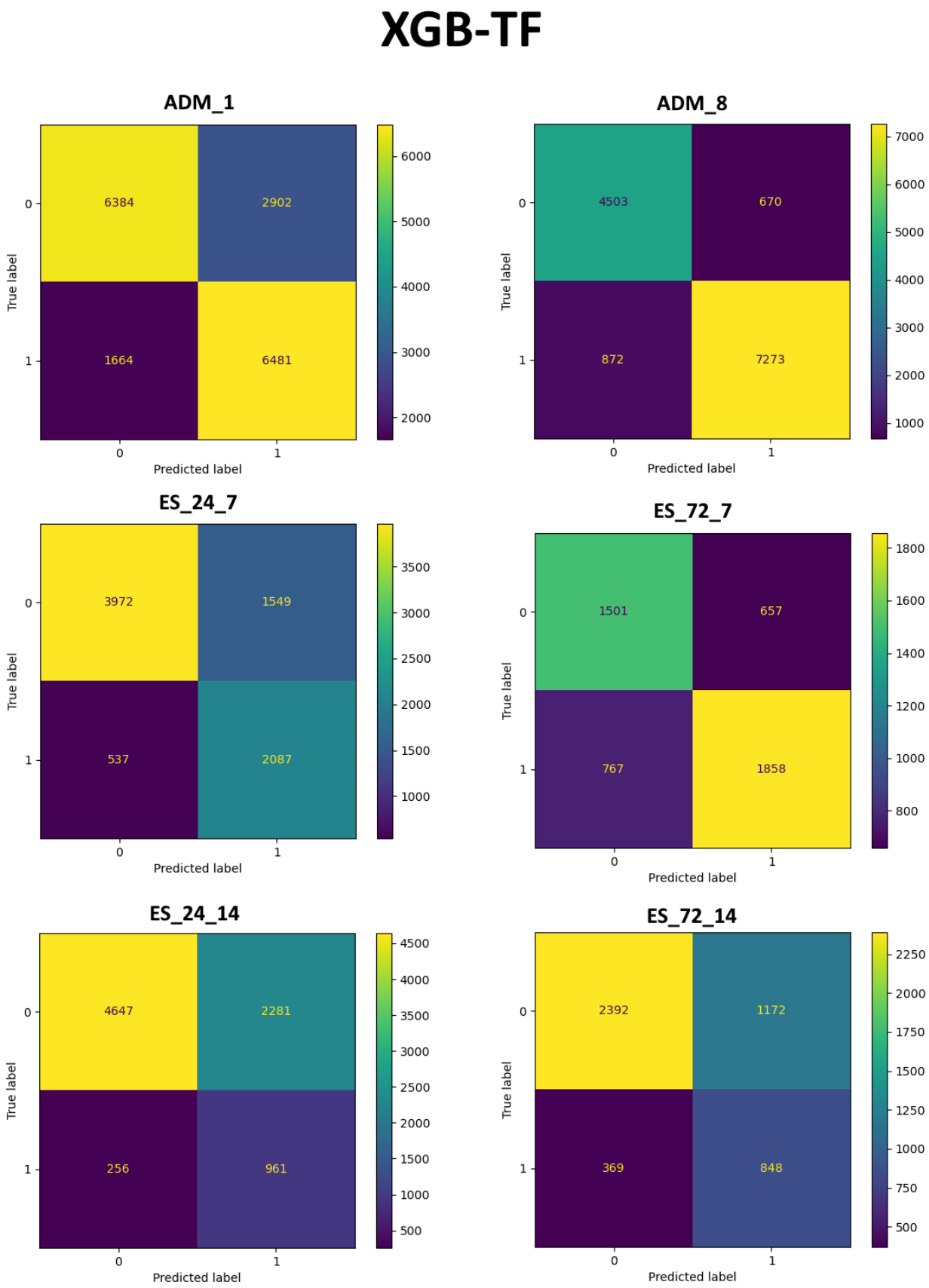



Figure 5.3: Confusion matrices for each task from the XGB-TF model.

language models were simply not trained to map very well given the very specific context and semantics, the TF-IDF model was able to slightly outperform them.

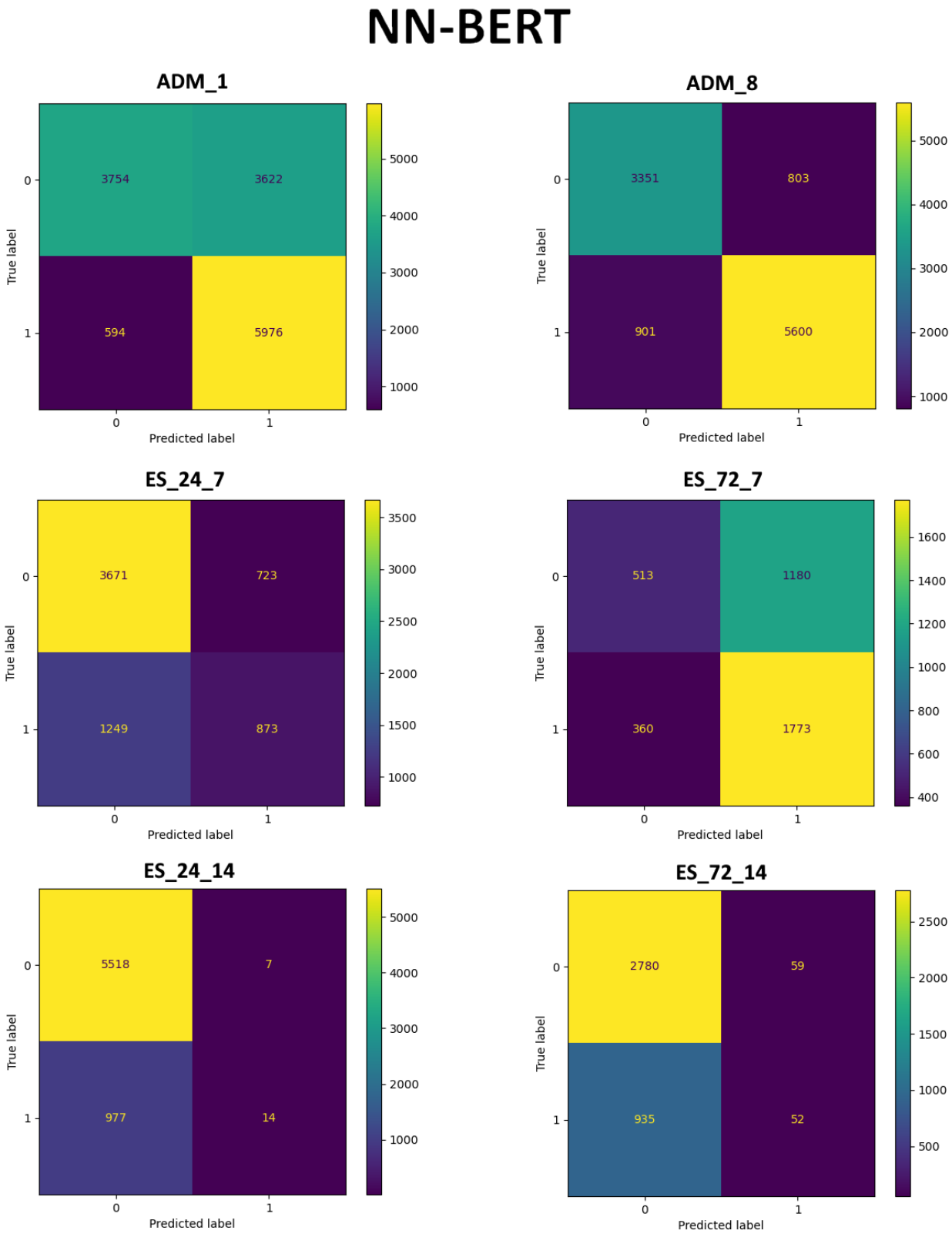



Figure 5.4: Confusion matrices for each task from the NN-BERT model.

Moreover, we believe that BERT managed to outperform Llama-3.1 because our chosen BERT model, BioBERT-pt, was trained on some Brazilian clinical notes and, as such, could encode meaning from the text better than the LLM, which was trained mainly for

use in English, and without a focus medical data. It was, however, unable to achieve better results than TF-IDF since the text is still quite complex.

As for how the simpler XGBoost architecture managed to outperform the more nuanced NN architecture, this is likely because XGBoost can learn better with fewer examples. As seen in Table 4.1 from Section 4.2, the two task datasets were the Admission Prediction sets, with $\tilde{6}$9,000 examples for prediction at one hour and $\tilde{5}$3,000 examples for prediction at eight hours, and these two achieved the closest results to XGB-TF's results. We believe that with more data, and perhaps more varieties of data, the NN architecture would overtake XGB performance-wise.

| Arch. | Task | Acc. | Prec. | Rec. | F1 | Arch. | Task | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **XGB-TF** | **ADM_1** | **0.74** | **0.74** | **0.74** | **0.73** | NN-BERT | ADM_1 | 0.72 | 0.72 | 0.72 | 0.72 |
| | **ADM_8** | **0.88** | **0.88** | **0.88** | **0.88** | | ADM_8 | 0.80 | 0.84 | 0.84 | 0.84 |
| | **ES_24_7** | **0.73** | **0.78** | **0.73** | **0.75** | | ES_24_7 | 0.70 | 0.68 | 0.70 | 0.68 |
| | **ES_72_7** | **0.70** | **0.70** | **0.70** | **0.70** | | ES_72_7 | 0.60 | 0.59 | 0.60 | 0.57 |
| | ES_24_14 | 0.69 | 0.85 | 0.69 | 0.73 | | **ES_24_14** | **0.85** | **0.82** | **0.85** | **0.78** |
| | **ES_72_14** | **0.67** | **0.74** | **0.67** | **0.69** | | ES_72_14 | 0.74 | 0.68 | 0.74 | 0.65 |
| XGB-LLM | **ADM_1** | 0.73 | **0.74** | 0.73 | **0.73** | NN-LLM | ADM_1 | 0.71 | 0.71 | 0.71 | 0.71 |
| | ADM_8 | 0.86 | 0.86 | 0.86 | 0.86 | | ADM_8 | 0.78 | 0.81 | 0.78 | 0.76 |
| | ES_24_7 | 0.70 | 0.74 | 0.70 | 0.71 | | ES_24_7 | 0.68 | 0.65 | 0.68 | 0.64 |
| | ES_72_7 | 0.65 | 0.65 | 0.65 | 0.65 | | ES_72_7 | 0.60 | 0.62 | 0.60 | 0.55 |
| | ES_24_14 | 0.67 | 0.83 | 0.67 | 0.71 | | **ES_24_14** | **0.85** | **0.83** | **0.85** | **0.78** |
| | ES_72_14 | 0.62 | 0.71 | 0.62 | 0.65 | | ES_72_14 | 0.74 | 0.79 | 0.74 | 0.64 |

Table 5.1: Results for each of the main patient flow architectures on the following tasks: ADM_1 (Admission Prediction at 1 Hour); ADM_8 (Admission Prediction at 8 Hours); ES_24_7 (7-Day Extended Stay Prediction at 24 Hour); ES_72_7 (7-Day Extended Stay Prediction at 72 Hour); ES_24_14 (14-Day Extended Stay Prediction at 24 Hour); ES_72_7 (7-Day Extended Stay Prediction at 72 Hour). The Precision, Recall, and F1 are weighted.

