2021

Multi-stream Longitudinal Data Analysis using Deep Learning

Sajjad Fouladvand
University of Kentucky, sjjd.fouladvand@gmail.com
Author ORCID Identifier:
https://orcid.org/0000-0002-9869-1836
Digital Object Identifier: https://doi.org/10.13023/etd.2021.395

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Sajjad Fouladvand, Student

Dr. Jin Chen, Major Professor

Dr. Simone Silvestri, Director of Graduate Studies

Multi-stream Longitudinal Data Analysis using Deep Learning

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Engineering at the
University of Kentucky

By
Sajjad Fouladvand
Lexington, Kentucky

Director: Jin Chen, Associate Professor of Computer Science
Lexington, Kentucky
2021

https://orcid.org/0000-0002-9869-1836

ABSTRACT OF DISSERTATION

Multi-stream Longitudinal Data Analysis using Deep Learning

Longitudinal healthcare data encompasses all tasks where patients information are collected at multiple follow-up times. Analyzing this data is critical in addressing many real world problems in healthcare such as disease prediction and prevention. In this thesis, technical challenges in analyzing longitudinal administrative claims data are addressed and novel deep learning based models are proposed for multi-stream data analysis and disease prediction tasks. These algorithms and frameworks are assessed mainly on substance use disorders prediction tasks and specifically designed to tackled these disorders. Substance use disorder is a public health crisis costing the US an estimated $740 billion annually in healthcare, lost workplace productivity, and crime. Early identification and engagement of individuals at risk of developing a substance use disorder is a critical unmet need in healthcare which can be achieved by producing automatic artificial intelligence based tools trained using big healthcare data. In fact, healthcare data can be harnessed together with artificial intelligence and machine learning to advance our understanding of factors that increase the propensity for developing different diseases as well as those that aid in the treatment of these disorders.

Here in, a disease prediction framework is first proposed based on recurrent neural networks. This framework includes three components: 1) data pre-processing, 2) disease prediction using long short term memory models, and 3) hypothesis exploration by varying the models and the inputs. This framework is assessed using two use cases: substance use disorder prediction and mild cognitive impairment prediction. Experimental results show that this proposed model can efficiently analyze patients' data and creates efficient disease prediction tools. Second, the limitations of current deep learning models including long short term memory models in claims data analysis are detected and addressed, and a novel model based on the transformer

models is proposed. In fact, leveraging the real-world longitudinal claims data, a novel multi-stream transformer model is proposed for predicting opioid use disorder as an important case of substance use disorders. This model is designed to simultaneously analyze multiple types of data streams, such as medications, diagnoses, procedures and demographics, by attending to segments within and across these data streams. The proposed model tested on the IBM MarketScan data showed significantly better performance than the traditional models and recently developed deep learning models.

KEYWORDS: Deep Learning, Longitudinal Data Analysis, Healthcare Data Analysis, Transformer, Recurrent Neural Networks.

Sajjad Fouladvand

October 26, 2021

Multi-stream Longitudinal Data Analysis using Deep Learning

By
Sajjad Fouladvand

Jin Chen
Director of Dissertation

Silvestri Simone
Director of Graduate Studies

October 26, 2021
Date

Dedicated to my mom and all other parents whom we lost during the Covid-19 pandemic.

ACKNOWLEDGMENTS

Undertaking this PhD has been a great experience and it would not have been successful without the support of several individuals in my life. First and foremost, I would like to express my sincere appreciation and gratitude to my PhD advisor, Dr. Jin Chen, for his constant support and guidance thorough my doctoral program. Dr. Chen's support started even before I start my PhD program when he patiently help me to work through my visa and travel issues. He continued this support thorough my PhD at the University of Kentucky during the past four years and guided me into the right direction. This dissertation would not have been successful without Dr. Chen and I am grateful for his help, support and ideas.

Besides my advisor, I would also like to express my deep appreciation for the rest of my doctoral committee: Dr. Jeffery Talbert, Dr. Licong Cui and Dr. Jinze Liu for their valuable comments and feedback which helped me to shape my PhD dissertation and implement my ideas. Further, I would also like to thank the University of Kentucky, the Department of Computer Science, and the Directors of Graduate Studies Dr. Simone Silvestri and Dr. Miroslaw (Mirek) Truszczynski that supported me during my Ph.D. study at the University of Kentucky.

During my PhD program I had the opportunity to work as a graduate research assistant in the Institute for Biomedical Informatics within the college of medicine where I had the honor to work with many wonderful scientists. My special thanks goes to Dr. Jeffery Talbert, Dr. Linda Dwoskin, Dr. Heather Bush, Dr. Amy L. Meadows, Dr. Lars E. Peterson, Dr. Barbara S. Nikolajczyk, and Dr. Ramakanth Kavuluru. I have truly enjoyed working with this interdisciplinary team and learned a lot from them. I would also like to thank the wonderful research and administrative staff at the Institute for Biomedical Informatics: Steve Roggenkamp, Connie Vaughn, Jill

Cioci, Darren W. Henderson, Brandon L. Muncy. Especially, I am grateful to Steve Roggenkam for his help with data extraction and I've learned a lot from him. I would like to extend my gratitude to Dr. Sunghwan Sohn who mentored me during my summer internship at Mayo Clinic, Rochester MN which contributed to a part of this dissertation. I also want to acknowledge my lab mates Taylor Smith, MD Selim, Lucas Liu and others. The help and support I received from my lab mates has played a very important role in finishing this dissertation.

Last but not least, I would like to express my sincere gratitude to my family, my mom, dad, brothers and sisters, for their endless love and support to me thorough my PhD study. They may have been physically far from me, but their love and support have been close during the past couple of years while I've been a student at the University of Kentucky. I would also like to thank my girlfriend Robin Thompson for all her support and love. In conclusion, I am grateful for the support and help I have received from all of those whom I did not mention explicitly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 Introduction

In modern artificial intelligence, longitudinal data encompasses all tasks where a response is observed on each subject repeatedly over time. What distinguishes longitudinal data from the traditional data sets is that in longitudinal data a temporal pattern of events is included in the data set in addition to the events' outcomes. The most common examples of longitudinal data include healthcare data, speech and natural language data. In healthcare data, patients outcomes and possibly treatments or procedures are collected at multiple follow-up times. Analyzing this longitudinal healthcare data and extracting meaningful knowledge from these growing data sets is critical in addressing many real world problems in healthcare. However, analyzing real-world healthcare data is a complicated task with computational challenges including high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity[1]. In particular, healthcare data is typically collected from multiple sources, and the subsequent data analysis requires simultaneous analysis of the temporal correlation among multiple streams of different data sources such as medications, diagnoses, and procedures. This study, addresses the challenges in analyzing longitudinal healthcare data using deep learning models and proposes a novel deep learning based model to analyze large scale claims data sets. Claims data, also known as administrative data, are large healthcare data that include person-specific clinical utilization, expenditures and enrollment across inpatient, outpatient, prescription drug and carve-out services.

1.1 Longitudinal healthcare data analysis using deep learning

Deep learning models have demonstrated great potentials in addressing some of the challenges in analyzing longitudinal and sequential data and created promising data analysis tools. Recurrent neural networks (RNNs) such as long short term memory models (LSTMs)[2] are a variety of deep learning model that have shown promising performance in longitudinal and sequential data analysis and gained popularity in many natural language processing (NLP) tasks. RNNs have an important characteristic that enables them to deal with the challenges in longitudinal and sequential data analyses. They are capable of extracting contextual information from past time steps and pass this information forward, which helps them to efficiently model long term dependencies in input sequences[3]. Nevertheless, the network architecture and

design preclude RNNs from processing long streams in a reasonable amount of time[4]. Attention mechanism was introduced in RNNs to increase their capacity in capturing long range dependencies more efficiently[4–6]. Attention-based models bridge the gap between different states in RNNs using a context vector. Successful applications of multiple attention layers led to the transformer model[7], which removed recurrence in RNNs relying entirely on the attention mechanism. These novel models yielded superior sequence analysis and outperforms state of the art models in almost all of the natural language processing tasks.

Both RNN and transformer have been successfully used to create longitudinal healthcare data analysis tools during the past couple of years. Among the deep learning based healthcare data analysis tools, Doctor AI[8], RETAIN[9], and DeepCare[10] modeled multiple data streams including medications, diagnoses, and procedures using recurrent neural network models such as long short term memory models[2]. Doctor AI concatenated multi-hot input vectors to predict subsequent visit events[8]. RETAIN used two separated RNNs to generate attentions at the visit level and the variable level as well[9]. These applications demonstrate that RNNs are promising in longitudinal and sequential healthcare data analysis, since RNNs are capable of extracting contextual information from past time steps and pass this information forward; this helps to efficiently model long-term dependencies in patients' data[3]. However, as mentioned above the network architecture and design preclude RNNs from processing long streams in a reasonable amount of time[4] and attention based models including the transformer were introduced to address this issue. Transformers were applied on longitudinal electronic health records (EHR)[11] to predict patients' outcomes in the future. There are also several models that have been successfully applied on EHR data without significantly changing the network architecture or loss[12–14]. Of course, the typical transformer's structure can be altered to better fit the special needs of solving healthcare problems[11,15]. Choi et al. proposed a transformer model for healthcare data analysis by utilizing the conditional probabilities calculated from the encounter records to guide the self-attention mechanism in the transformer[11]. BEHRT[15] was developed based on BERT[16], a popular transformer model for NLP tasks, for analyzing EHR data. BEHRT considers the patients' existing diagnoses and demographic data to predict their future diagnoses.

Similar to RNNs, transformers have been modified to model multiple data streams. Li et al developed a two-stream transformer to analyze both time-over-channel and channel-over-time features in human activity recognition tasks[17]. Two parallel, yet separate transformers were used to handle two input streams. Another multi-stream

transformer has been developed to generate effective self-attentions for speech recognition[18]. They parallelized multiple self-attention encoders to process different input speech frames. Gomez et al. developed a multi-channel transformer for sign language translation using one self-attention encoder[19]. Their model finds the attentions across three different channels, i.e. hand shapes, mouthing, and upper body pose. A more recent work[20] showed that "transformer is all you need" by using multiple transformer encoders. The encoded outputs can be concatenated using a joint decoder that enables simultaneous model training. There are also works that analyze multi-stream data using transformer by simply stacking or parallelizing multiple transformer models[21,22].

1.2 Motivations

Longitudinal healthcare data analysis is critical in addressing many real world problems in healthcare including substance use disorder (SUD). Early identification and engagement of individuals at risk of developing a SUD is a critical unmet need and public health emergency[23,24]. The 2018 National Survey of Drug Use and Health reports that over 20 million people aged 12 and over have a SUD, with over 2 million reporting an opioid use disorder (OUD). Individuals with OUD often do not seek treatment or have internalized stigma about OUD that limits identification through traditional means, such as screening and clinical interview[25]. Significant disparities limit access to treatment for OUD resulting in less than 20% of all individuals with OUD receiving any form of treatment in the past year[26]. Since individuals are reluctant to seek treatment, one solution is to identify patients otherwise engaged in the healthcare system where a majority of ambulatory care is provided, primary care[27]. While there are currently tools developed to predict aberrant behavior when prescribing opioids[28] or to predict OUD from a general primary care population[29], there are only a few clinical tools, such as the Opioid Risk Tool[30], developed for assessing the risk of OUD. Typical clinician workflow does not allow for comprehensive OUD screening, but available administrative and clinical data have the potential to help clinicians identify and screen higher risk patients providing an opportunity for primary care professionals to play a greater role in increasing OUD and SUD detection, treatment, and prevention. Healthcare data including claims data is a growing source of information that can be harnessed together with deep learning and machine learning to advance our understanding of factors that increase the propensity for developing OUDs and SUDs as well as those that aid in the treatment of the

disorders[31,32].

Although the recently developed deep learning based models such as transformer models have shown promising performance, the potential of applying transformers on claim data analysis has not yet been fully explored. One of the major limitations of the current deep learning models is the lack of capacity to model multiple data streams within the self-attention layer. The transformer was originally designed to process one data stream, which is mostly an order of words in a NLP task, at a time. The modified transformers either can only handle multiple streams at intra-stream level or they are not suitable to solve SUD and OUD prediction problems targeted in this thesis. Here, OUD prediction is complex data analysis task that includes not only finding long term effects of prescription opioids such as morphine and fentanyl, history of diagnoses such as mood disorders, but also the hidden associations between patient's prescriptions and diagnoses, since these input streams are highly correlated with each other. Identifying the relationships within and between input streams may reveal hidden patterns leading to an increased classification ability and interpretability for OUDs. Moreover, the medication application patterns and the interactions between medications across different visits as well as the patient's diagnoses and procedures patterns thorough his/her medical history may carry important information that should be extracted in order to develop precise and sensitive OUD identification and prediction tools.

This thesis, first proposes a framework to efficiently use RNNs including long short term memory models and the transformer in analyzing claims data sets and develop disease prediction tools. Then, the limitations of using transformers in healthcare data analysis are addressed and a novel transformer model to analyze longitudinal healthcare data collected from multiple sources is proposed. Leveraging the real-world longitudinal claims data, this work proposes a novel multi-stream transformer for longitudinal claims data analysis and disease prediction. The proposed model is designed to simultaneously analyze multiple types of data streams, such as medications, diagnoses, procedures and demographics, by attending to segments within and across these data streams. Multiple real-world problems in healthcare including substance use disorder, opioid use disorder and mild cognitive impairment are used to assess the proposed LSTM and transformer based models.

1.3 Contributions

The purpose of this thesis is to address challenges in analyzing longitudinal claims data using deep learning and develop novel models for disease predictions. To this end, multiple deep learning based models are proposed. First, this work addresses the technical challenges in analyzing temporal claims data and presents a new framework to perform disease prediction tasks using recurrent neural networks. This framework has three components: 1) data pre-processing, 2) disease prediction using RNNs, and 3) hypothesis exploration by varying the model and inputs. Second, this thesis extends and applies the transformer to real-world longitudinal claims data analysis tasks. Transformer traditionally processes one stream of inputs which is mostly an order of words in a NLP task. However, claims data analysis tasks require the model to analyze multiple input streams at the same time and extract the associations within and between all input sequences. In healthcare, the data is normally from multiple resources and each patient's data include multiple streams such as medications, diagnoses, procedures and demographics. Further, the attention mechanism in transformer can be utilized to make the deep learning models more transparent in applications such as healthcare data analysis where finding the association between events plays an important role in decision making. This work addresses these challenges and proposes a novel transformer to efficiently process the claims data and produce disease prediction models. The contributions of this thesis are:

- Producing an open source tool to convert large scale administrative claims data to a temporal format appropriate for training deep learning models.
- Proposing and developing a novel multi-stream transformer model for disease prediction using large administrative claims data. This proposed model is designed to simultaneously analyze multiple types of data streams including medications, diagnoses, procedures and demographics by attending to segments within and across these input data streams.
- Training multiple machine learning and deep learning models using data from more than 390K patients over 10 years to predict the onset of opioid use disorder.
- Creating a framework to predict the onset of substance use disorder among adolescents with attention deficiency hyperactivity disorder using long short term memory models.

- Proposing a long short term memory based framework to analyze the potential for patient clustering using routinely-collected EHRs and predict the progression from cognitively unimpaired to mild cognitive impairment.
- Producing risk identification and data visualization tools for longitudinal claims data.

1.4 Overview of Thesis

This thesis is organized into five chapters: fundamentals are presented in chapter 2, chapter 3 introduces the recurrent neural network based models developed in this thesis to perform disease prediction tasks. In this chapter, substance use disorder and mild cognitive impairment are used as use case diseases to predict their onset using the developed RNN based models. Chapter 4 describes the proposed multi-stream transformer models. This chapter is focused on predicting opioid use disorder using the proposed multi-stream transformer. Chapter 5 provides detailed discussions on the methods and results as well as the limitations of the current work and some suggestions to improve upon this work and guidelines for future works.

Chapter 2 Technical Background

In this chapter, material and methods that are used in building the proposed models are described. Recurrent neural networks are presented in section 2.4. This chapter also covers attention mechanisms and backgrounds on the transformer model in sections 2.6 and 2.7, respectively. Classical machine learning models including random forest, visualization and clustering methods as well as the data set used in this work are also covered in the current chapter.

2.1 IBM MarketScan Data Set

Here in, the large-scale administrative records in the IBM Health MarketScan Commercial Claims database (formerly known as Truven) for the years 2009 to 2020 were used to train and test baseline and proposed algorithms. Data include person-specific clinical utilization, expenditures and enrollment across inpatient, outpatient, prescription drug and carve-out services. This database contains about 30 million enrollees annually across the US, and these enrollees are nationally representative of the US population with respect to sex (50% female), regional distribution, and age.

Though the population is disproportionately privately insured and middle class, the very large sample supports well-powered subgroup analysis. The IBM MarketScan databases link paid claims and encounter data to detailed enrollee information across sites, types of providers, and over time. Historically, more than 20 billion service records are available each year. These data represent the medical experience of insured employees and their dependents for active employees, early retirees and Medicare-eligible retirees with employer-provided Medicare Supplemental plans. To provide an example of the sample size available when selecting specific subgroups, Figure 2.1 shows the 2004 data cohort for members with attention deficiency hyperactivity disorder (ADHD) and prescribed stimulants, along with their yearly follow-up. In 2004 there were 283,421 members with ADHD, with the sample size reduced each successive year due to follow-up. Each successive year would have a similar cohort and similar follow-up, with the combined sample population yielding over 500,000 members that will produce well powered subgroup analysis. Further, Table 2.1 provides some basic statistics on the amount of data available from IBM MarketScan data that is available in this work.

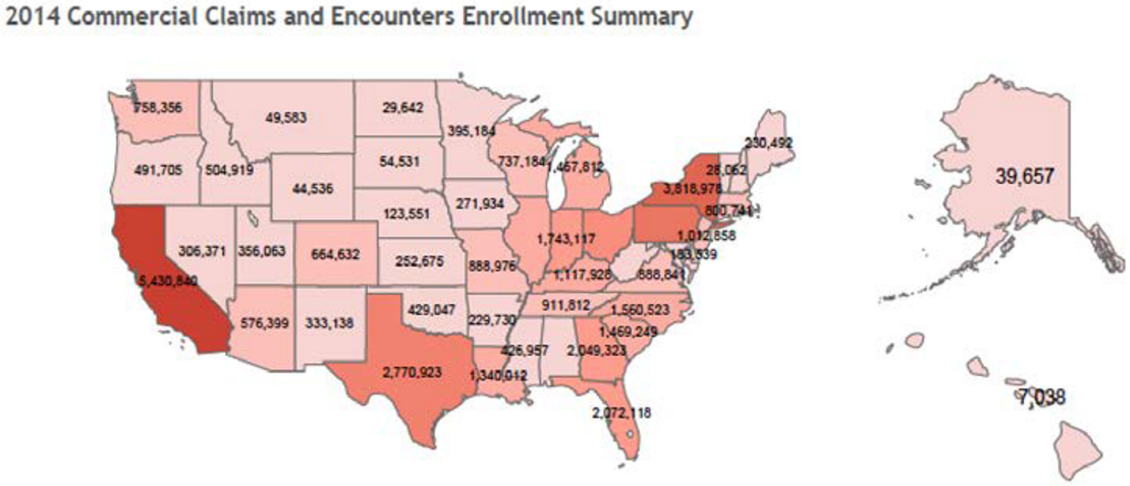


Figure 2.1: IBM MarketScan Health Analytics sample size for 2014 for each state.