### 5.2.2 Ablation Studies

From the ablation studies, we can see that even though there is much less structured data than text data, the models still benefit from its addition, even if slightly. The text by itself showed better results in general, but this was expected. Most interesting is how much the NN architectures benefited from the holistic use of the data. The Admission Prediction at 8 Hours task, for example, achieved an F1 of 0.72 with only structured information and 0.70 with only text information but achieved a result of 0.84 with both. This is an impressive result, showing that our holistic use of data is indeed worth pursuing.

| Arch. | Task | Acc. | Prec. | Rec. | F1 | Arch. | Task | Acc. | Prec. | Rec. | F1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XGB-TF STRUCT | ADM_1 | 0.68 | 0.72 | 0.68 | 0.67 | NN-BERT STRUCT | ADM_1 | 0.60 | 0.62 | 0.60 | 0.59 |
| | ADM_8 | 0.83 | 0.83 | 0.83 | 0.83 | | ADM_8 | 0.73 | 0.72 | 0.73 | 0.72 |
| | ES_24_7 | 0.69 | 0.73 | 0.69 | 0.70 | | ES_24_7 | 0.58 | 0.55 | 0.58 | 0.56 |
| | ES_72_7 | 0.64 | 0.65 | 0.64 | 0.64 | | ES_72_7 | 0.56 | 0.55 | 0.56 | 0.52 |
| | ES_24_14 | 0.66 | 0.84 | 0.66 | 0.71 | | ES_24_14 | 0.62 | 0.78 | 0.62 | 0.67 |
| | ES_72_14 | 0.64 | 0.71 | 0.63 | 0.66 | | ES_72_14 | 0.60 | 0.59 | 0.60 | 0.60 |
| XGB-TF TEXT | **ADM_1** | **0.70** | **0.75** | **0.70** | **0.69** | NN-BERT TEXT | ADM_1 | 0.70 | 0.75 | 0.70 | 0.69 |
| | **ADM_8** | **0.87** | **0.87** | **0.87** | **0.87** | | ADM_8 | 0.70 | 0.73 | 0.70 | 0.70 |
| | **ES_24_7** | **0.73** | **0.78** | **0.73** | **0.74** | | ES_24_7 | 0.56 | 0.55 | 0.56 | 0.56 |
| | **ES_72_7** | **0.69** | **0.70** | **0.69** | **0.69** | | ES_72_7 | 0.44 | 0.47 | 0.44 | 0.41 |
| | ES_24_14 | 0.68 | 0.84 | 0.68 | 0.73 | | **ES_24_14** | **0.85** | **0.82** | **0.85** | **0.78** |
| | **ES_72_14** | **0.66** | **0.74** | **0.66** | **0.68** | | ES_72_14 | 0.55 | 0.67 | 0.54 | 0.56 |

Table 5.2: Ablation studies on the best performing architectures XGB-TF and NN-BERT. Results are on the following test sets: ADM_1 (Admission Prediction at 1 Hour); ADM_8 (Admission Prediction at 8 Hours); ES_24_7 (7-Day Extended Stay Prediction at 24 Hour); ES_72_7 (7-Day Extended Stay Prediction at 72 Hour); ES_24_14 (14-Day Extended Stay Prediction at 24 Hour); ES_72_7 (7-Day Extended Stay Prediction at 72 Hour);. The Precision, Recall, and F1 are weighted.

### 5.2.3 Discussion

We see that Patient Flow tasks are very difficult to solve, but some of our architectures performed quite well for several tasks. Given the relative lack of information compared to the usual source material for testing data, the MIMIC collections, which have minute by minute vital sign readings, standardized exam information, and very extensive pre hospital, post hospital, and during-stay note taking, among other details, our data engineering and learning models performed well.

Our efforts reveal the strengths of simple architectures, as well as some weaknesses of more complex architectures for the available data. We can also see some limitations to our data and methods when it comes to attempting to predict events too far into the future. The first thing we should acknowledge is the excellent performance of both methods for the Admission Prediction by 8 Hours task, with the highest accuracy and F1 achieved by the XGB-Term Frequency architecture.

We also observed that, regardless of architecture, using available data holistically resulted in improved performance for the architecture. We performed a study on the reasoning behind the predictions of the XGB architecture on the Admission Prediction tasks using SHAP and found that while the text data is indeed very important, as the ablation results bear out, all of the structured data is also important to predicting patient flow. For Admission Prediction by 1 Hour, for example, we found that whether or not a patient was an Emergency patient was very important to prediction. Seeing as there is a general dearth of data at one hour, the model found that an emergency patient was much more likely to stay longer than 24 hours. For Admission Prediction at 8 Hours, on the other hand, the pres-

ence of surgical events overtakes whether the patient's visit started as an emergency. This is an example of a logical explanation that can be given to medical professionals to engender trust in such systems as well, and is pictured in Figure 5.5.

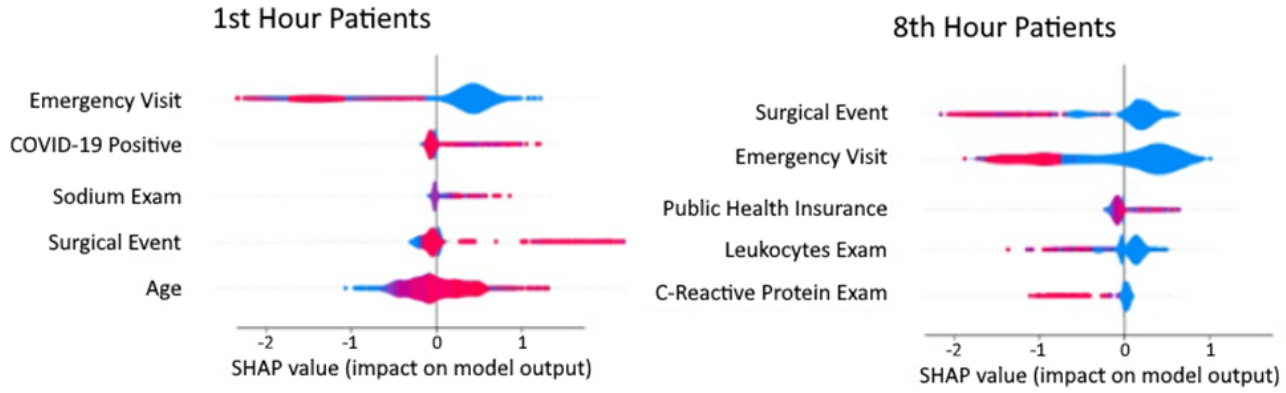



Figure 5.5: SHAP beeswarm plots for the five most important features for each kind of patient. Blue indicates a low value or False (i.e., blue Surgical Event means the patient did not have a Surgical Event while a blue age means a lower age. Red indicates a high value or True. A widening of the line means more patients at that level of effect, and the closer to the edges, the more significant the impact. Negative impact indicates that the patient is more likely to leave, while positive impact indicates that the patient is more likely to be admitted.

It is notable that the XGB-Term Frequency architecture achieves the best scores for all tasks we could consider "successful": the Admission tasks and the 7-day Stay prediction tasks. This shows that classic machine learning can still outperform more complex architectures, especially in spaces with little training data and, we believe, most importantly, very noisy data. The textual data, in particular, is at once informative but written in very particular shorthand. As previously mentioned, we believe the BERT language models and Llama-3.1 couldn't quite extract more meaningful information than the simpler term frequency vectorizer. We have chosen three clinical notes, carefully redacted them as appropriate, and present snippets of them here as examples of the shorthand used in the notes in Table 5.3.

It is easy to see why language models would have trouble with these, especially smaller ones for trained for the Portuguese language or even larger ones that are not specialized for Multilingual or Portuguese writing, such as Llama-3.1. It is also true that BERT is limited by its rather short input length of 512 tokens. This resulted in especially rough truncations in the longer clinical note compilations fed to the models for tasks with 24 and 72 hours of input information. This was no problem for the TF-Vectorizer or LLaMa-3, as the former is not transformer-based, and the latter has a massive token input limit. LLaMa still seems to be limited by the jargon and the general semantic complexity of the notes.

Beyond that, there were some tasks that we could consider "unsuccessful" in that their results were not satisfactory. The 14-day stay prediction task, in particular, proved difficult to train for both classic and neural network based architectures. The confusion ma-

| Example Clinical Note |
| --- |
| Conteúdo da anotação: ###CIT#### PO de cirurgia cardíaca - troca de valvula metálica por biológica (aórtica e mitral) em *** com Dr ***# Nega alergias# Em uso de: Furosemida 20mg , AAS 100mg , Omeprazol 20mg , Tramadol 12 / 12h , Paco 08 / 08hHDAX: Paciente refere quadro de dor torácica posterior intensa acompanhado de dispneia , edema em membros inferioresX (mais em MIE) [...] |
| - NEGA COMORBIDADES - NEGA ALERGIAS DOR EPIGÁSITRCA INICIO ESSA NOITE, AGORA NA MADRUGADA, FORTE INTENSIDADE, IRRADIAÇÃO PARA DORSO, FORMIGAMENTO NAS MÃOS. DOR LEVE AGORA. NEGA OUTRAS QUEIXAS. BEG AP: MV+, SEM RA AC; 2T, RR AB; RHA+, DOR EPIGÁSTRICAX, SEM IRRITAÇÃO PERITONEAL EXT: PERFUNDIDAS DOR ABDOMINAL, GASTRITE? ECO, LABS, EQU, ANALGESIA |
| BACTERIÚRIA: VIDE NOTA; BILIRRUBINA: NEGATIVO; CILINDROS: AUSENTES; CORPOS CETÔNICOS: NEGATIVO; CRISTAIS: AUSENTES; CÉLULAS EPITELIAIS ESCAMOSAS: RARAS; CÉLULAS TRANSICIONAIS: AUSENTES; CÉLULAS TUBULARES RENAIS: AUSENTES; ESTERASE LEUCOCITÁRIA: NEGATIVO; GLICOSE: NEGATIVO; HEMOGLOBINA: NEGATIVO; LEVEDURAS: AUSENTES; MUCO: AUSENTES; NITRITO: NEGATIVO; OBSERVAÇÃO:PRESENÇA DE GRÂNULOS DE FOSFATO AMORFO NA AMOSTRA EXAMINADA. AMOSTRA ENVIADA AOLABORATÓRIO.; PROTEÍNAS (ALBUMINA): NEGATIVO; UROBILINOGÊNIO: NORMAL; |

Table 5.3: Examples of clinical notes.

trices show that the models were seemingly unable to learn useful patterns from the data. This may have been caused by the relative dearth of examples for these tasks in particular, as well as the apparent higher difficulty of the task.

It is also notable that each medical professional has different preferences when it comes to discharging patients, and different hospitals have slightly different protocols when it comes to the same. This means that, beyond their physical readiness to leave, there is a level of arbitrariness to the data that further impedes pattern learning. This perhaps speaks to the imprudence of using such systems in any given hospital that was not trained with data from that same hospital, as these details may skew results one way or the other. It also implies a new avenue of research into predicting exaggerated stays, i.e., when a patient is kept in hospital longer than necessary. This is, of course, a topic that must be approached more delicately and with more oversight from medical professionals, given the dangers of early discharge.

Ultimately, however, the relative success of the classic XGB architecture is an important result. With it, we have shown a very cheap, easy to implement, method of acquiring meaningful predictions which can be acted upon with some certainty by hospital administration staff. The low computational requirements are a massive plus when thinking about adding more stress onto important hospital computer systems, and the cheap buy-in can entice administrators to adopt admission and extended stay prediction systems as auxiliary systems to help with duties such as bed allocation and cost savings.

# 6. AUTOMATED DATASET ANNOTATION

The Automated Dataset Annotation task is subdivided into the prompts used to create the annotated training sets. Five different prompts were developed and tested with our SDoH-GPT architecture: one Zero-Shot prompt and four Two-Shot prompts. They were used to automatically annotate one training set from the categories that were selected from each data collection. The training sets were then used to train an XGB model to predict the answer for each of the tasks mentioned in Section 4.3.

We used the GPT-3.5 LLM as a base model for our prompting strategy. This LLM was used to refine our corpus into a usable silver annotated test set, as shown by the successful training of XGBoost models on the task of Social Determinant of Health Extraction from MIMIC-III discharge summaries. Figure 6.1 presents the SDoH-GPT architecture.

## 6.1 Detailing our Prompting Strategy

Developing a straightforward Zero-Shot prompt is the first step of our annotation system, which we have named SDoH-GPT. GPT-3.5 was employed using the Azure OpenAI (See Supplementary Material for more model parameter settings). This prompt is composed of a set of three instructions and a query. The instructions are as follows: a succinct roleplaying instruction designed to contextualize the GPT model, a General Task Instruction that explains the task, and SDoH Specific Instruction that explicitly states which kinds of information must be extracted.

To facilitate straightforward responses from GPT-3.5, our queries were formulated as Yes/No questions. As an illustration, consider the query structure for Economics prompts: "Does the social history indicate that the patient is currently unemployed or retired? Answer with yes or no as the first word." This format prompts GPT-3.5 to respond with either "Yes" or "No." The results were categorized into True Positive and True Negative groups based on gold standard human annotations. This process continued until True Positive and True Negative Zero-Shot prompt-annotated sample groups each comprised a minimum of 50 samples. These new balanced 100 LLM annotations are called the Prototype Annotated Dataset, which is further categorized into four distinct groups by employing our 0-Shot SDoH-GPT: True Positives (positive samples correctly categorized by GPT-3.5), True Negatives (negative samples correctly categorized by GPT-3.5), False Positives (positive samples incorrectly categorized by GPT-3.5), and False Negatives (negative samples incorrectly categorized by GPT-3.5).