2.2 Classical Machine Learning Models

Here, random forest and other classical machine learning models used as baselines in this thesis are introduced.

2.2.1 Random Forest

Random forest[33] are ensemble models that can be used to solve prediction tasks such as disease prediction. This model operates by constructing a multitude of decision trees at training time and has been used extensively to solve predictions tasks in healthcare data analysis or served as a baseline when deep learning models are used to create predictive modelings. The goal is to create a predictive model to predict $Y \in \mathbb{R}$ given a training data set $S_n = (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ of independent random variables distributed as the independent prototype pair (X, Y)[34]. For each tree T_j in a forest including M trees, the predicted value for the input sample x is denoted by $m_n(x; \Theta_j, D_n)$, where $\Theta_1, ..., \Theta_M$ are independent random variables, distributed the same as a generic random variable Θ.

$$m_n(x; \Theta_j, S_n) = \sum_{i \in S\prime_n(\Theta_j)} \frac{\mathbb{1}_{x_i \in A_n(x; \Theta_j, S_n)} Y_i}{N_n(x; \Theta_j, S_n)} \tag{2.1}$$

Where $S\prime$ is a subset of data and $A_n(x; \Theta_j, S_n)$ is the leaf containing x and

Table 2.1: Statistics of the subset of IBM MarketScan data available in this work. The data includes prescriptions, diagnoses, procedures and demographics information for a large and nationally representative sample of the US population.

Measurement	**Number of records/patients or date**
Demographics	
Female patients	84,114,853
Minimum birth date	1889
Maximum birth date	2020
Data availibility	
Earliest service/fill date	01/01/2009
Latest service/fill date	06/30/2020
Number of records	
Number of prescription records	3,865,125,856
Number of diagnose records	8,171,102,764
Number of procedure records	10,917,467,123
Number of patients	
Unique patients in demographics table	164,148,434
Unique patients in prescriptions table	95,841,261
Unique patients in diagnoses table	133,958,766
Unique patients in procedures table	133,370,031

$N_n(x; \Theta_j, S_n)$ is all the points that are included in $A_n(x; \Theta_j, S_n)$. The random forest algorithm then combine all the trees to create the final random forest:

$$m_M, n(x, \Theta_1, \Theta_2, ..., \Theta_M, S_n) = \frac{1}{M} \sum_{j=1}^{M} m_n(x; \Theta_j, S_n) \tag{2.2}$$

2.3 Clustering and Visualization

In this work, multiple methods have been use to cluster and visualize the data sets at different stages of the projects. This section provides a background on the main clustering and visualization methods used thorough this thesis.

2.3.1 K-Means clustering

The k-means algorithm[35] is an unsupervised learning algorithms that has been utilized in many problem domains. In k-means clustering algorithm, n input patterns are divided into k clusters in which each pattern belongs to the cluster with the nearest mean (cluster center). The k-means algorithm places the cluster centers as distant as possible from each other. This method can be summarized in the following steps:

- Initialize K points within the feature space of the training samples randomly, as initial clusters' centers.
- Assign each training sample x to the cluster with minimum euclidean distance between the cluster center and x (Equation 2.3).
- When all training samples have been assigned, recalculate new means for each cluster and consider these mean values as new centers for the new clusters.
- Repeat Steps 2 and 3 until the centers no longer move. This algorithm minimizes the error function in Equation 2.4.

$$d(x,u) = \sqrt{\sum_{i=1}^{n}(x_i - u_i)^2} \tag{2.3}$$

$$j = \sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{(j)} - c_j\|^2 \tag{2.4}$$

In Equation 2.4, k and n are the number of clusters and the number of training samples, respectively. Also, $\sum_{j=1}^{k}\sum_{i=1}^{n}\|x_i^{(j)} - c_j\|^2$ indicates the Euclidean distance from the sample $x_i^{(j)}$ to the cluster center c_j. Then, a new sample is characterized by:

$$f(x) = min_j(x - c_j)^2 \tag{2.5}$$

2.3.2 t-Distributed Stochastic Neighbor Embedding

t-Distributed Stochastic Neighbor Embedding (t-SNE)[36] is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data. The technique is a variation of stochastic neighbor embedding[37] that is much easier to optimize, and produces significantly better visualizations by reducing the tendency to crowd points together in the center of the map. The t-SNE algorithm calculates a similarity measure between pairs of instances in the high dimensional

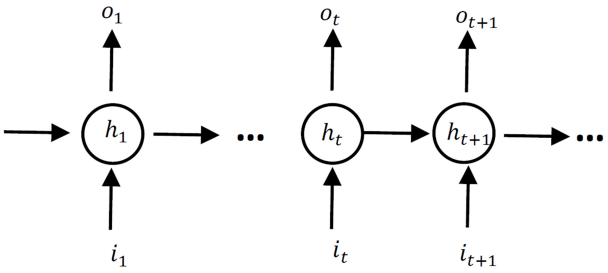


Figure 2.2: Unrolled structure of RNNs. The circles present hidden layers, i_t , o_t and h_t are respectively input, output and hidden state at time step t.

space and in the low dimensional space. This algorithm attempts to optimize visualization in a low dimensional space by matching the distributions using KL divergence. These pairwise similarities in higher and lower dimensional spaces are modeled using conditional probabilities[38]:

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)} \tag{2.6}$$

Where $p_{j|i}$ is the conditional probability that a point j is a neighbor of point i in the higher dimensional space and it models the target distribution of pairwise similarities in the lower dimensional embedding space using a Student's t-distribution around each data point to overcome the over- crowding problem in the Gaussian distribution:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}} \tag{2.7}$$

t-SNE then minimizes the KL divergence between the distributions, which conserves the local structure of data points across the higher and lower dimensional spaces.

2.4 Recurrent Neural Networks

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them powerful models in processing longitudinal data. Figure 2.2 shows a full RNN structure, where the circles represent the network layers and the solid lines represent the weighted connections[3].

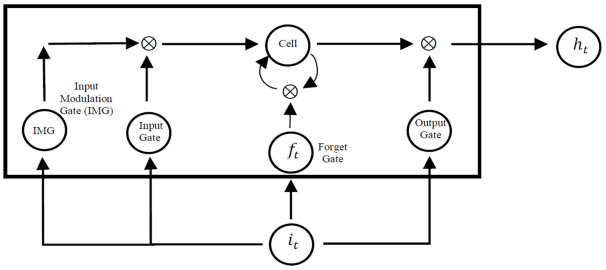


Figure 2.3: A LSTM memory cell. Each LSTM unit includes the input gate, the output gate and the forget gate.

RNN computes a set of hidden states $h = (h_1, h_2, ..., h_T)$ as well as output vectors $y = (y_1, y_2, ..., y_T)$ for a given input sequence $x = (x_1, x_2, ..., x_T)$ when T is the number of time steps. RNN iterates over the input sequence and computes the hidden states and outputs using the following equations[39]:

$$h_t = F(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2.8}$$

$$y_t = W_{hy}h_t + b_y \tag{2.9}$$

Where the W is a trainable weight matrix, the b is a trainable bias vector and F is the hidden layer function. In the following section, Long Short Term Memory model which is one of the most popular RNN models have been described.

2.5 Long Short Term Memory Model

The long short term memory model is one of the most powerful and popular RNN models. The LSTM has already been deployed successfully in analyzing temporal data in many biomedical applications[40–43]. In particular, the LSTM alleviates the vanishing gradient problems[44] and bridges time intervals in a sequence. In a LSTM model, nodes are replaced with a unit called memory cell as shown in Figure 2.3[45]. A memory cell includes the input gate, output gate, and forget gate. The input gate decides how to update the cell state using the new input and the output gate determines how to filter the output. The forget gate decides which information the LSTM is going to forget; it considers both i_t and h_t and then, utilizing a sigmoid function, it generates a matrix with elements between 0 and 1. The previous cell state will be element-wisely multiplied by the numbers generated by the forget gate to determine how much information the LSTM unit wants to keep.

The most common LSTM architecture is given by the following equations[3]:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{2.10}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{2.11}$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{2.12}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{2.13}$$

$$h_t = o_t tanh(c_t) \tag{2.14}$$

Although LSTMs could partially solve vanishing gradients problems, it still suffers from the complexity of sequential path from older past cells to the current one. Moreover, LSTM's sequential nature precludes parallelization within training examples, which causes low computational efficiency in training using longer sequence lengths. This limitation motivated researchers to improve efficacy of RNNs using methods such as factorization tricks and attention mechanisms.

2.6 Attention Mechanism

Traditional sequence-to-sequence modeling include encoder and decoder where encoder map the input sequence to a fixed length vector and the decoder consumes this vector to generate the target sequence. However, it has been shown that the use of a fixed-length vector and simply feeding that into a decoder precludes improving the performance of this basic encoder–decoder architecture. Attention mechanism[5,6] boosted the RNNs performances in sequence modeling. Attention-based models bridge the gap between hidden states in encoder and the decoder using a context vector. In fact, the context vector computes the probability distribution of input sequence values at different time steps for each output that decoder generate each at each step.

The idea of a attentional model is to consider all the hidden states when deriving the context vector c_t. In this model type, a variable-length alignment vector a_t, whose

size equals the number of time steps on the source side, is derived by comparing the current target hidden state h_t with each source hidden state $\overline{h}_s$:

$$\alpha_t(s) = align(h_t, \overline{h}_s) = \frac{exp(score(h_t, \overline{h}_s))}{\sum_{s'}(exp(score(h_t, \overline{h}_{s'})))} \tag{2.15}$$

The score function is referred as a content-based function for which three different alternatives are considered in[5]:

$$score(h_t, \overline{h}_s) = \begin{cases} {h_t}^\top \overline{h}_s & \text{dot} \\ {h_t}^\top W_a \overline{h}_s & \text{general} \\ v_a^\top tanh(W_a[h_t; \overline{h}_s] & \text{concat} \end{cases} \tag{2.16}$$

Attention-based models showed promising performance in many NLP tasks and they later led to the transformers which outperformed all previous models in many language modeling problems.

2.7 Transformer: Attention is All You Need

The transformer was first proposed[7] based solely on attentions. Sequential nature of recurrent neural networks is a barrier to parallelization during training using long sequences. Attention mechanisms have been introduced to be used in conjunction with RNNs to model longer dependencies in sequences more efficiency. Transformers are model architectures that removed recurrence and instead relied entirely on an attention mechanism to draw global dependencies between input and output.Transformers were superior in quality while being more parallelizable and requiring significantly less time to train. They led to powerful neural machine translation models such as GPT[46], BERT[16], XLNet[47] and ALBERT[48].

Transformers, as neural sequence transduction models[49], consists of two main components: encoder and decoder. The encoder maps an input sequence $x = (x_1, ..., x_n)$ to a new representation $z = (z_1, ..., z_n)$. The decoder part auto-regressively generate the output $y = (y_1, ..., y_m)$ using z and previous outputs at each step. Figure 2.4 shows the overall architecture of the tsransformer. Both encoder and decoder layers include similar components: Multi-Head Attention, layer normalization, and position-wise fully connected feed-forward network. Multi-Head attention plays the most critical role within encoder and decoder layers.

The input to the attention function in transformer is set of queries, keys and values. The function calculates the weighted sum of input values where the weight

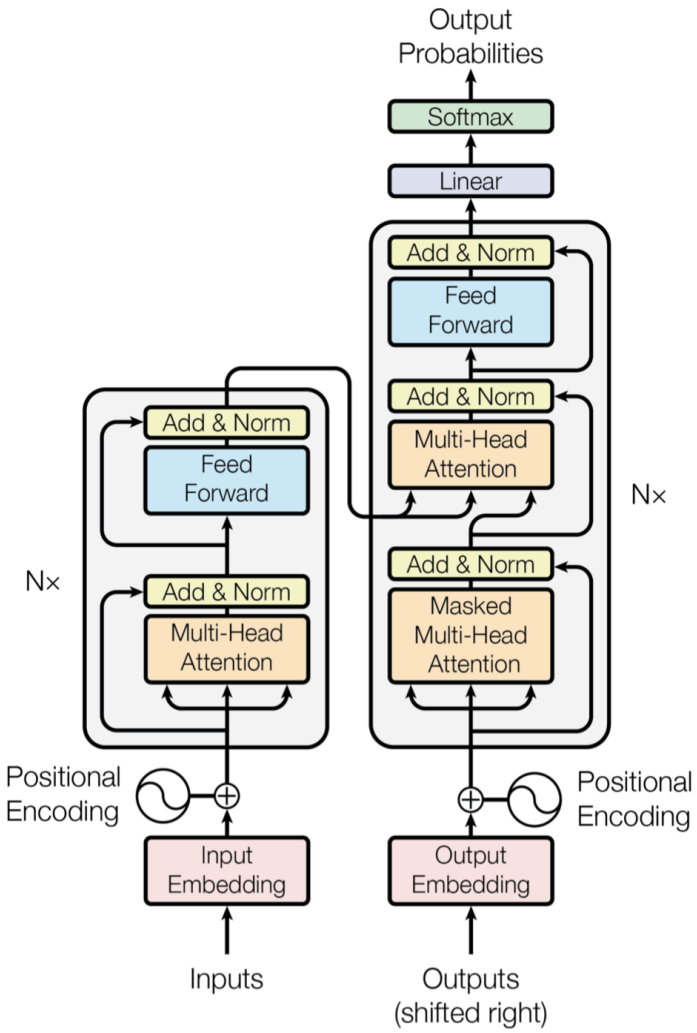


Figure 2.4: The Transformer - model architecture.

assigned to each value is computed using a soft-max of normalized dot product of query and key vectors. Mathematically speaking, the output vectors are computed as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.17}$$

Where, d_k is the dimension of the model and Q, K, and V are query, key and value vectors, respectively.

The transformer use another mechanism called multi-headed attention to first enables the model to focus on different positions of the input and second, provides the attention layer multiple representation subspaces. Multi-headed attention is simply running the attention procedure described above in parallel and concatenating the outputs:

$$MultiHead(Q, K, V) = Concat(head_1, ..., hean_h)W^O$$
$$where\ head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2.18)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}, W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$ *and* $W^O \in \mathbb{R}^{h_{dv} \times d_{model}}$.

2.8 Conclusion

This section provided detailed information on material and methods including IBM MarketScan data, classical machine learning models, recurrent neural networks, long short term memory model, and transformer models which are all parts of material and methods used in this thesis. In the following chapters, the above mention methods will be used to introduce the novel methods and frameworks proposed in this thesis.

Chapter 3 Longitudinal Data Analysis using LSTM

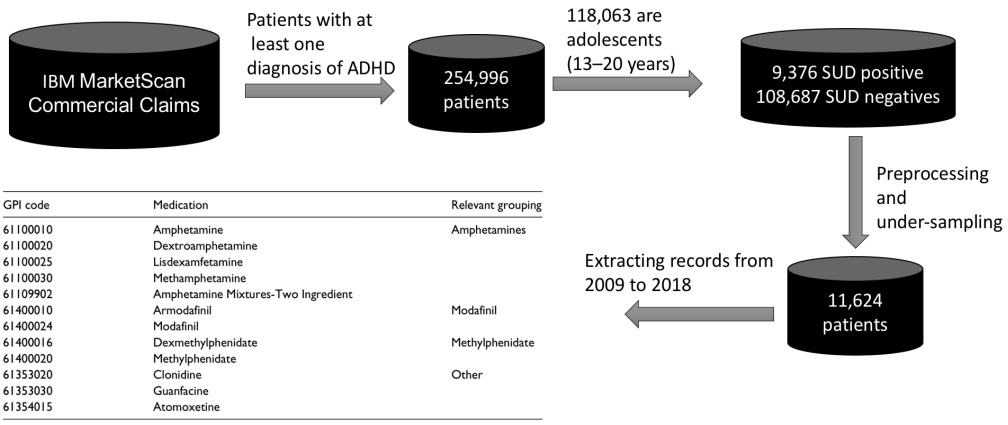
This chapter compares the performance of traditional machine learning and statistical methods to a proposed recurrent neural networks based model on two disease prediction tasks: SUD prediction in claims data and MCI prediction in EHR data. Bayesian statistics, Support Vector Machines, Random Forests, and Bayesian Statistics were used as traditional models to be compared with the deep learning models.

3.1 LSTM prediction of SUD

Here in, a deep learning based framework is proposed to capture the temporal patterns in patients information and predict the onset of substance use disorder. One of the potential risk factors for the development of SUD is an attention deficit hyperactivity disorder (ADHD) diagnosis. ADHD is one of the most prevalent neuropsychiatric disorders, with 6-10% of children (aged 2-17) received a diagnosis[50]. Given that 62% of individuals diagnosed with ADHD receive medication therapies[51], it is critical to systematically assess the long-term association of ADHD medication on subsequent risk of SUD.

The Spontaneously Hypertensive Rat (SHR) model of ADHD has been extensively used to assess the impact of exposure to specific ADHD medications[52–58]. Using the SHR model, it has recently been reported that never medicated rats with an ADHD phenotype self-administered more cocaine than those without an ADHD phenotype. Moreover, the ADHD medication introduced during adolescence was an important factor associated with a further increase in cocaine self-administration in young adulthood[52–58]. These preclinical findings suggest that medication type and age of medication initiation are critical factors in the relationship between ADHD pharmacotherapy and subsequent SUDs.

In this study, the technical challenges in analyzing temporal medication data were addressed and present a new framework to predict the long-term impact of ADHD medication initiated during adolescence. This framework has three components: 1) data pre-processing, 2) SUD prediction using RNNs, and 3) hypothesis exploration by varying RNN and inputs. Experimental results show that temporal medication features of ADHD medication initiation during adolescence, rather than stationary features (e.g., medication type, age, sex), are the most important factors on the health consequences related to SUD.



GPI code	Medication	Relevant grouping
61100010	Amphetamine	Amphetamines
61100020	Dextroamphetamine	
61100025	Lisdexamfetamine	
61100030	Methamphetamine	
61109902	Amphetamine Mixtures-Two Ingredient	
61400010	Armodafinil	Modafinil
61400024	Modafinil	
61400016	Dexmethylphenidate	Methylphenidate
61400020	Methylphenidate	
61353020	Clonidine	Other
61353030	Guanfacine	
61354015	Atomoxetine	

Figure 3.1: Cohort extraction and data pre-procesing schema.

3.1.1 Data Pre-processing

All the ADHD-related records from IBM MarketScan between Jan 2009 and Dec 2015 were extracted. Figure 3.1 shows the flowchart of cohort selection and data pre-processing. All the 254,996 individuals who had an International Classification of Disease (ICD-9) diagnosis of ADHD (ICD-9 code 314.X) were selected; among them 136,933 are children (6-12 years) and 118,063 are adolescents (13-20 years) onset exposure to ADHD medication. For each of the enrollees with an ADHD diagnosis, all the ADHD medication records between Jan 2009 and Dec 2015 were extracted. In total, 11,778,912 records from IBM MarketScan were extracted.