A randomly selected single sample from each group, together with its social history, human-annotated label, and explanation, are systematically organized into a Shot.

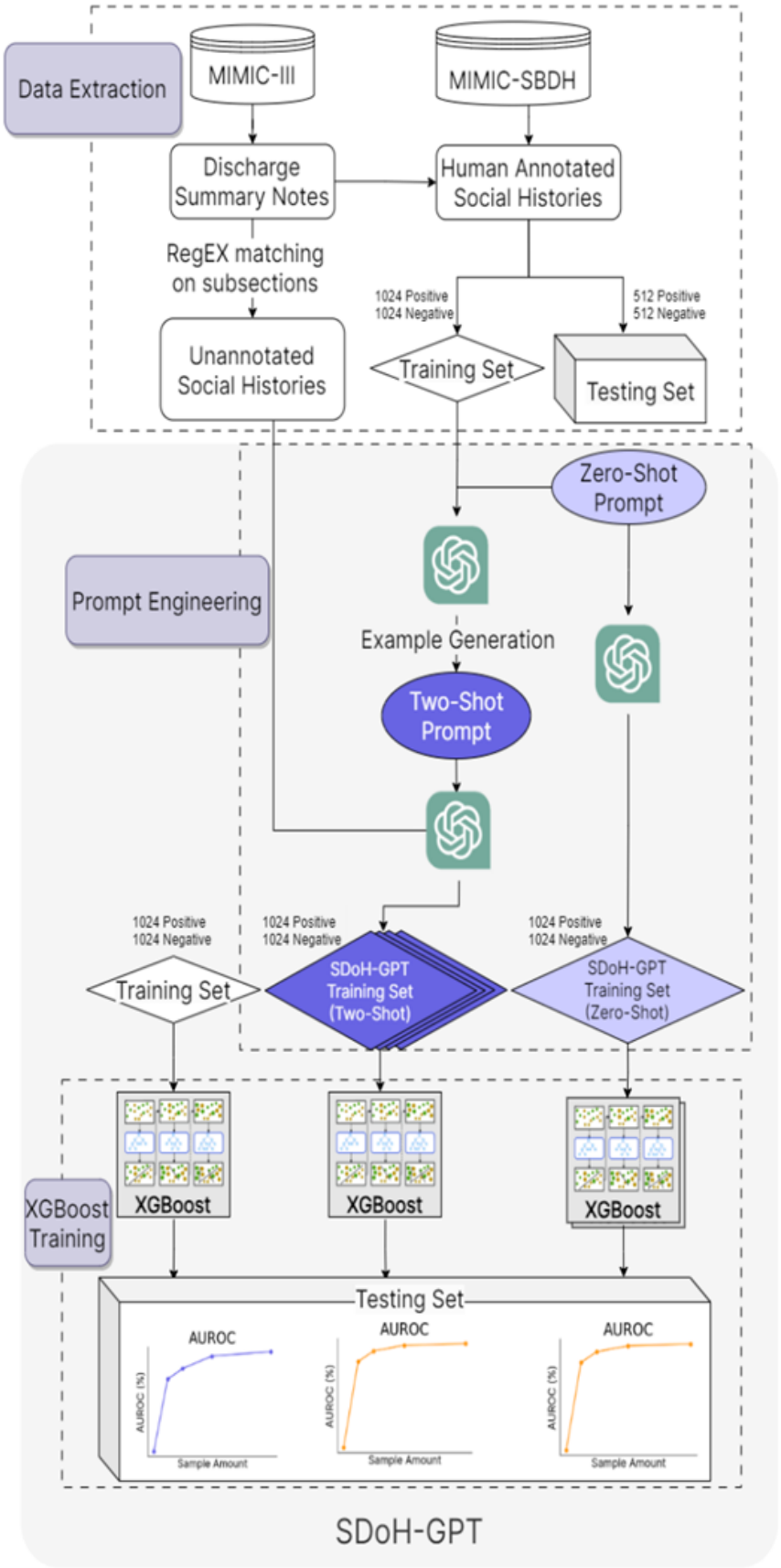



Figure 6.1: An overview of SDoH-GPT.

The explanation refers to the human-created reasoning for the gold standard labeling as-

sociated with each sample, which was used to create natural language explanations that detailed why each sample was labeled as positive or negative. Four kinds of Two Shots are generated: the Easy pair (E), consisting of one True Positive example and one True Negative example (i.e., 2-Shot E); the Easy-Explained pair (E+Expl), mirroring the examples in the Easy pair, but adding explanations (i.e., 2-Shot E+Expl); the Hard pair (H), comprising one False Negative example and one False Positive example (i.e., 2-Shot H); and the Hard-Explained pair (H+Expl), replicating the examples in the Hard pair, but including explanations (i.e., 2-Shot H+Expl).

## 6.2 Detailing our XGBoost Model Training

We used two kinds of training datasets to train XGBoost classifiers: Human annotated training data from MIMIC-SDBH and SDoH-GPT-annotated training data. The SDoH-GPT-annotated training datasets were created using the Zero-Shot and four kinds of Two-Shot prompts. Each prompt was employed to annotate a balanced training dataset, for a total of five SDoH-GPT training datasets. These six training sets have 1024 positive and 1024 negative examples for each MIMIC SDoH category: Human-Annotated Training Set (From MIMIC-SBDH), 0-Shot SDoH-GPT Training Set, 2-Shot E SDoH-GPT Training Set, 2-Shot E+Expl SDoH-GPT Training Set, 2-Shot H SDoH-GPT Training Set, and 2-Shot H+Expl SDoH-GPT Training Set. The XGBoost classifier trained on human annotations is called XGBoost-Human, while the XGBoost classifier trained on SDoH-GPT annotations is called XGBoost-SDoH-GPT.

The input data for XGBoost is a 3000-integer array representing word frequencies in social history samples. The top 3000 words, excluding stop words, are selected from each training dataset using SciKit Learn's "CountVectorizer" [73] and NLTK's stop word list. Six balanced training sets of 2048 samples for each SDoH category were further sub-sampled to smaller balanced datasets with 16, 32, 64, 128, 256, 512, 1024, and 2048 samples. In total, 48 XGBoost models were trained and tested on a 1024-sample balanced dataset.

## 6.3 Results

The results for the Automated Dataset Annotation task are presented within this section, including all permutations of the test. The AUROCs for all five categories by number of examples is summarized in Figure 6.2. The section ends with a discussion of the results and an analysis of what can be gleaned from them.

We primarily evaluated SDoH-GPT on MIMIC-III discharge summaries to classify the top three most lexically complex SDoH categories: Community, Economics, and To-

bacco Use. Secondarily, we validated the approach using two similar but differently formatted data sets: Sleep Notes and Suicide Reports. These validation sets were processed using specific prompts to identify mentions of sleep apnea in the sleep notes or associations with job problems in the suicide report. A binary 'Yes' or 'No' classifications balanced dataset was created through random sampling. The results to be discussed will include the accuracy of annotation, as well as the estimated cost of annotation in both time and price.

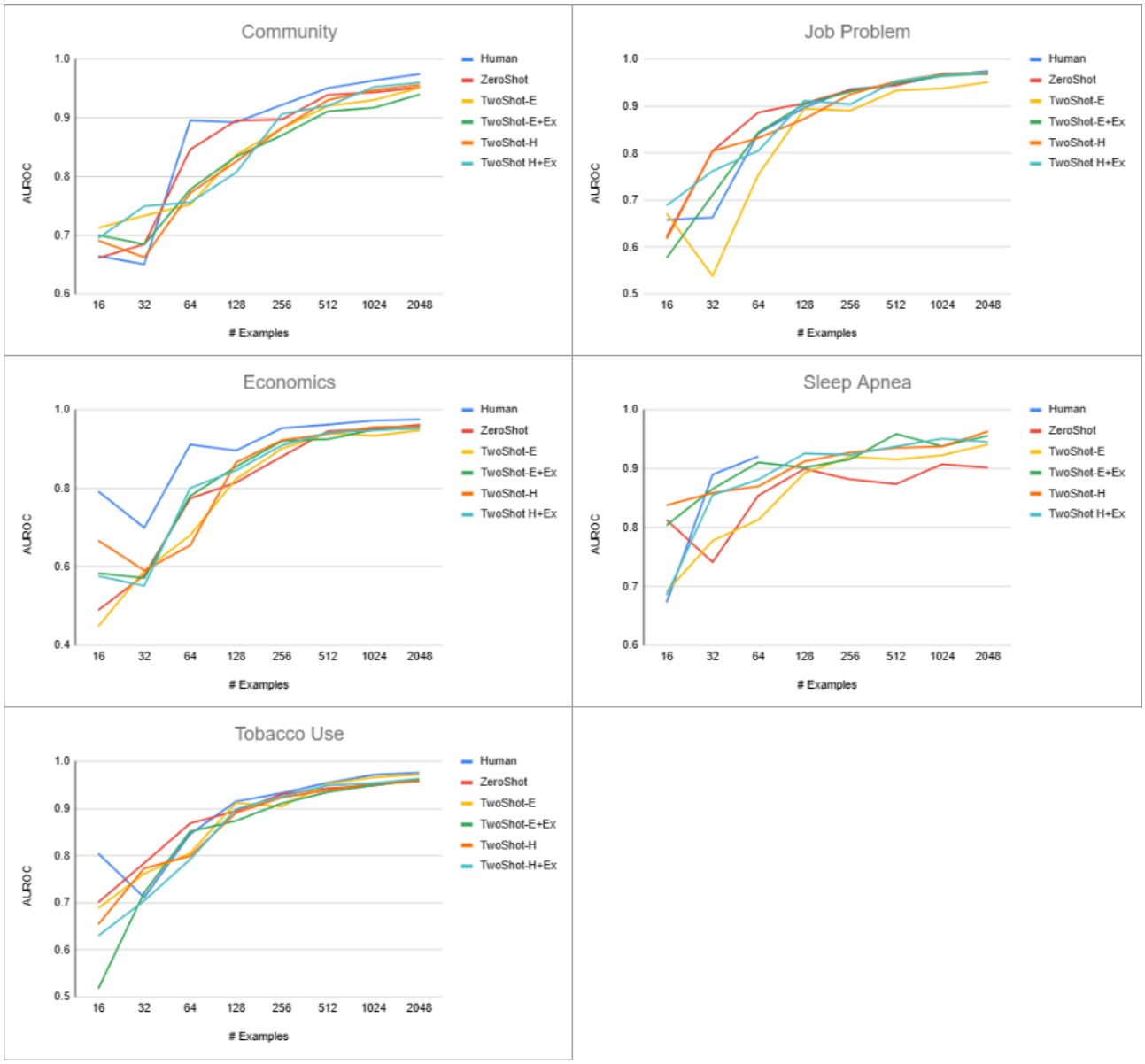



Figure 6.2: AUROC by number of examples for all 5 task categories.