The original format of the prescription and professional service encounter claims in IBM MarketScan is a table where each row is a visit and columns are enrollee ID, date of visit, and prescription. If an enrollee has multiple visits, each visit will occupy a row in the table. To facilitate further study of the temporal patterns in the data, the IBM MarketScan format was converted into an enrollee-time matrix $X(P, T)$, where P is the complete set of ADHD enrollees and T is the set of time points between Jan 2009 and Dec 2015 (by month), each cell x_{ij} records the medications an enrollee p_i took at time t_j. Figure 3.2 shows the cohort definition and describes the process mentioned above. Here twelve ADHD medications were considered and categorized into four medication groups, i.e. amphetamines, methylphenidate, modafinil, and others, based on the first eight digits of the generic product indicator (see Table 3.2

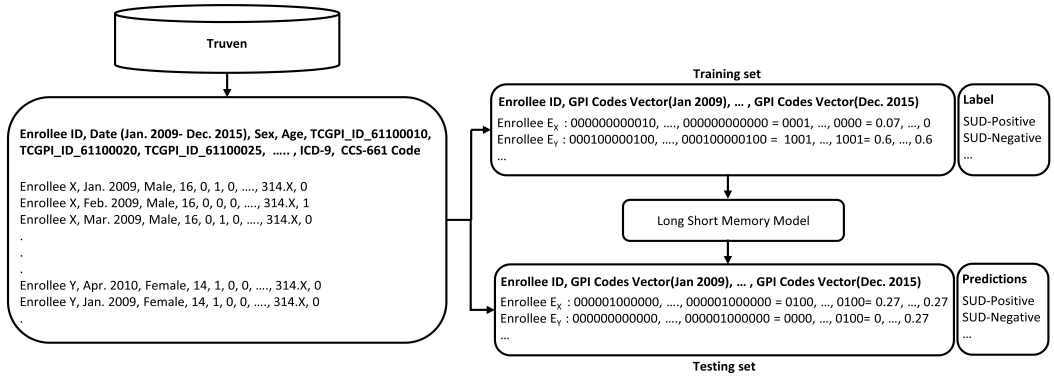


Figure 3.2: Cohort definition: All ADHD-related records from IBM MarketScan between January 2009 to December 2015 have been extracted. The SUD-cohort includes ADHD individuals who has been diagnosed as having a SUD for the first time after he/she has received a prescription for an ADHD medication for at least five months. The SUD-control cohort includes ADHD individuals without any diagnosis of SUD during the follow up period.

for details). Given the monthly subscription records in IBM MarketScan, the finest temporal resolution of T is by month.

Enrollee's demographic information including age and sex, were directly plugged into the longitudinal ADHD medication records, and then trained the LSTM. Specifically, for enrollee p_i at time step t_j, the age of p_i at t_j were converted into to a bit vector, which was then concatenated the vector with the ADHD medication record in cell $x(i, j)$. Similarly, an additional bit was used to incorporate the sex information. This new format allows to model both temporal and stationary features simultaneously.

If an enrollee in P has been diagnosed as having a SUD for the first time after he/she has received a prescription for an ADHD medication for at least five months were determined. If this condition is met, the enrollee is labeled as *ADHD-SUD positive* (label $y_i = 1$ for enrollee p_i) and all the ADHD medications after the first SUD diagnosis will be removed from $X(P, T)$; otherwise the enrollee is labeled as *ADHD-SUD negative* (label $y_i = 0$ for enrollee p_i). The extracted IBM MarketScan data include detailed enrollee-level information over time, ready for assessing the long-term impact of ADHD medications. Figure 3.3 shows a simplified example of the longitudinal medical record data that have been created for an enrollee, and Figure 3.4 visualizes the entire medical record data. Each row in Figure 3.4 has a similar format to what is described in Figure 3.3; each pixel in each row is related to a time stamp (one month) and its color shows the ADHD medication that an enrollee

Patients ID	Time step 0	Time step 1	...	Time step $n-1$	Time step n	Label
ID1	ADHD Med. Code	ADHD Med. Code	...	ADHD Med. Code	ADHD Med. Code	SUD Positive/Negative

Figure 3.3: A simplified example of the longitudinal medical record data. Each entry corresponds to a time step (a specific month) and shows the ADHD medications that enrollee was prescribed. If the enrollee did not use any medication during a specific month, the entry will be 0.

was prescribed.

Further, to determine type and pattern of medication exposure during adolescence, male and female IBM MarketScan enrollees who initiated any ADHD medication between 13-20 years old were selected (Table 3.1). Medication utilization was based on the first medication prescribed (i.e., initiated medication) and the medication prescribed in $\geq 50\%$ of medicated months (i.e., primary medication). Amphetamine was the first prescribed medication for over half, and methylphenidate was the first prescribed medication in over a third of male and female adolescents. Initial medication was the primary medication 90% of the time. No sex differences were found, with the exception of a greater lag time from ADHD diagnosis to first ADHD medication in males compared to females (1.6 and 0.3 months, respectively; not shown). Number and length of gaps in medication use averaged 1.5 breaks for a total of 3 months over the course of 12 months. Among enrollees who switched medications, an average of 1.6 medication switches occurred (Figure 3.5, heat map for visualization of 16-year old sample). When amphetamine was both the initial and primary medication, an SUD diagnosis was found in 10.2% of enrollees. In contrast, when methylphenidate was both the initial and primary medication, an SUD diagnosis was found in 8.2% of enrollees. Although preliminary, this suggests that amphetamine increases risk for SUD. It is important to consider other factors, such as role of ADHD severity in medication prescribing, as it may be disease severity, rather than specific medication that increases SUD risk.

On average, every enrollee has 74.4 visits and among them, the average number of visits for ADHD-SUD positive enrollee is 61.2. Regarding the type of ADHD medications that enrollees used, 39.3% of enrollees used amphetamine as their initial ADHD medication while 52.9% initiated with methylphenidate. Only a small portion (almost one percent) of enrollees were prescribed both as their first prescriptions. A medication is called the primary medication for an enrollee if the prescriptions for that medication occupy more than 80% of that enrollees ADHD medication prescriptions.

Table 3.1: Utilization of ADHD Medications initiated during adolescence (13-20 years) by sex. Amphetamine was the most common initiated and primary medication prescribed during adolescence, followed by methylphenidate. Primary medication designation is defined as the medication that was prescribed on $\geq 50\%$ of medicated months. *On average, the primary medication was prescribed to enrollees on $\geq 95\%$ of medicated months.

Medication History	**Males (n=137, 921)**	**Females (n=88,834)**
-	n(%)	n (%)
Initiated Medication		
Amphetamine	73,262 (53.1%)	51,134 (57.6%)
Methylphenidate	51,751 (37.5%)	30, 592 (34.4%)
Other	12, 037 (8.7%)	6, 702 (7.5%)
None	871 (0.6 %)	406 (0.5%)
Primary Medication		
Amphetamine	75, 461 (54.7 %)	53, 309 (60.0 %)
Methylphenidate	48, 646 (32.3 %)	27, 733 (31.2%)
Other	10, 968 (8.0%)	5, 732 (6.5%)
None	2, 846 (2.1 %)	2.060 (2.3%)

Given the definition, 33.2% of enrollees are prescribed amphetamine as the primary ADHD medication and 40.2% of enrollees are prescribed methylphenidate as the primary ADHD medication. According to the above basic statistics, there is no significant bias in the data towards SUD. More sophisticated methods are required to extract knowledge from the IBM MarketScan data.

Note, although adolescent ADHD medication initiators in the IBM MarketScan data are represented less than child-initiators, for those who are ADHD-SUD positive, data from adolescent enrollees are abundant, whereas data from children are insufficient for this analysis (see Figure 3.6). Therefore, although a limited subset of IBM MarketScan data is used in this study, it is still appropriate to study the impact of the initiation of ADHD medication in adolescent enrollees.

Furthermore, two problems in the IBM MarketScan data were addressed first, i.e. data sparsity and label imbalance problems. First, data sparsity was reduced. This is an important problem since in the enrollee medication record matrix $X(P, T)$, 96.6% of medication records are simply empty (noted as zero). To address this, all

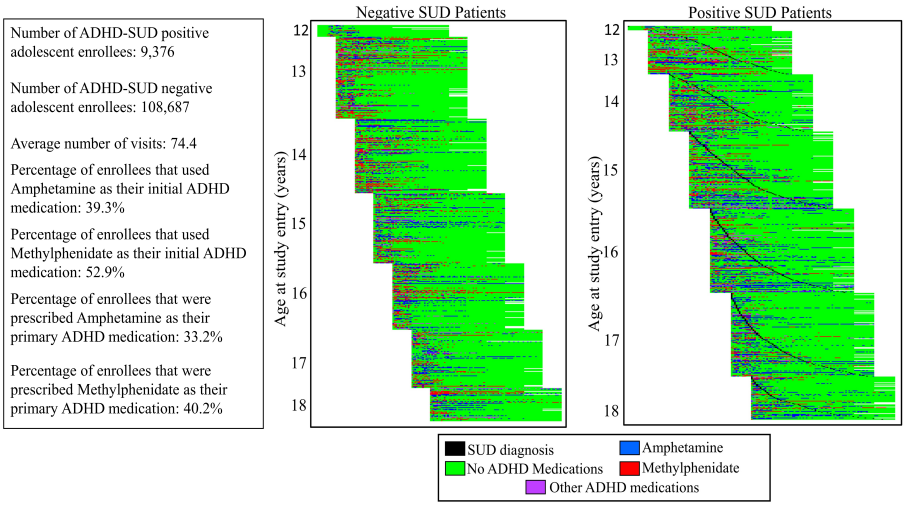


Figure 3.4: Visualization of the longitudinal data. Left figure describes SUD negative enrollees and right figure describes SUD positive enrollees. Enrollees are sorted based on their ages and in the right figure (SUD positives) they are also sorted based on the SUD diagnosis dates. Note, black dots highlights the time stamp when the ADHD enrollee is diagnosed with SUD.

Table 3.2: Generic product indicator of ADHD medications.

GPI Code	Medication	Relevant Grouping
61100010	Amphetamine	Amphetamines
61100020	Dextroamphetamine	
61100025	Lisdexamfetamine	
61100030	Methamphetamine	
61109902	Amphetamine Mixtures-Two Ingredient	
61400010	Armodafinil	Modafinil
61400024	Modafinil	
61400016	Dexmethylphenidate	Methylphenidate
61400020	Methylphenidate	
61353020	Clonidine	Other
61353030	Guanfacine	
61354015	Atomoxetine	

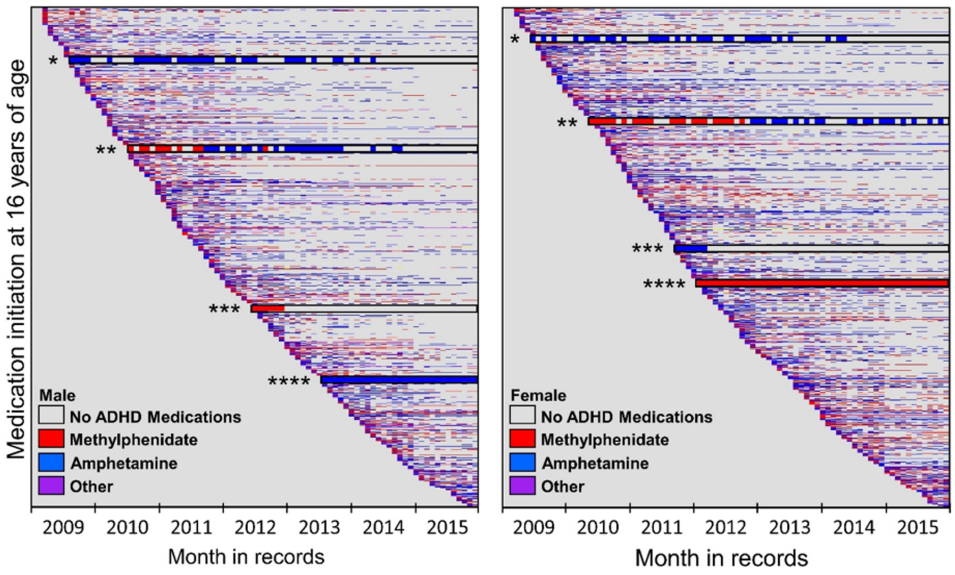


Figure 3.5: Heat map of IBM MarketScan members who initiated ADHD medication at 16 years of age to demonstrate sex differences in the utilization of ADHD medication across time (See Fig. 3-A, green box). Medication utilization in a sample of 1,000 male (left) and 1,000 female (right) enrollees with an ADHD diagnosis who initiated ADHD medication at 16 years. Each row represents one enrollee and colors represent the absence or presence of medication (methylphenidate, amphetamine, other) exposure on a monthly basis. Heat maps illustrate variability in medication utilization characteristics, including change in number and length of gaps*, medication types**, discontinuation of medication*** and continuous medication****.

the empty records were removed before the first ADHD medication record or after the last ADHD medication record. In addition, empty sequences, sequences in which the enrollee used ADHD medication for less than five months, and enrollees who started using ADHD medication less than five months prior to being diagnosed with SUD were all removed to further reduce noise and remove outliers. In the literature, three, six and twelve months break are commonly used as a baseline cut of value to determine whether the medicine initialization is authentic[59,60].

However, long ADHD breaks usually happen during the summer break which is approximately 2.5 to 3 months in the United States. It is clear that three-month break is too short and six or twelve-month break is too long. Alternatively, the five-month break was used based on our previous experiences with Spontaneously Hypertensive Rat (SHR) model of ADHD[52–58]. Using the five-month break also slightly enlarged the data set size from 10,836 to 11,462 (almost 6% increase) compared with using the six-month breaks. Second, the label imbalance problem were addressed. The

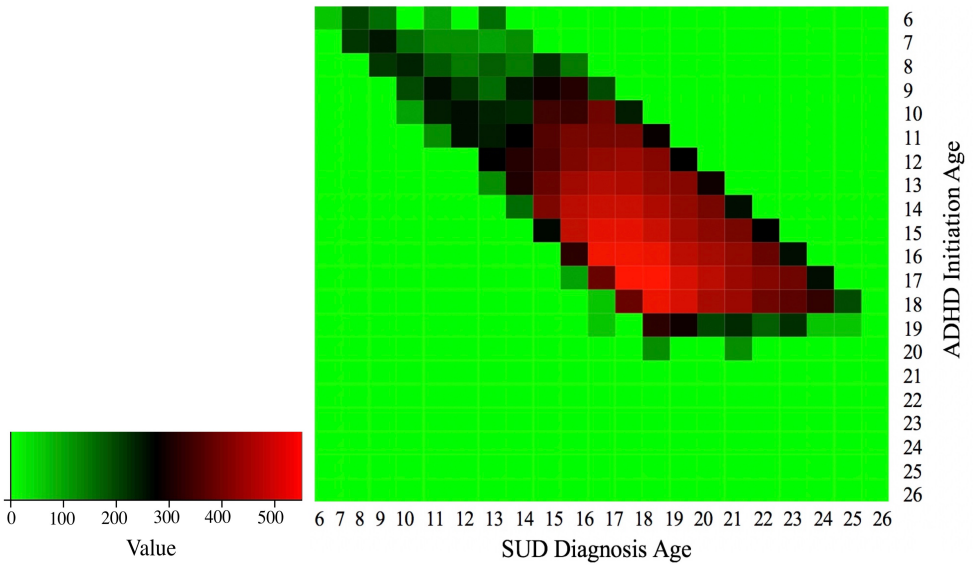


Figure 3.6: Visualization of data availability in the IBM MarketScan data. Red indicates abundant medical records, black indicates insufficient medical records, and green means NA. Only the ADHD enrollees who develop SUD are included. Figure shows that in the IBM MarketScan data, data from adolescent ADHD enrollees are abundant, while data from children with ADHD are insufficient for analysis.

IBM MarketScan data include in total 118,063 enrollees with adolescent initiation of ADHD medication, 9,376 of them are ADHD-SUD positive and 108,687 of them are ADHD-SUD negative.

The positive and the negative datasets are highly unbalanced and this can significantly affect training and optimization of any type of neural networks[61,62]. In this study, a data level approach called focused under-sampling[63] were adopted to deal with the class imbalance problem. Zero-padding were first applied to the input sequences so that the sequence lengths are the same, which makes it feasible for sequence comparison. Note, LSTM model handles variable length sequences and this zero padding strategy is only used during the pre-processing phase. Then, the cosine similarity[64] were applied to compare these sequences. The cosine similarity were defined between two sequences u and v as (assume each timestamp is i.i.d):

$$cosine(u, v) = 1 - \frac{(u.v)}{||u||_2||v||_2} \tag{3.1}$$

For every SUD-positive sample, one (or five) of the most similar SUD-negative samples from the ADHD-SUD negative enrollee pool were selected using cosine simi-

larity (Equation 3.1), thus a balanced (or a slightly unbalanced) dataset was generated to train the models.

3.1.2 Experimental Settings

All the LSTM models were deployed on the TensorFlow platform[65] and were trained using eight GeForce GTX 1080 GPUs. Batch size, learning rate, number of hidden neurons and number of epochs were set to be 256, 0.09, 75 and 100, respectively. Support Vector Machine (SVM)[66], Random Forest (RF) models[67], and Bayesian Statistics as well as two dummy models which label all validation samples as either negatives or positives were compared with the LSTM models.

For systematic performance testing, each time 90 percent of the data were randomly selected as training data and the other 10 percent as testing data and this process was repeated 20 times. Averaged performance on the unseen folds was reported as the final result. Parameters of SVM and RF were tuned using a 20-Fold cross-validation and the averages of the best validation performance were reported in Table 3.4. Note that the performance of the baselines were even poorer than the validation performance. Accuracy, precision, recall, specificity and F1-score were used to evaluate and compare different models[68]. In all models, ADHD individuals with SUD were considered as the positive samples and ADHD individuals without SUD as the negative samples.

3.1.3 Bayesian statistics

IBM MarketScan is a complex national biomedical dataset. To better understand the data, Bayesian statistics were used to explore this data. Specifically, the conditional probabilities of SUD given different types of ADHD medications were computed and compared.

First, based on data availability, the IBM MarketScan data was stratified into six subsets based on ADHD medication initiation age (13-18 years). For every subset, the probability of the SUD diagnosis for enrollees with ADHD were calculated for two types of initial ADHD medications, i.e. methylphenidate, $P(SUD \mid M_{me})$ and amphetamine, $P(SUD \mid M_{am})$. These estimated probabilities (prevalence rates) can be included in a ratio to describe the relative probability (prevalence) of SUD for enrollees initiating on these medications. Table 3.3 shows an observed age-effect where those in the oldest age strata who initiated on amphetamine have elevated

Table 3.3: Estimated age-stratified probabilities. $P(M_{am})$ and $P(M_{me})$ are the probabilities of receiving amphetamine or methylphenidate as the first ADHD medication, respectively; and $P(M_{me} \mid SUD)$ and $P(M_{am} \mid SUD)$ are the probabilities of initiating on methylphenidate or amphetamine as the first ADHD medication within the population of enrollees with SUD.

Initiation age	$P(M_{me} \mid SUD)$	$P(M_{am} \mid SUD)$	$P(M_{me})$	$P(M_{am})$	$\frac{P(SUD\|M_{me})}{P(SUD\|M_{am})}$
13	0.58	0.41	0.59	0.41	0.99
14	0.52	0.48	0.56	0.44	0.93
15	0.52	0.48	0.54	0.46	0.97
16	0.47	0.53	0.51	0.49	0.93
17	0.42	0.58	0.47	0.53	0.91
18	0.37	0.63	0.43	0.57	0.90

probabilities (prevalence) of SUD, relative to those who initiated on methylphenidate (last column provides relative probability ratios or prevalence rate ratios).