### 6.3.1 MIMIC-III Discharge Summaries

The best-performing configurations, determined based on AUROC scores, were 2-Shot H+Expl for Community, 0-Shot SDoH-GPT for Economics, and 2-Shot E for Tobacco

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Community | 16 | 0.6646 | 0.6614 | 0.7128 | 0.7002 | 0.691 | 0.6952 |
| | 32 | 0.6506 | 0.6848 | 0.7336 | 0.6847 | 0.6625 | 0.7498 |
| | 64 | 0.8961 | 0.8465 | 0.7524 | 0.7783 | 0.7723 | 0.7561 |
| | 128 | 0.893 | 0.8959 | 0.8375 | 0.8343 | 0.8259 | 0.8072 |
| | 256 | 0.9227 | 0.8976 | 0.883 | 0.8708 | 0.883 | 0.9072 |
| | 512 | 0.9512 | 0.9398 | 0.9207 | 0.9117 | 0.9308 | 0.921 |
| | 1024 | 0.9642 | 0.944 | 0.9308 | 0.9176 | 0.9476 | 0.9531 |
| | 2048 | 0.9752 | 0.952 | 0.9518 | 0.9402 | 0.9565 | 0.9608 |
| Economics | 16 | 0.7926 | 0.4898 | 0.4489 | 0.5841 | 0.6671 | 0.5765 |
| | 32 | 0.6999 | 0.5794 | 0.5908 | 0.5715 | 0.5909 | 0.5519 |
| | 64 | 0.9122 | 0.7748 | 0.6811 | 0.7812 | 0.6552 | 0.8013 |
| | 128 | 0.8968 | 0.8154 | 0.8247 | 0.8558 | 0.8668 | 0.8463 |
| | 256 | 0.9543 | 0.8819 | 0.9018 | 0.9212 | 0.9235 | 0.9103 |
| | 512 | 0.9631 | 0.9465 | 0.9411 | 0.926 | 0.9397 | 0.9427 |
| | 1024 | 0.9732 | 0.9522 | 0.9348 | 0.95 | 0.9568 | 0.9481 |
| | 2048 | 0.9762 | 0.9626 | 0.9482 | 0.9538 | 0.9592 | 0.9553 |
| Tobacco Use | 16 | 0.8047 | 0.7012 | 0.6886 | 0.5179 | 0.6545 | 0.6302 |
| | 32 | 0.7116 | 0.7846 | 0.7621 | 0.7219 | 0.7735 | 0.7047 |
| | 64 | 0.8459 | 0.8689 | 0.8055 | 0.8517 | 0.7995 | 0.7922 |
| | 128 | 0.9154 | 0.8952 | 0.9123 | 0.8744 | 0.8911 | 0.8994 |
| | 256 | 0.9334 | 0.931 | 0.9048 | 0.9116 | 0.9242 | 0.9259 |
| | 512 | 0.9553 | 0.9436 | 0.9534 | 0.9351 | 0.9393 | 0.95 |
| | 1024 | 0.9723 | 0.9493 | 0.967 | 0.9504 | 0.9539 | 0.9543 |
| | 2048 | 0.9768 | 0.9615 | 0.9728 | 0.9601 | 0.958 | 0.9641 |

Table 6.1: Performance of XGBoost models trained on human annotations and automated SDoH-GPT annotations for the three MIMIC-SBDH categories. This table shows AUROC results, measuring the correctness of annotation.

Use. Given the high expense and complexity inherent in human annotation, related studies typically have a limited number of SDoH annotations: 1,000 [8], 1,576 [57] and 500 [105]. As seen in Table 6.4, assuming only 512 human annotations are available for Community, 2-Shot H+Expl SDoH-GPT can efficiently generate an additional 2,048 annotations at approximately one-fifth of the cost and one-third of the time required for human annotation. Furthermore, given that SDoH-GPT annotated about 2,000 samples per 10 minutes after the initial 12 minutes necessary to annotate the human samples necessary to seed the method, by the time humans annotate 512 samples, SDoH-GPT will have annotated around 6,000, still at about one-fifth of the cost. Our results for MIMIC can be seen in Tables 6.1, 6.2, and 6.4.

We have shown our architecture to be able to train the XGBoost model to achieve AUROC scores that are higher than for human annotations for lower costs, even if the human annotations outperform SDoH-GPT for the same number of annotated samples. SDoH-GPT requires only 100 human annotations to seed the automated annotation pipeline, and then it can annotate as many examples as there exist for the task, demonstrating significant cost-effectiveness and efficiency. Our studies revealed minimal variance in AUROC scores between 0-Shot SDoH-GPT and various 2-Shot SDoH-GPT trained on 2048 annotations,

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Community | 16 | $3.98 | $24.87 | $24.87 | $24.87 | $24.87 | $24.87 |
| | 32 | $7.95 | $24.87 | $24.88 | $24.89 | $24.89 | $24.89 |
| | 64 | $15.91 | $24.89 | $24.91 | $24.91 | $24.91 | $24.92 |
| | 128 | $31.81 | $24.91 | $24.95 | $24.96 | $24.97 | $24.98 |
| | 256 | $63.62 | $24.96 | $25.04 | $25.07 | $25.07 | $25.10 |
| | 512 | $127.24 | $25.06 | $25.23 | $25.27 | $25.28 | $25.33 |
| | 1024 | $254.48 | $25.26 | $25.59 | $25.68 | $25.70 | $25.81 |
| | 2048 | $508.97 | $25.67 | $26.32 | $26.50 | $26.55 | $26.75 |
| Economics | 16 | $5.72 | $35.78 | $35.78 | $35.79 | $35.79 | $35.79 |
| | 32 | $11.45 | $35.78 | $35.80 | $35.81 | $35.81 | $35.82 |
| | 64 | $22.89 | $35.80 | $35.83 | $35.84 | $35.85 | $35.86 |
| | 128 | $45.78 | $35.83 | $35.89 | $35.92 | $35.94 | $35.95 |
| | 256 | $91.57 | $35.89 | $36.01 | $36.07 | $36.10 | $36.13 |
| | 512 | $183.14 | $36.00 | $36.25 | $36.37 | $36.44 | $36.50 |
| | 1024 | $366.28 | $36.24 | $36.73 | $36.97 | $37.11 | $37.23 |
| | 2048 | $732.56 | $36.70 | $37.70 | $38.17 | $38.45 | $38.69 |
| Tobacco Use | 16 | $4.68 | $29.23 | $29.24 | $29.24 | $29.24 | $29.24 |
| | 32 | $9.35 | $29.24 | $29.26 | $29.26 | $29.26 | $29.26 |
| | 64 | $18.70 | $29.25 | $29.28 | $29.29 | $29.29 | $29.29 |
| | 128 | $37.41 | $29.28 | $29.34 | $29.35 | $29.36 | $29.36 |
| | 256 | $74.82 | $29.33 | $29.46 | $29.47 | $29.50 | $29.49 |
| | 512 | $149.63 | $29.44 | $29.70 | $29.72 | $29.78 | $29.75 |
| | 1024 | $299.27 | $29.65 | $30.17 | $30.21 | $30.33 | $30.27 |
| | 2048 | $598.54 | $30.07 | $31.11 | $31.20 | $31.43 | $31.31 |

Table 6.2: Price for every SDoH-GPT annotation compared to human annotations for all 8 sample sizes tested in the three MIMIC-III categories. This table shows values in USD and is adjusted for the currency's value in 2023.

suggesting that additional shots do not necessarily enhance performance. Employing 2-Shot SDoH-GPT directly for annotating the Testing Set without XGBoost yielded an AUROC score nearly equivalent to that achieved by using XGBoost trained on 2048 human annotations, reducing the need for extensive manual annotation by a factor of twenty to maintain comparable AUROC scores.

To better compare our findings to other works that used LLMs to annotate similar categories of SDoH, we also tested how well our prompts perform without training an XGBoost to predict subsequent samples more cheaply. For this, we annotated 1024 samples pre-annotated by MIMIC-SBDH and calculated an F1 score, as well as the Cohen's kappa for the accordance between human annotations and SDoH-GPT. F1 and Cohen's kappa results are presented in Table 6.3.

We attained a notably higher F1 score than a previous state of the art in Ramachandran et al. (2023)'s [79] GPT-4 architecture while using our best 2-Shot SDoH-GPT: 0.889 in Economics, which is 0.086 higher; 0.975 in Tobacco Use, surpassing theirs by 0.15; and 0.935 in Community, a significant improvement of 0.345 over their Living Status results, which is the closest equivalent to our Community category. Several factors potentially contribute to the differences:

| Prompt | Community | | Economics | | Tobacco Use | |
|---|---|---|---|---|---|---|
| | F1 | Cohen's kappa | F1 | Cohen's kappa | F1 | Cohen's kappa |
| 0-Shot | 0.9335 | 0.8613 | 0.8745 | 0.7930 | 0.9596 | 0.9043 |
| 2-Shot E | **0.9351** | 0.8125 | 0.7957 | 0.7168 | 0.9596 | 0.9063 |
| 2-Shot E+Ex | 0.9337 | 0.8301 | 0.8637 | 0.7539 | 0.9596 | **0.9219** |
| 2-Shot H | 0.9259 | 0.8418 | 0.7628 | **0.8164** | 0.9656 | **0.9219** |
| 2-Shot H+Ex | 0.9251 | **0.8672** | **0.8893** | 0.7832 | **0.9675** | **0.9219** |

Table 6.3: F1 and Cohen's kappa for the SDoH-GPT prompts (without training XGBoost for further annotation) in the three MIMIC-SBDH categories.

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Community | 16 | 0:02:01 | 0:12:38 | 0:12:38 | 0:12:38 | 0:12:38 | 0:12:38 |
| | 32 | 0:04:01 | 0:12:42 | 0:12:42 | 0:12:42 | 0:12:42 | 0:12:42 |
| | 64 | 0:08:03 | 0:12:50 | 0:12:50 | 0:12:50 | 0:12:50 | 0:12:50 |
| | 128 | 0:16:05 | 0:13:06 | 0:13:06 | 0:13:06 | 0:13:06 | 0:13:06 |
| | 256 | 0:32:11 | 0:13:38 | 0:13:38 | 0:13:38 | 0:13:38 | 0:13:38 |
| | 512 | 1:04:21 | 0:14:42 | 0:14:42 | 0:14:42 | 0:14:42 | 0:14:42 |
| | 1024 | 2:08:42 | 0:16:50 | 0:16:50 | 0:16:50 | 0:16:50 | 0:16:50 |
| | 2048 | 4:17:24 | 0:21:06 | 0:21:06 | 0:21:06 | 0:21:06 | 0:21:06 |
| Economics | 16 | 0:02:54 | 0:18:09 | 0:18:09 | 0:18:09 | 0:18:09 | 0:18:09 |
| | 32 | 0:05:47 | 0:18:13 | 0:18:13 | 0:18:13 | 0:18:13 | 0:18:13 |
| | 64 | 0:11:35 | 0:18:21 | 0:18:21 | 0:18:21 | 0:18:21 | 0:18:21 |
| | 128 | 0:23:09 | 0:18:37 | 0:18:37 | 0:18:37 | 0:18:37 | 0:18:37 |
| | 256 | 0:46:19 | 0:19:09 | 0:19:09 | 0:19:09 | 0:19:09 | 0:19:09 |
| | 512 | 1:32:37 | 0:20:13 | 0:20:13 | 0:20:13 | 0:20:13 | 0:20:13 |
| | 1024 | 3:05:14 | 0:22:21 | 0:22:21 | 0:22:21 | 0:22:21 | 0:22:21 |
| | 2048 | 6:10:29 | 0:26:37 | 0:26:37 | 0:26:37 | 0:26:37 | 0:26:37 |
| Tobacco Use | 16 | 0:02:22 | 0:14:51 | 0:14:51 | 0:14:51 | 0:14:51 | 0:14:51 |
| | 32 | 0:04:44 | 0:14:55 | 0:14:55 | 0:14:55 | 0:14:55 | 0:14:55 |
| | 64 | 0:09:28 | 0:15:03 | 0:15:03 | 0:15:03 | 0:15:03 | 0:15:03 |
| | 128 | 0:18:55 | 0:15:19 | 0:15:19 | 0:15:19 | 0:15:19 | 0:15:19 |
| | 256 | 0:37:50 | 0:15:51 | 0:15:51 | 0:15:51 | 0:15:51 | 0:15:51 |
| | 512 | 1:15:41 | 0:16:55 | 0:16:55 | 0:16:55 | 0:16:55 | 0:16:55 |
| | 1024 | 2:31:21 | 0:19:03 | 0:19:03 | 0:19:03 | 0:19:03 | 0:19:03 |
| | 2048 | 5:02:42 | 0:23:19 | 0:23:19 | 0:23:19 | 0:23:19 | 0:23:19 |

Table 6.4: Time cost for every SDoH-GPT annotation compared to human annotations for all 8 sample sizes tested in the three MIMIC-III categories. This table shows values in H:MM:SS, where H is Hours, M is minutes, and S is Seconds.

1. **Variance in GPT prompt structure:** Their query in GPT prompts was to annotate discharge summaries in the BRAT standoff format, while ours were pure SDoH categorization;

2. **Dataset differences:** Their study contains MIMIC III and an additional dataset from the University of Washington, while ours are MIMIC III and two other datasets;

3. **Instruction guideline:** Their prompt included a lengthy instruction, exceeding 1,000 characters, while our instruction employed a more concise instruction, limited to a few sentences;

4. **Two-shot learning:** Their prompt did not have two examples to facilitate two-shot learning.

Moreover, Guevara et al. (2024) [41] used a manual annotation of 200 medical notes from MIMIC-III and fine-tuned Flan-T5 18 million parameters using LoRA. It reported 0.44 F1 in Community and 0.55 in Economics for MIMIC III. However, our F1 scores in these two categories nearly doubled. Assuming only 256 human annotations are available, SDoH-GPT can effortlessly generate more annotations quickly and cheaply.

### 6.3.2 Validation Sets: Sleep Notes and Suicide Reports

The validation sets behaved similarly, despite the text being arranged rather differently and tasks being only somewhat similar in form. In general, XGBoost models trained with SDoH-GPT annotations perform comparably to models trained with the same number of human annotations at significantly reduced time and computational cost.

These results are interesting for the Suicide Reports dataset since it had much larger text inputs than the MIMIC-III dataset, thus showing that our SDoH-GPT method is effective even for larger text sizes. This also results in much larger time and money expenditures to annotate these texts, emphasizing how much cheaper our automated annotation can be while not losing effectiveness, as can be seen in Tables 6.5, 6.6, and 6.7.

The Sleep Notes had very few annotations, and we were able to collect only 226 human annotations for sleep apnea (64 of which were used as a Training Set and 162 of which were used as a Testing Set). The peak AUROC achieved by an XGBoost model trained on human annotations is 0.922 at 64 training examples. By spending $12.13 more USD and 14.75 more minutes to annotate 2048 examples with SDoH-GPT, we can increase AUROC by 0.0414 points with the 2-Shot H SDoH-GPT prompt. This demonstrates the substantial utility of SDoH-GPT in areas where human annotations are scarce and expensive to obtain. SDoH-GPT markedly enhanced performance with minimal additional effort.

### 6.3.3 Discussion

Our proposed framework of SDoH-GPT is a novel way of leveraging LLM and XGBoost classifiers to efficiently extract SDoH from unstructured medical notes. Our approach demonstrated remarkable efficiency, achieving up to a ten-fold reduction in annotation time and up to a twenty-fold reduction in annotation price compared to traditional methods. Remarkably, it also maintained strong alignment with human annotators, as evidenced by Cohen's kappa scores, as seen in Table 6.3. The impressive results highlight the potential

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Job Problem | 16 | 0.6581 | 0.6216 | 0.6719 | 0.5769 | 0.6174 | 0.6886 |
| | 32 | 0.6631 | 0.8048 | 0.5384 | 0.7113 | 0.8049 | 0.7621 |
| | 64 | 0.8427 | 0.8874 | 0.7549 | 0.8451 | 0.8327 | 0.8055 |
| | 128 | 0.8972 | 0.9065 | 0.8948 | 0.9045 | 0.8734 | 0.9123 |
| | 256 | 0.9365 | 0.9347 | 0.8911 | 0.9317 | 0.9257 | 0.9048 |
| | 512 | 0.9455 | 0.9449 | 0.934 | 0.9496 | 0.9537 | 0.9534 |
| | 1024 | 0.9662 | 0.9702 | 0.938 | 0.9647 | 0.9688 | 0.967 |
| | 2048 | 0.9752 | 0.9691 | 0.9519 | 0.9706 | 0.9731 | 0.9728 |
| Sleep Apnea | 16 | 0.6728 | 0.8132 | 0.6913 | 0.8043 | 0.8384 | 0.6852 |
| | 32 | 0.8903 | 0.7414 | 0.7779 | 0.866 | 0.8592 | 0.8557 |
| | 64 | 0.9215 | 0.8549 | 0.8136 | 0.911 | 0.8706 | 0.882 |
| | 128 | N/A | 0.9 | 0.8929 | 0.9024 | 0.9127 | 0.9262 |
| | 256 | N/A | 0.8822 | 0.9209 | 0.9163 | 0.9279 | 0.924 |
| | 512 | N/A | 0.8743 | 0.916 | 0.9596 | 0.9358 | 0.9383 |
| | 1024 | N/A | 0.908 | 0.9233 | 0.9386 | 0.9383 | 0.9515 |
| | 2048 | N/A | 0.9023 | 0.9418 | 0.9566 | 0.9639 | 0.9457 |

Table 6.5: Performance of XGBoost models trained on human annotations and automated SDoH-GPT annotations for the two validation categories. This table shows AUROC results, measuring the correctness of annotation.

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Job Problem | 16 | $30.56 | $191.04 | $191.07 | $191.08 | $191.07 | $191.08 |
| | 32 | $61.13 | $191.06 | $191.11 | $191.12 | $191.12 | $191.13 |
| | 64 | $122.26 | $191.09 | $191.19 | $191.22 | $191.22 | $191.23 |
| | 128 | $244.51 | $191.16 | $191.35 | $191.42 | $191.41 | $191.43 |
| | 256 | $489.03 | $191.30 | $191.68 | $191.80 | $191.80 | $191.83 |
| | 512 | $978.06 | $191.57 | $192.33 | $192.58 | $192.58 | $192.62 |
| | 1024 | $1,956.11 | $192.12 | $193.63 | $194.14 | $194.13 | $194.22 |
| | 2048 | $3,912.23 | $193.22 | $196.24 | $197.25 | $197.24 | $197.42 |
| Sleep Apnea | 16 | $12.65 | $58.54 | $58.55 | $58.55 | $58.55 | $58.56 |
| | 32 | $25.31 | $58.55 | $58.57 | $58.58 | $58.57 | $58.59 |
| | 64 | $50.62 | $58.57 | $58.62 | $58.63 | $58.61 | $58.66 |
| | 128 | $101.24 | $58.61 | $58.71 | $58.73 | $58.70 | $58.79 |
| | 256 | $202.48 | $58.70 | $58.89 | $58.92 | $58.87 | $59.06 |
| | 512 | $404.96 | $58.87 | $59.24 | $59.32 | $59.20 | $59.58 |
| | 1024 | $809.92 | $59.20 | $59.96 | $60.11 | $59.88 | $60.64 |
| | 2048 | $1,619.83 | $59.88 | $61.39 | $61.69 | $61.23 | $62.75 |

Table 6.6: Price for annotation using each of the SDoH-GPT prompts for the two validation categories. These values are in USD and are adjusted for the currency's value in 2023.

of integrating LLM-based annotation with cost-effective machine learning models to enhance the scalability and accessibility of SDoH data analysis in clinical and public health contexts, minimizing the reliance on extensive human annotation and significantly reducing costs.

Our study establishes 2-Shot SDoH-GPT as a more effective approach compared to Ramachandran et al. (2023) [79], achieving significantly higher F1 scores in Economics, Tobacco Use, and Community categories. This success reflects key advancements in our methodology. Our concise prompts are optimized for SDoH categorization, avoiding ex-

| Task | # Examples | Human | 0-Shot | 2-Shot E | 2-Shot E+Ex | 2-Shot H | 2-Shot H+Ex |
|---|---|---|---|---|---|---|---|
| Job Problem | 16 | 0:15:27 | 1:36:40 | 1:36:40 | 1:36:40 | 1:36:40 | 1:36:40 |
| | 32 | 0:30:55 | 1:36:44 | 1:36:44 | 1:36:44 | 1:36:44 | 1:36:44 |
| | 64 | 1:01:50 | 1:36:52 | 1:36:52 | 1:36:52 | 1:36:52 | 1:36:52 |
| | 128 | 2:03:40 | 1:37:08 | 1:37:08 | 1:37:08 | 1:37:08 | 1:37:08 |
| | 256 | 4:07:19 | 1:37:40 | 1:37:40 | 1:37:40 | 1:37:40 | 1:37:40 |
| | 512 | 8:14:38 | 1:38:44 | 1:38:44 | 1:38:44 | 1:38:44 | 1:38:44 |
| | 1024 | 16:29:16 | 1:40:52 | 1:40:52 | 1:40:52 | 1:40:52 | 1:40:52 |
| | 2048 | 32:58:32 | 1:45:08 | 1:45:08 | 1:45:08 | 1:45:08 | 1:45:08 |
| Sleep Apnea | 16 | 0:23:25 | 0:29:51 | 0:29:40 | 0:29:40 | 0:29:40 | 0:29:40 |
| | 32 | 0:23:26 | 0:29:51 | 0:29:44 | 0:29:44 | 0:29:44 | 0:29:44 |
| | 64 | 0:23:27 | 0:29:51 | 0:29:52 | 0:29:52 | 0:29:52 | 0:29:52 |
| | 128 | 0:23:29 | 0:29:51 | 0:30:08 | 0:30:08 | 0:30:08 | 0:30:08 |
| | 256 | 0:23:34 | 0:29:51 | 0:30:40 | 0:30:40 | 0:30:40 | 0:30:40 |
| | 512 | 0:23:44 | 0:29:51 | 0:31:44 | 0:31:44 | 0:31:44 | 0:31:44 |
| | 1024 | 0:24:03 | 0:29:51 | 0:33:52 | 0:33:52 | 0:33:52 | 0:33:52 |
| | 2048 | 0:24:41 | 0:29:51 | 0:38:08 | 0:38:08 | 0:38:08 | 0:38:08 |

Table 6.7: Time costs for annotation using each of the SDoH-GPT prompts for the two validation categories. This table shows values in H:MM:SS, where H is Hours, M is minutes, and S is Seconds.

cessive complexity instructions. Our use of two-shot learning enables the model to better contextualize examples, a feature absent in their approach. We achieved superior results without requiring extensive fine-tuning, demonstrating the cost-efficiency and scalability of our framework. Additionally, we address the challenges of high lexical complexity across SDoH categories, ensuring consistent performance across diverse datasets, including MIMIC-III, Suicide Reports, and Sleep Notes.

Despite the robust performance of SDoH-GPT, our error analysis reveals certain patterns and areas for improvement. We employed SDoH-GPT to classify the Testing Set in MIMIC-SDBH to conduct a thorough error analysis. We found four categories of errors: Human error (i.e., errors in human annotations); SDoH-GPT error (i.e., errors in SDoH-GPT annotations); Extraction error (i.e., incorrect extractions of social histories from discharge summaries using Regular Expression algorithms), and Ambiguity (i.e., hard to decide). Results from this analysis can be seen in Figure 6.3.

Markedly, the Economics category has the most errors, both human and from SDoH-GPT. This is likely because the Economics category has by far the largest lexical complexity of the MIMIC-SBDH categories, as per Ahsan et al. (2021) [2]. It is reasonable that this complicates both human and automated annotation efforts.

As for ambiguity, we examined the more confusing and ambiguous examples and identified several types of ambiguity:

1. **Contextual Misinterpretations:** SDoH-GPT identified some patients as community present, misinterpreting contexts such as “lost family”, or “deceased parents”;

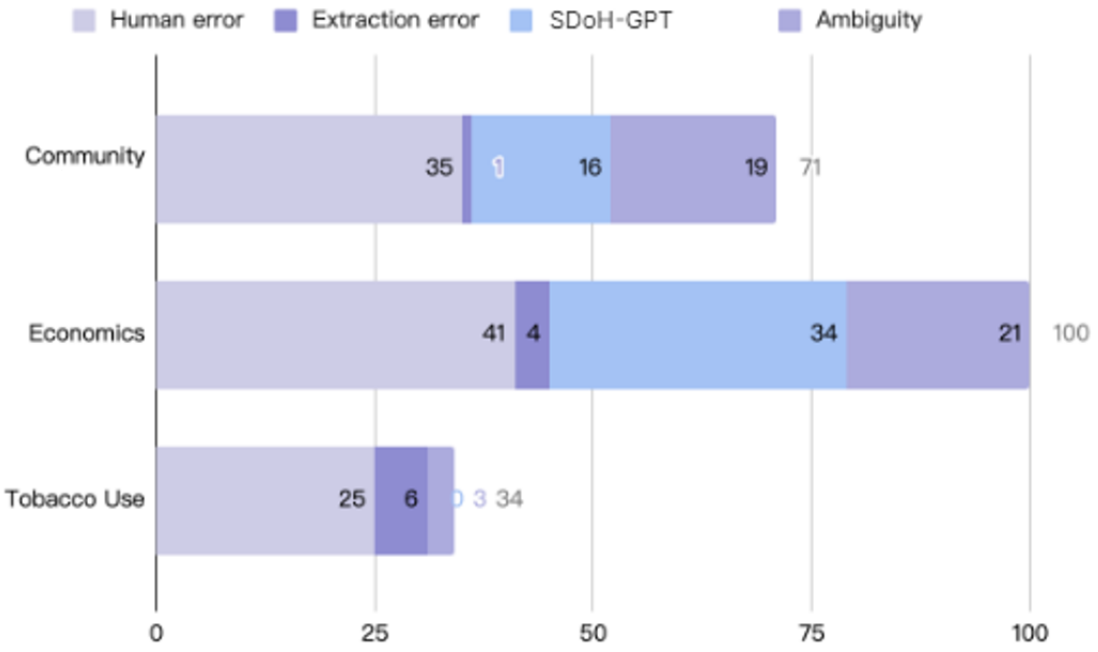



Figure 6.3: The errors we classified during our analysis of the MIMIC-SBDH human annotation comparison against SDoH-GPT.