However, these estimates are limited by only describing medication at initiation. This summary measure does not encompass a more comprehensive picture of medication utilization, and the over-simplification could miss important relationships between age at initiation, medication patterns, and subsequent SUDs in enrollees with ADHD.

3.1.4 Deep learning versus traditional methods

Here in, the SUD prediction performances of LSTMs, SVM, RF, and two dummy models were compared. Performance of the LSTM models, dummy models, and traditional classification models (SVM and RF) are provided in Table 3.4. In Table 3.4, the models "All-Negatives", "All-Positives", "RF", "SVM" and "LSTM" are all trained and tested on the balanced dataset. In Table 3.4, regarding the balanced IBM MarketScan data, the LSTM model has the highest accuracy (0.84), precision (0.96), specificity (0.97) and F1-Score (0.82). The results indicate that LSTM captures important factors in the IBM MarketScan data providing increased power to predict the development of SUD, while SVM and RF miss such factors.

In addition, model performances were tested on two unbalanced IBM MarketScan datasets, i.e. Unblnc1 (positive-negative ratio being 0.20) and Unblnc2 (positive-negative ratio being 0.08 when the complete IBM MarketScan data were used). Table 3.4 shows the model performances on Unblnc1 and Unblnc2; X-Unblnc1 repre-

Table 3.4: Performance of SUD prediction using traditional classification models and the LSTM model.

Model	Size of Data	Accuracy	Precision	Recall	Specificity	F1-Score
All-Negatives	11,624	0.50	0.00	0.00	1.00	0.00
All-Positives	11,624	0.50	0.50	1.00	0.00	0.67
RF	11,624	0.60 ±0.02	0.60 ±0.03	0.66 ±0.08	0.54 ±0.09	0.62 ±0.04
SVM	11,624	0.56 ±0.03	0.54 ±0.03	0.93 ±0.02	0.17 ±0.03	0.68 ±0.03
LSTM	11,624	0.84 ±0.01	0.96 ±0.03	0.72 ±0.02	0.97 ±0.03	0.82 ±0.01
SVM-Unblnc1	34,872	0.73 ±0.20	0.30 ±0.18	0.19 ±0.31	0.84 ±0.30	0.11 ±0.10
SVM-Unblnc2	78,493	0.80 ±0.29	0.41 ±0.40	0.16 ±0.35	0.85 ±0.34	0.04 ±0.04
RF-Unblnc1	34,872	0.84 ±0.01	0.74 ±0.11	0.09 ±0.02	0.99 ±0.01	0.16 ±0.03
RF-Unblnc2	78,493	0.93 ±0.01	0.93 ±0.07	0.06 ±0.01	0.99 ±0.01	0.10 ±0.03
LSTM-Unblnc1	34,872	0.95 ±0.01	1.00 ±0.01	0.68 ±0.02	1.00 ±0.00	0.81 ±0.02
LSTM-Unblnc2	78,493	0.97 ±0.01	1.00 ±0.00	0.62 ±0.15	1.00 ±0.00	0.75 ±0.16

sents a model X trained on Unblnc1 dataset. Table 3.4 shows that LSTM maintains high performance on both unbalanced datasets (LSTM-Unblnc1: precision=1.00, recall=0.68 and F1-score=0.81 and LSTM-Unblnc2: precision=1.00, recall=0.62 and F1-score=0.75). The high performance of the LSTM model indicates that the temporal medication records in the unbalanced IBM MarketScan data encode critical factors that provide an increased power to predict the development of SUD in adolescent ADHD enrollees, which were captured by the LSTM model.

Figure 3.7a illustrates the robustness of the LSTM model. The performance of the LSTM model on both the training and the validation datasets remains stable when different learning rates were applied. Figure 3.7b shows the model performance over a wide range of number of hidden neurons, and Figure 3.7c shows how the variation of the number of epochs affected the LSTM model performance. In both experiments,

the LSTM model performance remained stable.

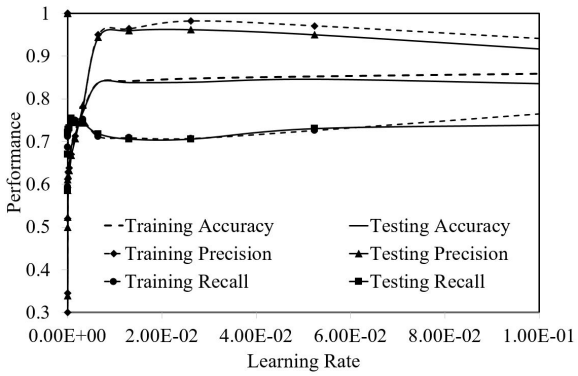
3.1.5 Hypothesis testing in longitudinal data using deep learning

To predict SUD from the temporal ADHD medication records $X(P,T)$, the LSTM model were adopted. Besides the temporal ADHD medication records, each enrollee has rich stationary features such as sex, ADHD initiation age, and medication type. Such stationary features could be critical factors towards the development of SUD. Two different approach were used here to test the functionality of the stationary features. In the first approach, stationary features, including age and gender, were encoded into the longitudinal ADHD medication records. In the second approach, the IBM MarketScan data was separated using each stationary feature as a partitioning criterion, such as separating enrollees into age groups, changing the granularity of the medication records, or ignoring medication-to-medication differences. Finally, LSTM models were trained on these datasets and compared model performance. If the model performance is significantly different than the model trained using complete data, then the selected enrollees regrouping criteria may be a critical factor towards the development of SUD because it provides an increased power to predict the development of SUD.

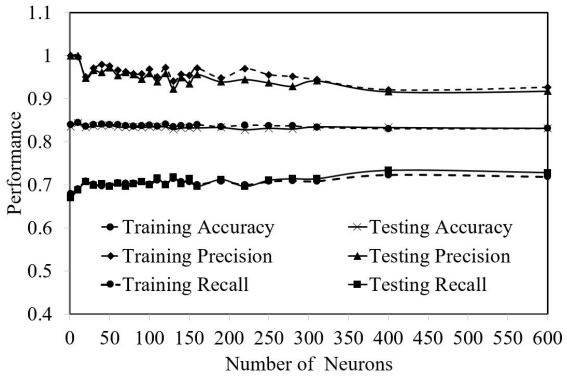
Note, an enrollee may be prescribed multiple medications at the same time point. Cell $x(i,j)$ in $X(P,T)$ is a vector of bits, each representing an ADHD medication and a bit is 1 if the corresponding medication is prescribed. Since the LSTM requires the input to be either a word or a value, the vector of bits were converted into an integer and then apply the Min-Max normalization to vanish the effect of scaling and map all values in $[0,1]$.

Knowing that in the temporal medication records there are critical factors that increase the power to predict SUD in enrollees who initiate ADHD medication during adolescence, the next step is to explore all of the important factors suggested from the literature (such as age of ADHD medication initiation and demographic information). To this end, demographic features, age and sex, as well as medication types were encoded into the temporal medication records. Table 3.5 shows the performance of a LSTM model trained only using the temporal pattern medication records (LSTM-NoDemo) and a LSTM model trained using additional demographic features (LSTM-Demo). As shown in Table 3.5, incorporating demographic features into the LSTM model neither increased nor decreased the model performance significantly.

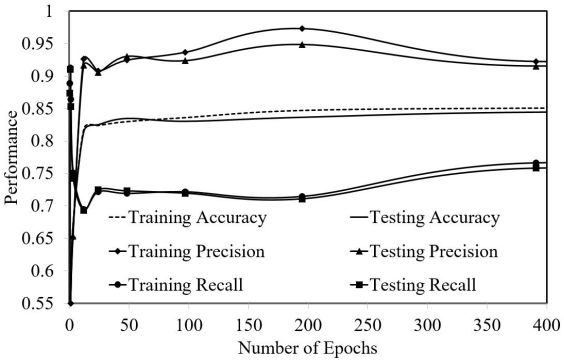
The effect of stationary features other than the demographic information, such as



(a) Impact of the learning rate on the LSTM model.



(b) Impact of the number of neurons on the LSTM model.



(c) Impact of the number of epochs on the LSTM model.

Figure 3.7: (a) The number of epochs is fixed to be 100. The learning rate varies in a wide range. (b) The learning rate and number of epochs are fixed to be 0.09 and 100, respectively. The number of neurins is changed from 1 to 600. (c) The learning rate and number of neurons are fixed to be 0.09 and 75, respectively. The number of epochs is changed from 0.1 to 400.

Table 3.5: Effect of encoding stationary features on the LSTM model performance on predicting SUD.

Model	Size of Data	Accuracy	Precision	Recall	Specificity	F1-Score
LSTM Demo	11,624	0.83 ±0.01	0.95 ±0.04	0.70 ±0.03	0.96 ±0.04	0.80 ±0.02
LSTM NoDemo	11,624	0.84 ±0.01	0.96 ±0.03	0.72 ±0.02	0.97 ±0.03	0.82 ±0.01

medication type, initiation age, sex, and temporal pattern of using medications on developing SUD were further tested in this study. The IBM MarketScan data was separated using each stationary feature as a partitioning criterion and then the LSTM models were trained on the separated data and compared the model performance.

First, whether the medication application pattern is a key factor for predicting SUD in adolescent ADHD enrollees was tested. Specifically, an enrollee may be prescribed a medication during a few months, stop for a few months and then start again being prescribed the medication. To test whether the on-off medication application pattern is a critical factor, a new dataset was generated by removing all the gaps from the ADHD medication records, and trained a LSTM model called LSTM-NoGaps using the newly generated data. For comparison, another dataset was generated, in which only the binary on-off medication application patterns were preserved. This dataset is used to train a LSTM called LSTM-OnOf. Table 3.6 shows the performance of LSTM-NoGaps and LSTM-OnOf models.