2. **Temporal Conditions:** human annotators mistakenly treated the subjects as currently employed, neglecting past tense;

3. **Evolving Status:** SDoH-GPT's could annotate based on a limited scope, such as the initial part of the note, leading to the annotation of community present when the note then goes to say that such connections do not exist;

4. **Implicit Statements:** Some cases suggest daily access to community services for the patient, yet this does not explicitly imply anything about community presence, and SDoH-GPT marked those for community presence. Daily access to healthcare services cannot directly infer community presence [9].

5. **Incomplete Information:** Sometimes sentences can be unclear, such as the mention of adult children, yet no mention of whether they are present in the lives of the patients. Nevertheless, SDoH-GPT classified those as community present, which can be questionable.

Furthermore, as SDoH data in medical notes are often brief and lack context, ambiguous text is extremely common. There are several pivotal issues which are unclear to both human annotators and automatic annotation:

1. **Individualization:** The living conditions of a patient in an elder care facility can lead to diverse SDoH annotations [9];

2. **Incompleteness:** The absence of comprehensive information further compounds the ambiguity in determining SDoH status [77];

3. **Misclassification:** LLMs often classify mentions of "community support" as references to social activities, overlooking contexts where it might pertain to community healthcare services [90];

Ambiguity in SDoH stems from its inherently complex and multifaceted nature, requiring a nuanced understanding and context-specific analysis to ensure precise and meaningful annotation [57, 62, 37]. Effectively addressing these issues requires a comprehensive understanding of the medical domain, as well as the broader societal context pertinent to healthcare. By incorporating hard to annotate examples into SDoH-GPT prompts, we improve the model's ability to handle complex scenarios. Through iterative evaluation, including error analysis of SDoH-GPT and human annotations, we identify key areas for improvement, such as refining prompts and addressing annotation inconsistencies over multiple patient visits.

# 7. CONCLUSION

In this work, we presented our contributions to the field of Brazilian medical computation. We worked alongside the Institute for Artificial Intelligence in Healthcare in organizing the data collected from their NoHarm system into the largest collection of tertiary care data in the Portuguese language at the time, which we called BRATECA [20]. We then used the data to create test sets for Admission Prediction [21] and Extended Stay Prediction [22] and showed that these datasets can be transformed and used with classic machine learning and neural networks, reaching good results.

After achieving these results, we looked into new ways to automatically augment our data as well as new ways to more quickly and efficiently annotate raw data, such as those provided in the BRATECA dataset. For this, we partnered with the AI Health Lab at the University of Texas at Austin [1] and investigated such automation using LLMs to aid in annotation and augmentation of data on the widely used MIMIC-III dataset, which has been annotated for several tasks, including our chosen task of Social Determinants of Health Prediction from free text. Our efforts resulted in the SDoH-GPT architecture, which uses LLMs and XGBoost models to cheaply, quickly, and correctly annotate unstructured text data with SDoH categories. This research is currently under review for publication in the Journal of the American Medical Informatics Association (JAMIA)[2].

Finally, we analyzed the usefulness of the holistic use of heterogeneous data as opposed to using only structured table data or unstructured text data. We showed that, for our best performing task, Admission Prediction, using both kinds of data is beneficial but also that attempting to cherry-pick structured data is usually detrimental. This is true even in BRATECA, a dataset with much more unstructured text data than structured table data. This work is currently being peer reviewed for publication in the 20th World Congress on Medical and Health Informatics (MEDINFO2025)[3].

These results have limitations which ought to be pointed out, however. BRATECA, despite being the largest collection of tertiary care data from Brazilian hospitals currently openly accessible for research purposes, is poor in structured data when compared to other datasets such as MIMIC. Of course, MIMIC is a very mature collection that has seen four iterations over many years of development while BRATECA is a much more recent creation, but the lack of data remains. The strength of BRATECA is in its vast unstructured text data, but unfortunately, it is yet unannotated for any tasks. Annotations are expensive and time consuming as they require expertise to perform. We hope that, by further advancing our SDoH-GPT annotation architecture and adapting to BRATECA and other Brazilian collections, we can further advance the state of annotations for many tasks in Brazilian data.

[1] https://aihealth.ischool.utexas.edu/
[2] https://academic.oup.com/jamia
[3] https://medinfo2025.org/

As an example of the kinds of advancements we will be able to pursue with more data, once we have more temporal data from a patient's multiple visits to the hospital, we may be able to integrate important tasks such as patient readmission predictions. We would then be able to add historical data to our holistic approach to data use. As a reminder, BRATECA data currently includes less than a year of hospital data.

The patient flow architectures achieved better than expected results given a dearth of structured time-series data, which is what state-of-the-art prediction architectures for these tasks require to achieve the best results. The most advanced tools used for these tasks, such as attention layers and CNNs, failed to yield results because of this dearth of information, especially the lack of time-series data, such as vital sign data, which is crucial to most solutions implemented using MIMIC data. These remain a possible avenue in future research once the BRATECA collection has matured and more data relevant to these approaches has been added to it.

Still, these architectures are merely a first step to further explorations of these data in the reality of a Brazilian data environment. By showing that even computationally cheaper models like XGBoost are effective for these tasks, it opens up the possibility for trial usage in real hospital environments, and should they perform as seen in our work, this area of research could gather more interest and funding momentum to encourage testing of more computationally expensive architectures such as the CNNs and Attention layers which have been shown to work with MIMIC but did not yield workable results with our data.

The Automated Annotation Generation architecture, SDoH-GPT, demonstrated competitive performance, as well as proving that there exists a cheaper and more timely way to annotate medical data. We could not use the SDoH-GPT method in BRATECA because we lacked human annotated data with which to compare our approach. In the future, once these annotations are created, we will be able to test and even adapt the SDoH-GPT method to BRATECA tasks.

Still, certain limitations must be acknowledged for its applicability in broader clinical settings. First, our SDoH categorization is binary (Yes or No), which may not adequately capture clinical complexity. Future work should explore multi-label or hierarchical classification frameworks to capture richer and more actionable insights. Second, our approach to SDoH annotation is confined to the categorical level and does not extend to sentence-level annotation of triggers and spans. This limits the model's ability to capture context-specific details in real world clinical contexts. Third, real world EHRs are far more fragmented, heterogeneous, and complex, requiring tools capable of handling multiple data formats and sections. To address these challenges, future work should aim to enhance SDoH-GPT's ability to handle this diversity while incorporating sentence-level annotation techniques for more precise and meaningful insights. Lastly, while our SDoH-GPT framework demonstrates generalizability across different tasks, including applications in domains such as suicide and sleep apnea, further targeted analysis is required to fully explore its adaptability

and potential in these areas. Future research should also explore tailored adaptations of the model, such as optimizing prompts and integrating domain-specific knowledge for these unique contexts. All of these avenues of future work would help in applying our SDoH-GPT method to BRATECA and other Brazilian collections as well.

Our work here, as well as much of the literature we've referenced, aims to slowly and safely introduce machine learning solutions into hospital environments. By targeting administrative tasks and improving their effectiveness, clarity, and trustworthiness, we hope that medical professionals and hospital administrators will begin trialing these solutions. Should they work well in practice, it is likely that access to new data will become less difficult, and also open up more possibilities for working with hospitals to collect better data.

The construction of real time data pipelines in hospitals, which is slowly taking place in Brazil, will also add new dimensions to this area of research. The possibility of real time updates to training data, the use of AI alerts to allow medical professionals to become aware of issues before they come to a head, and improved administrative organization with predictive patient flow models are slowly becoming reality as computational medicine and hospital AI develop.

We aim to continue developing BRATECA with more organization and more data. Patient safety and anonymity must remain paramount. Thus, much work must still be put into this project so that the data becomes both safer to handle and easier to access. We hope that the fruits of our work become part of the foundation of Brazilian computational medicine research.

# REFERENCES


[1] Agrawal, M.; Hegselmann, S.; Lang, H.; Kim, Y.; Sontag, D. “Large language models are few-shot clinical information extractors”, *ArXiv*, vol. arXiv:2205.12689, Nov 2022, pp. 26.

[2] Ahsan, H.; Ohnuki, E.; Mitra, A.; You, H. “Mimic-sbdh: a dataset for social and behavioral determinants of health”. In: Proceedings of the 6th Machine Learning for Healthcare Conference, 2021, pp. 391–413.

[3] Akbik, A.; Blythe, D.; Vollgraf, R. “Contextual string embeddings for sequence labeling”. In: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 1638–1649.

[4] Alghatani, K.; Ammar, N.; Rezgui, A.; Shaban-Nejad, A. “Predicting intensive care unit length of stay and mortality using patient vital signs: Machine learning model development and validation”, *JMIR Medical Informatics*, vol. 9–5, Jun 2021, pp. 23.

[5] Bandeira, L. T.; Consoli, B. S.; Vieira, R.; Bordin, R. H. “Semantic textual similarity for abridging clinical notes in brazilian electronic health records”. In: Proceedings of the Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2023, pp. 224–228.

[6] Bejan, C. A.; Angiolillo, J.; Conway, D.; Nash, R.; Shirey-Rice, J. K.; Lipworth, L.; Cronin, R. M.; Pulley, J.; Kripalani, S.; Barkin, S.; et al.. “Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records”, *Journal of the American Medical Informatics Association*, vol. 25–1, Jul 2018, pp. 61–71.

[7] Bertsimas, D.; Pauphilet, J.; Stevens, J.; Tandon, M. “Predicting inpatient flow at a major hospital using interpretable analytics”, *Manufacturing & Service Operations Management*, vol. 24–6, Jun 2021, pp. 2797–3306.

[8] Bhate, N. J.; Mittal, A.; He, Z.; Luo, X. “Zero-shot learning with minimum instruction to extract social determinants and family history from clinical notes using gpt model”. In: Proceedings of the IEEE International Conference on Big Data, 2023, pp. 1476–1480.

[9] Boamah, S. A.; Weldrick, R.; Lee, T.-S. J.; Taylor, N. “Social isolation among older adults in long-term care: A scoping review”, *Journal of Aging and Health*, vol. 33–8, Mar 2021, pp. 618–632.

[10] Bradley, A. P. “The use of the area under the roc curve in the evaluation of machine learning algorithms”, *Pattern recognition*, vol. 30–7, Jul 1997, pp. 1145–1159.

[11] Braveman, P.; Gottlieb, L. “The social determinants of health: it’s time to consider the causes of the causes”, *Public Health Reports*, vol. 129–1, Jan 2014, pp. 19–31.

[12] Brin, D.; Sorin, V.; Vaid, A.; Soroush, A.; Glicksberg, B. S.; Charney, A. W.; Nadkarni, G.; Klang, E. “Comparing chatgpt and gpt-4 performance in usmle soft skill assessments”, *Scientific Reports*, vol. 13–1, Oct 2023, pp. 5.

[13] Brink, A.; Alsma, J.; van Attekum, L. A.; Bramer, W. M.; Zietse, R.; Lingsma, H.; Schuit, S. C. “Predicting inhospital admission at the emergency department: a systematic review”, *Emergency Medicine Journal*, vol. 39–3, Oct 2022, pp. 191–198.

[14] Brysbaert, M. “How many words do we read per minute? a review and meta-analysis of reading rate”, *Journal of memory and language*, vol. 109–1, Dec 2019, pp. 8.

[15] Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; Varoquaux, G. “API design for machine learning software: experiences from the scikit-learn project”. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[16] Carrell, D. S.; Cronkite, D. J.; Malin, B. A.; Aberdeen, J. S.; Hirschman, L. “Is the juice worth the squeeze? costs and benefits of multiple human annotators for clinical text de-identification”, *Methods of information in medicine*, vol. 55–4, Jan 2016, pp. 356–364.

[17] Carter, J. V.; Pan, J.; Rai, S. N.; Galandiuk, S. “Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves”, *Surgery*, vol. 159–6, Jun 2016, pp. 1638–1645.

[18] Che, Z.; Cheng, Y.; Sun, Z.; Liu, Y. “Exploiting convolutional neural network for risk prediction with medical feature embedding”, *ArXiv*, vol. abs/1701.07474, Jan 2017, pp. 5.

[19] Che, Z.; Kale, D.; Li, W.; Bahadori, M. T.; Liu, Y. “Deep computational phenotyping”. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 507–516.

[20] Consoli, B.; Dias, H.; Ulbrich, A.; Vieira, R.; Bordini, R. “Brateca (brazilian tertiary care dataset): a clinical information dataset for the portuguese language”. In: Proceedings of the 13th Conference on Language Resources and Evaluation, 2022, pp. 5609—-5616.

[21] Consoli, B.; Viera, R.; Bordini, R. H.; Manssour, I. H. “Predicting inpatient admissions in brazilian hospitals”. In: Proceedings of the SIMPÓSIO BRASILEIRO DE COMPUTAÇÃO APLICADA À SAÚDE, 2024, pp. 284–295.

[22] Consoli, B. S.; Vieira, R.; Bordini, R. H. “Benchmarking the brateca clinical data collection for prediction tasks”. In: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies, 2023, pp. 338–345.

[23] Conway, M.; Keyhani, S.; Christensen, L.; South, B. R.; Vali, M.; Walter, L. C.; Mowery, D. L.; Abdelrahman, S.; Chapman, W. W. “Moonstone: a novel natural language processing system for inferring social risk from clinical narratives”, *Journal of Biomedical Semantics*, vol. 10–6, Apr 2019, pp. 1–10.

[24] Cui, L.; Xie, X.; Shen, Z. “Prediction task guided representation learning of medical codes in ehr”, *Journal of Biomedical Informatics*, vol. 84–1, Aug 2018, pp. 1–10.

[25] Cusido, J.; Comalrena, J.; Alavi, H.; Llunas, L. “Predicting hospital admissions to reduce crowding in the emergency departments”, *Applied Sciences*, vol. 12–21, Oct 2022, pp. 16.

[26] D. P. dos Santos, H.; D. P. S. Ulbrich, A. H.; Vieira, R. “Evaluation of a prescription outlier detection system in hospital’s pharmacy services”. In: Proceedings of the 12th International Workshop on Biomedical and Health Informatics, 2021, pp. 2862–2868.

[27] D. P. S. Ulbrich, A. H.; Aline Maciel dos Santos, K.; Dias Pereira dos Santos, H.; Zanella Lazaretto, F. “Analysis of pharmaceutical interventions performed with decision support using artificial intelligence in brazilian hospitals”. In: Proceedings of the 13th Brazilian Congress of Hospital Pharmacy, 2021, pp. 16–16.

[28] da Saúde, M. “TERMINOLOGIA BÁSICA EM SAÚDE”. Secretaria Nacional de Organização e Desenvolvimento de Serviços de Saúde, 1987, 52p.

[29] De Hond, A. A.; Steyerberg, E. W.; Van Calster, B. “Interpreting area under the receiver operating characteristic curve”, *The Lancet Digital Health*, vol. 4–12, Dec 2022, pp. 853–855.

[30] de Melo, T.; Figueiredo, C. M. “A first public dataset from brazilian twitter and news on covid-19 in portuguese”, *Data in Brief*, vol. 32–1, Aug 2020, pp. 8.

[31] Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. “BERT: pre-training of deep bidirectional transformers for language understanding”. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[32] Dhar, T.; Dey, N.; Borra, S.; Sherratt, R. S. “Challenges of deep learning in medical image analysis—improving explainability and trust”, *IEEE Transactions on Technology and Society*, vol. 4–1, Jan 2023, pp. 68–75.

[33] Ding, B.; Qin, C.; Liu, L.; Chia, Y. K.; Joty, S.; Li, B.; Bing, L. “Is gpt-3 a good data annotator?”, *ArXiv*, vol. arXiv:2212.10450, Jun 2022, pp. 21.

[34] Dorr, D. A.; Phillips, W.; Phansalkar, S.; Sims, S. A.; Hurdle, J. F. “Assessing the difficulty and time cost of de-identification in clinical narratives”, *Methods of information in medicine*, vol. 45–3, Aug 2006, pp. 246–252.

[35] e Oliveira, L. E. S.; Peters, A. C.; da Silva, A. M. P.; Gebeluca, C. P.; Gumiel, Y. B.; Cintho, L. M. M.; Carvalho, D. R.; Hasan, S. A.; Moro, C. M. C. “Semclinbr - a multi institutional and multi specialty semantically annotated corpus for portuguese clinical NLP tasks”, *Journal of Biomedical Semantics*, vol. 13–1, May 2022, pp. 19.

[36] Feliciana Silva, F.; Macedo da Silva Bonfante, G.; Reis, I. A.; André da Rocha, H.; Pereira Lana, A.; Leal Cherchiglia, M. “Hospitalizations and length of stay of cancer patients: A cohort study in the brazilian public health system”, *PLOS ONE*, vol. 15–5, May 2020, pp. 1–13.

[37] Feller, D. J.; Don’t Walk, O. J. B.; Zucker, J.; Yin, M. T.; Gordon, P.; Elhadad, N.; et al.. “Detecting social and behavioral determinants of health with structured and free-text clinical data”, *Applied clinical informatics*, vol. 11–1, Mar 2020, pp. 172–181.

[38] Gilson, A.; Safranek, C. W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R. A.; Chartash, D.; et al.. “How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment”, *JMIR Medical Education*, vol. 9–1, Dec 2023, pp. 9.

[39] Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; Stanley, H. E. “Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals”, *circulation*, vol. 101–23, Jun 2000, pp. 215–220.

[40] Grave, E.; Mikolov, T.; Joulin, A.; Bojanowski, P. “Bag of tricks for efficient text classification”. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017, pp. 427–431.

[41] Guevara, M.; Chen, S.; Thomas, S.; Chaunzwa, T. L.; Franco, I.; Kann, B. H.; Moningi, S.; Qian, J. M.; Goldstein, M.; Harper, S.; et al.. “Large language models to identify social determinants of health in electronic health records”, *NPJ Digital Medicine*, vol. 7–1, Jan 2024, pp. 6.

[42] Guo, C.; Chen, J. “Big data analytics in healthcare”, *Knowledge technology and systems: Toward establishing knowledge systems science*, vol. 34–1, Jan 2023, pp. 27–70.

[43] Hatef, E.; Rouhizadeh, M.; Tia, I.; Lasser, E.; Hill-Briggs, F.; Marsteller, J.; Kharrazi, H.; et al.. "Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system", *JMIR medical informatics*, vol. 7–3, Aug 2019, pp. 14.

[44] Hong, W. S.; Haimovich, A. D.; Taylor, R. A. "Predicting hospital admission at emergency department triage using machine learning", *PLOS ONE*, vol. 13–7, Jul 2018, pp. 1–13.

[45] Huang, Y.; Yang, X.; Xu, C. "Time-guided high-order attention model of longitudinal heterogeneous healthcare data", *ArXiv*, vol. abs/1912.00773, Aug 2019, pp. 57–70.

[46] Imamalieva, D. "Legal difficulties associated with the use of big data in healthcare: Civil law and cyberlaw review", *Medicine, Law & Society*, vol. 17–1, Apr 2024, pp. 22.

[47] Jaotombo, F.; Pauly, V.; Fond, G.; Orleans, V.; Auquier, P.; Ghattas, B.; Boyer, L. "Machine-learning prediction for hospital length of stay using a french medico-administrative database", *Journal of Market Access & Health Policy*, vol. 11–1, Nov 2023, pp. 11.

[48] Jiang, L. Y.; Liu, X. C.; Nejatian, N. P.; Nasir-Moin, M.; Wang, D.; Abidin, A.; Eaton, K.; Riina, H. A.; Laufer, I.; Punjabi, P.; et al.. "Health system-scale language models are all-purpose prediction engines", *Nature*, vol. 619–7969, Jun 2023, pp. 357–362.

[49] Johnson, A.; Bulgarelli, L.; Pollard, T.; Horng, S.; Celi, L.; Mark, R. "Mimic-iv, a freely accessible electronic health record dataset", *Scientific Data*, vol. 10–1, Jan 2023, pp. 9.

[50] Johnson, A.; Pollard, T.; Shen, L.; Lehman, L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.; Mark, R. "Mimic-iii, a freely accessible critical care database", *Scientific Data*, vol. 3–1, May 2016, pp. 5.

[51] Kadri, F.; Dairi, A.; Harrou, F.; Sun, Y. "Towards accurate prediction of patient length of stay at emergency department: a gan-driven deep learning framework", *Journal of Ambient Intelligence and Humanized Computing*, vol. 14–1, Feb 2023, pp. 11481–11495.

[52] Knevel, R.; Liao, K. P. "From real-world electronic health record data to real-world results using artificial intelligence", *Annals of the Rheumatic Diseases*, vol. 82–3, Mar 2023, pp. 306–311.

[53] Kurtz, P.; Peres, I.; Soares, M.; Soares, M.; Salluh, J. I. F.; Bozza, F. A. "Hospital length of stay and 30-day mortality prediction in stroke: A machine learning analysis of 17,000 icu admissions in brazil", *Neurocritical Care*, vol. 37–2, Apr 2022, pp. 313–321.

[54] Lazaretto, F.; Ulbrich, A. H.; dos Santos, K.; dos Santos, H. “Análise das intervenções farmacêuticas realizadas com suporte à decisão utilizando inteligência artificial em hospitais brasileiros”. In: Proceedings of the 13th Brazilian Congress of Hospital Pharmacy and Health Services, 2021, pp. 5609–5616.

[55] LeCun, Y.; Bengio, Y.; Hinton, G. E. “Deep learning”, *Nature*, vol. 521–7553, May 2015, pp. 436–444.

[56] Lee, C. H.; Yoon, H.-J. “Medical big data: promise and challenges”, *Kidney research and clinical practice*, vol. 36–1, Mar 2017, pp. 3.

[57] Lituiev, D. S.; Lacar, B.; Pak, S.; Abramowitsch, P. L.; De Marchis, E. H.; Peterson, T. A. “Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients”, *Journal of the American Medical Informatics Association*, vol. 30–8, May 2023, pp. 1438–1447.

[58] Liu, G. S. “Surveillance for violent deaths—national violent death reporting system, 48 states, the district of columbia, and puerto rico, 2020”, *Morbidity and Mortality Weekly Report Surveillance Summaries*, vol. 72–5, May 2023, pp. 1–38.

[59] Liu, L.; Li, H.; Hu, Z.; Shi, H.; Wang, Z.; Tang, J.; Zhang, M. “Learning hierarchical representations of electronic health records for clinical outcome prediction”. In: Proceedings of the American Medical Informatics Association Annual Symposium, 2019, pp. 597–606.

[60] Liu, N.; Gao, R.; Yuan, J.; Park, C.; Xing, S.; Gou, S. “Gru-tv: Time-and velocity-aware gru for patient representation on multivariate clinical time-series data”, *ArXiv*, vol. arXiv:2205.04892, Oct 2022, pp. 11.

[61] Liu, X.; Wang, H.; He, T.; Liao, Y.; Jian, C. “Recent advances in representation learning for electronic health records: A systematic review”, *Journal of Physics: Conference Series*, vol. 2188–1, Feb 2022, pp. 14.

[62] Lybarger, K.; Ostendorf, M.; Yetisgen, M. “Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction”, *Journal of Biomedical Informatics*, vol. 113-1, Jan 2021, pp. 11.

[63] Ma, F.; You, Q.; Xiao, H.; Chitta, R.; Zhou, J.; Gao, J. “Kame: Knowledge-based attention model for diagnosis prediction in healthcare”. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 743–752.

[64] Magnan, S. “Social determinants of health 101 for health care: five plus five”, *National Academy of Medicine Perspectives*, vol. 2021–1, Jun 2021, pp. 36.

[65] McHugh, M. L. “Interrater reliability: the kappa statistic”, *Biochemia medica*, vol. 22–3, 7 2012, pp. 276–282.

[66] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; Dean, J. “Distributed representations of words and phrases and their compositionality”. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, 2013, pp. 3111–3119.

[67] Miotto, R.; Li, L.; Kidd, B. “Deep patient: An unsupervised representation to predict the future of patients from the electronic health records”, *Scientific Reports*, vol. 6–1, May 2016, pp. 10.

[68] Natália Boff Medeiros, Flávio Sanson Fogliatto, M. K. R.; Tortorella, G. L. “Predicting the length-of-stay of pediatric patients using machine learning algorithms”, *International Journal of Production Research*, vol. 63–2, Jun 2023, pp. 483–496.

[69] Nazario-Johnson, L.; Zaki, H. A.; Tung, G. A. “Use of large language models to predict neuroimaging”, *Journal of the American College of Radiology*, vol. 20–10, Oct 2023, pp. 1004–1009.

[70] Niero, L. H. P.; Guilherme, I. R.; Oliveira, L. E. S. e.; de Araújo Filho, G. M. “Psybertpt: A clinical entity recognition model for psychiatric narratives”. In: Proceedings of the 36th International Symposium on Computer-Based Medical Systems, 2023, pp. 672–677.

[71] Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. “Explanation of machine learning models using shapley additive explanation and application for real data in hospital”, *Computer Methods and Programs in Biomedicine*, vol. 214–1, Feb 2022, pp. 4.

[72] Patra, B. G.; Sharma, M. M.; Vekaria, V.; Adekkanattu, P.; Patterson, O. V.; Glicksberg, B.; Lepow, L. A.; Ryu, E.; Biernacka, J. M.; Furmanchuk, A.; et al.. “Extracting social determinants of health from electronic health records using natural language processing: a systematic review”, *Journal of the American Medical Informatics Association*, vol. 28–12, Oct 2021, pp. 2716–2727.

[73] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al.. “Scikit-learn: Machine learning in python”, *The Journal of Machine Learning Research*, vol. 12–1, Jan 2011, pp. 2825–2830.

[74] Peres, I. T.; Hamacher, S.; Cyrino Oliveira, F. L.; Bozza, F. A.; Salluh, J. I. F. “Data-driven methodology to predict the icu length of stay: A multicentre study of 99,492

admissions in 109 brazilian units", *Anaesthesia Critical Care & Pain Medicine*, vol. 41–6, Dec 2022, pp. 4.

[75] Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. "Deep contextualized word representations". In: Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies, 2018, pp. 2227–2237.

[76] Pollard, T.; Johnson, A.; Raffa, J.; Celi, L.; Mark, R.; Badawi, O. "The eicu collaborative research database, a freely available multi-center database for critical care research", *Scientific Data*, vol. 5–1, Sep 2018, pp. 13.

[77] Portacolone, E.; Nguyen, T. T.; Bowers, B. J.; Johnson, J. K.; Kotwal, A. A.; Stone, R. I.; Keiser, S.; Tran, T.; Rivera, E.; Martinez, P.; et al.. "Perceptions of the role of living alone in providing services to patients with cognitive impairment", *JAMA Network Open*, vol. 6–8, Aug 2023, pp. 13.

[78] Rajkomar, A.; Oren, E.; Chen, K.; ai, A. M.; Hajaj, N.; Hardt, M.; Liu, P. J.; Liu, X.; Marcus, J.; Sun, M.; Sundberg, P.; Yee, H.; Zhang, K.; Zhang, Y.; Flores, G.; Duggan, G. E.; Irvine, J.; Le, Q.; Litsch, K.; Mossin, A.; Tansuwan, J.; Wang, D.; Wexler, J.; Wilson, J.; Ludwig, D.; Volchenboum, S. L.; Chou, K.; Pearson, M.; Madabushi, S.; Shah, N. H.; Butte, A. J.; Howell, M. D.; Cui, C.; Corrado, G. S.; Dean, J. "Scalable and accurate deep learning with electronic health records", *Nature Digital Medicine*, vol. 1–18, May 2018, pp. 10.

[79] Ramachandran, G. K.; Fu, Y.; Han, B.; Lybarger, K.; Dobbins, N. J.; Uzuner, Ö.; Yetisgen, M. "Prompt-based extraction of social determinants of health using few-shot learning", *ArXiv*, vol. arXiv:2306.07170, Jan 2023, pp. 9.

[80] Reddy, S. "Explainability and artificial intelligence in medicine", *The Lancet Digital Health*, vol. 4–4, Apr 2022, pp. 214–215.

[81] Reis, E. P.; Paiva, J.; Bueno da Silva, M. C.; Sousa Ribeiro, G. A.and Fornasiero Paiva, V.; Bulgarelli, L.; Lee, H.; dos Santos, P. V.; Brito, B.; Amaral, L.; Beraldo, G.; Haidar Filho, J. N.; Teles, G.; Szarf, G.; Pollard, T.; Johnson, A.; Celi, L. A.; Amaro, E. "Brax, a brazilian labeled chest x-ray dataset", *Scientific Data*, vol. 9–1, Aug 2022, pp. 8.

[82] Reshetnyak, E.; Ntamatungiro, M.; Pinheiro, L. C.; Howard, V. J.; Carson, A. P.; Martin, K. D.; Safford, M. M. "Impact of multiple social determinants of health on incident stroke", *Stroke*, vol. 51–8, Jul 2020, pp. 2445–2453.

[83] Rongali, S.; Rose, A. J.; McManus, D. D.; Bajracharya, A. S.; Kapoor, A.; Granillo, E.; Yu, H. "Learning latent space representations to predict patient outcomes:

Model development and validation", *Journal of Medical Internet Research*, vol. 22–3, Mar 2020, pp. 13.

[84] Santos, J.; dos Santos, H. D.; Tabalipa, F.; Vieira, R. "De-identification of clinical notes using contextualized language models and a token classifier". In: Proceedings of the 10th Brazilian Conference on Intelligent Systems, 2021, pp. 33–41.

[85] Savova, G. K.; Masanz, J. J.; Ogren, P. V.; Zheng, J.; Sohn, S.; Kipper-Schuler, K. C.; Chute, C. G. "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications", *Journal of the American Medical Informatics Association*, vol. 17–5, Sep 2010, pp. 507–513.

[86] Schneider, E. T. R.; de Souza, J. V. A.; Knafou, J.; Oliveira, L. E. S. e.; Copara, J.; Gumiel, Y. B.; Oliveira, L. F. A. d.; Paraiso, E. C.; Teodoro, D.; Barra, C. M. C. M. "BioBERTpt - a Portuguese neural language model for clinical named entity recognition". In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, 2020, pp. 65–72.

[87] Shamout, F.; Zhu, T.; Clifton, D. A. "Machine learning for clinical outcome prediction", *IEEE Reviews in Biomedical Engineering*, vol. 14–1, Jul 2021, pp. 116—-126.

[88] Si, Y.; Du, J.; Li, Z.; Jiang, X.; Miller, T.; Wang, F.; Zheng, W. J.; Roberts, K. "Deep representation learning of patient data from electronic health records (ehr): A systematic review", *Journal of Biomedical Informatics*, vol. 115–1, Mar 2021, pp. 13.

[89] Si, Y.; Roberts, K. "Deep patient representation of clinical notes via multi-task learning for mortality prediction". In: Proceedings of the AMIA Joint Summits on Translational Science, 2019, pp. 779–788.

[90] Singh, R.; Javed, Z.; Yahya, T.; Valero-Elizondo, J.; Acquah, I.; Hyder, A. A.; Maqsood, M. H.; Amin, Z.; Al-Kindi, S.; Cainzos-Achirica, M.; et al.. "Community and social context: an important social determinant of cardiovascular disease", *Methodist Debakey Cardiovascular Journal*, vol. 17–4, Sep 2021, pp. 15.

[91] Sivarajkumar, S.; Tam, T. Y. C.; Mohammad, H. A.; Viggiano, S.; Oniani, D.; Visweswaran, S.; Wang, Y. "Extraction of sleep information from clinical notes of alzheimer's disease patients using natural language processing", *Journal of the American Medical Informatics Association*, vol. 31–10, Oct 2024, pp. 2217–2227.

[92] Song, H.; Rajan, D.; Thiagarajan, J. J.; Spanias, A. "Attend and diagnose: Clinical time series analysis using attention models". In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 4091–4098.

[93] Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval", *Journal of documentation*, vol. 28–1, Jan 1972, pp. 11–21.

[94] Spasic, I.; Nenadic, G.; et al.. “Clinical text data in machine learning: systematic review”, *JMIR medical informatics*, vol. 8–3, Mar 2020, pp. 19.

[95] Stojanovic, J.; Gligorijevic, D.; Radosavljevic, V.; Djuric, N.; Grbovic, M.; Obradovic, Z. “Modeling healthcare quality via compact representations of electronic health records”, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14–3, Jul 2017, pp. 545–554.

[96] Suresh, H.; Hunt, N.; Johnson, A. E. W.; Celi, L. A.; Szolovits, P.; Ghassemi, M. “Clinical intervention prediction and understanding with deep neural networks”. In: Proceedings of the Machine Learning for Health Care Conference, 2017, pp. 322–337.

[97] Sushil, M.; Kennedy, V. E.; Mandair, D.; Miao, B. Y.; Zack, T.; Butte, A. J. “Coral: expert-curated oncology reports to advance language model inference”, *NEJM AI*, vol. 1–4, Mar 2024, pp. 15.

[98] Wade, D. “Ethics of collecting and using healthcare data”, *British Medical Journal*, vol. 334–7608, Jun 2007, pp. 1330–1331.

[99] Wang, M.; Pantell, M. S.; Gottlieb, L. M.; Adler-Milstein, J. “Documentation and review of social determinants of health data in the ehr: measures and associated insights”, *Journal of the American Medical Informatics Association*, vol. 28–12, Sep 2021, pp. 2608–2616.

[100] Wei, Q.; Franklin, A.; Cohen, T.; Xu, H. “Clinical text annotation–what factors are associated with the cost of time?” In: Proceedings of the AMIA Annual Symposium, 2018, pp. 1552–1560.

[101] Xu, Y.; Biswal, S.; Deshpande, S. R.; Maher, K. O.; Sun, J. “Raim: Recurrent attentive and intensive model of multimodal patient monitoring data”. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2565–2573.

[102] Yang, S.; Varghese, P.; Stephenson, E.; Tu, K.; Gronsbell, J. “Machine learning approaches for electronic health records phenotyping: a methodical review”, *Journal of the American Medical Informatics Association*, vol. 30–2, Nov 2022, pp. 367–381.

[103] Yèche, H.; Kuznetsova, R.; Zimmermann, M.; Hüser, M.; Lyu, X.; Faltys, M.; Rätsch, G. “Hirid-icu-benchmark–a comprehensive machine learning benchmark on high-resolution icu data”, *ArXiv*, vol. arXiv:2111.08536, Jan 2021, pp. 27.

[104] Yin, K.; Qian, D.; Cheung, W. K.; Fung, B. C. M.; Poon, J. “Learning phenotypes and dynamic patient representations via rnn regularized collective non-negative tensor factorization”. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 1246–1253.

[105] Yu, Z.; Yang, X.; Dang, C.; Wu, S.; Adekkanattu, P.; Pathak, J.; George, T. J.; Hogan, W. R.; Guo, Y.; Bian, J.; et al.. "A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models". In: Proceedings of the AMIA Annual Symposium, 2022, pp. 1225–1233.

[106] Zhang, E.; Robinson, R.; Pfahringer, B. "Deep holistic representation learning from ehr". In: Proceedings of the 12th International Symposium on Medical Information and Communication Technology (ISMICT), 2018, pp. 1–6.

[107] Zhang, J.; Gong, J.; Barnes, L. "Hcnn: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records". In: Proceedings of the IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies, 2017, pp. 214–221.

[108] Zhou, C.; Jia, Y.; Motani, M. "Optimizing autoencoders for learning deep representations from health data", *IEEE Journal of Biomedical and Health Informatics*, vol. 23–1, Jan 2019, pp. 103–111.

[109] Zhou, C.; Jia, Y.; Motani, M.; Chew, J. "Learning deep representations from heterogeneous patient data for predictive diagnosis". In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics, 2017, pp. 115–123.

[110] Zhou, J.; Wang, F.; Hu, J.; Ye, J. "From micro to macro: Data driven phenotyping by densification of longitudinal electronic medical records". In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 135–144.


Pontifícia Universidade Católica do Rio Grande do Sul
Pró-Reitoria de Pesquisa e Pós-Graduação
Av. Ipiranga, 6681 – Prédio 1 – Térreo
Porto Alegre – RS – Brasil
Fone: (51) 3320-3513
E-mail: propesq@pucrs.br
Site: www.pucrs.br