Table 3.6 shows that the model performance of LSTM-NoGaps is significantly worse than using the complete data. F1-Score drops from 0.81 to 0.53 when the gaps were removed. However, the model performance of LSTM-OnOff is similar to the LSTM model trained using the complete data. These results indicate that the on-off medication application patterns encoded in the temporal data are a critical factor for predicting the development of SUD in adolescent ADHD enrollees.

Second, duration of the medication application patterns was tested. Since in the IBM MarketScan data the longest ADHD medication records span seven years, a batch of new datasets was generated by only considering the ADHD medication records in the first x years (x varies from 1 to 7). Note, all of these new datasets include the same enrollees, but are across different ranges of time for their medication records. Figure 3.8 shows LSTM results on these 7 datasets. In Figure 3.8, LSTM-xyr represents a LSTM which is trained using the medication records in the first x

Table 3.6: Effect of medication application patterns.

Model	Size of Data	Accuracy	Precision	Recall	Specificity	F1-Score
LSTM NoGaps	11,624	0.56 ±0.01	0.58 ±0.06	0.55 ±0.21	0.57 ±0.21	0.53 ±0.11
LSTM OnOff	11,624	0.84 ±0.01	0.98 ±0.02	0.69 ±0.02	0.99 ±0.02	0.81 ±0.01

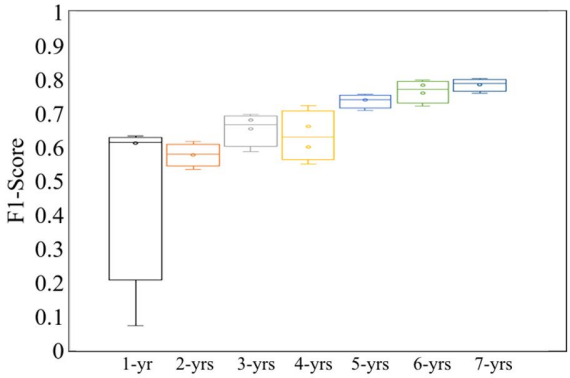


Figure 3.8: Impact of range of time of ADHD medication on SUD prediction. When the medication record range is increased from one year to seven years, the model performance increases.

years. Figure 3.8 shows that the F1-score of the LSTM-1yr model is low (average 0.48 and SD 0.27). However, the F1-score of the LSTM-2yr model increases significantly from 0.48 to 0.58, and continued to increase when longer medication records are used. These results indicate that the medication application patterns is a long-term pattern (at least longer than one year).

Since the long-term temporal medication application patterns appears to be a key factor in predicting subsequent SUD in adolescent ADHD enrollees, whether this association holds for different ADHD medication types or different sex groups was also tested. The LSTM-Methylphenidate model is trained using enrollee's data where methylphenidate is the primary medication ($\geq$ 80%). The LSTM-Amphetamine model is trained using enrollees' data where amphetamine is the primary medication. Table 3.7 shows that both models achieve a similar F1-Score (0.84 and 0.81) compared to the LSTM model trained incorporating both medications (0.82). These results indicate that long-term temporal medication application patterns exist using different ADHD medication types, but the exact pattern may be different. Table 3.7 also shows LSTM-male and LSTM-female models achieve the same F1 score (0.80),

Table 3.7: Performance of LSTM for adolescent ADHD enrollees with different primary medication or sex.

Model	Size of Data	Accuracy	Precision	Recall	Specificity	F1-Score
LSTM Meth.	2,076	0.85 ±0.03	0.95 ±0.03	0.75 ±0.04	0.96 ±0.02	0.84 ±0.03
LSTM Amph.	4,834	0.84 ±0.02	0.97 ±0.05	0.70 ±0.04	0.97 ±0.04	0.81 ±0.02
LSTM Male	8,468	0.83 ±0.02	0.94 ±0.07	0.71 ±0.05	0.94 ±0.07	0.80 ±0.02
LSTM Female	2,824	0.82 ±0.02	0.90 ±0.07	0.73 ±0.05	0.91 ±0.08	0.80 ±0.03

which is similar to the LSTM model trained incorporating both male and female enrollees (0.82). These results indicate that within different sex groups, the long-term temporal medication application patterns constitute important factors captured by the LSTM model.

Finally, the importance of the long-term temporal medication application patterns in different age groups was tested. Table 3.8 summarizes the LSTM model performance when applied to different age groups. All the F1-Scores are similar to each other and are similar to the F1-Score of the LSTM model trained using the entire data. These results indicate that in different age groups, the long-term temporal medication application patterns constitute the important factors captured by the LSTM model.

3.2 LSTM Prediction of MCI

Dementia is one of the most prevalent health problems in the aging population. It is estimated that by 2030, 75.6 million people will suffer from various types of dementia worldwide, and that this number will increase to 135.5 million people in 2050[69]. Such an increase will place a tremendous burden on patients, their families, society, and health care systems. Given that people with mild cognitive impairment are at an increased risk for dementia, predicting MCI risk and understanding the progression from cognitively unimpaired (CU) to MCI and dementia is a crucial task to help the aging population with their health needs.

In general, MCI is formally determined by health professionals through a comprehensive cognitive evaluation, together with clinical examination, medical history and

Table 3.8: Performance of LSTM for adolescent ADHD enrollees separated by age group.

Model	Size of Data	Accuracy	Precision	Recall	Specificity	F1-Score
LSTM Age13	1,704	0.81 ±0.03	0.87 ±0.08	0.74 ±0.06	0.88 ±0.09	0.79 ±0.03
LSTM Age14	2,040	0.81 ±0.03	0.88 ±0.04	0.73 ±0.04	0.89 ±0.04	0.80 ±0.03
LSTM Age15	2,396	0.82 ±0.03	0.92 ±0.06	0.70 ±0.04	0.93 ±0.06	0.79 ±0.03
LSTM Age16	2,282	0.82 ±0.04	0.88 ±0.10	0.76 ±0.06	0.88 ±0.12	0.81 ±0.03
LSTM Age17	2,002	0.82 ±0.02	0.89 ±0.04	0.74 ±0.04	0.91 ±0.04	0.81 ±0.02
LSTM Age18	1,062	0.83 ±0.04	0.87 ±0.08	0.79 ±0.07	0.87 ±0.10	0.82 ±0.04

often the input of an informant (an individual that know the patient very well) to understand changes in cognition and daily function. However, this is not routinely performed in many primary care visits which results in a delay of timely diagnosis, misses opportunities for appropriate care plans, and leads to adverse clinical outcomes. Electronic health records (EHRs), especially clinical free text, contain valuable information that is routinely recorded as part of clinical care. This rich information may be used to identify patterns predicting the development of MCI and dementia. Previous studies have shown that some signals of cognitive decline exist in EHRs, years before the clinician diagnoses of cognitive impairment[70,71].

Recently, deep learning models have demonstrated their capabilities in analyzing EHR data[72]. EHR data are a growing source of information that can be harnessed to provide an earlier diagnosis and to identify those at greatest risk for developing MCI. However, little is known about systematically analyzing patient data related to MCI in routinely-collected EHRs and how their temporal patterns are associated with the development of MCI. Although predicting MCI and understanding the progression from CU to MCI utilizing EHRs is a challenging and largely unexplored task, deep learning models can be used to capture temporal characteristics of patient data in EHRs to predict early stages of MCI. In this study, the application of EHR data analysis and deep learning models for predicting and clustering MCI patients has been systematically studies.

3.2.1 MCI prediction in the literature

Multiple studies to understand the progression from CU to MCI has been conducted. In a study by Pankratz et al.(2015)[73] demographic, clinical, and neuropsychological measures implemented in Cox proportional hazard models were applied to predict progression from CU to MCI. The authors showed that MCI risk factors presented in their previous studies[74–77] can be used to predict MCI using multivariate models. They used the Mayo Clinic Study of Aging (MCSA) cohort[78,79] and developed an augmented model capable of predicting progression from CU to MCI with AUC of 0.70. In a research conducted by Albert et al. (2018)[80], 224 CU participant were analyzed and followed up to detect measures or combination of measures that can be used to predict MCI. These researchers analyzed various MCI risk factors from different domains including cognitive, cerebrospinal fluid, magnetic resonance imaging (MRI), and genetic domain. They utilized time-dependent receiver operating characteristic and showed the feasibility of MCI prediction (best AUC using all variables > 0.83).

Another line of research is the prediction of dementia and understanding progression from MCI to dementia. Biomarkers, genetics, brain imaging as well as demographic and various variables related to individual's lifestyle have been used in the literature to predict progression from MCI to dementia with AUC ranging from 0.48 to 0.91[81]. Further, clinical variables and primary care data have been analyzed using logistic regression to predict progression of MCI to Alzheimer's disease (AD) dementia[82–85]. Ramakers et al.(2007)[86] used general practice data to create risk prediction models for dementia with sensitivity of 0.58 and specificity of 0.98 in the year before diagnosis.

With the availability of public databases related to AD dementia, such as the North American Alzheimer's Disease Neuroimaging Initiative (ADNI) and the European's AddNeuroMed Stud, big data has been utilized in AD research the past couple of years. Most of this research focused on diagnosing AD dementia, identify those at greatest risk of MCI, and predicting progression from MCI to AD dementia. Much of this work has utilized the non-community-based ADNI dataset and traditional machine learning models including support vector machines (SVMs), logistic regression and random forest[87]. SVM and linear models have also been used in the literature[88] to discriminate patients with AD from MCI and CU patients in ADNI dataset. Moreover, SVMs have been utilized in separating patients with AD dementia or MCI from CU patients using MRIs[89,90], and predicting progression from MCI to AD dementia using AddNeuroMed dataset[91]. MRI images from ADNI data set has also been used

in a study utilizing deep learning models conducted by Li et al.(2014)[92]. These results show that convolutional neural network based deep learning models can detect AD progression. In addition, neuroimaging, machine learning and deep learning has been used to predict conversion from MCI to AD in ADNI[93]. A comprehensive review of applying big data prospective and machine learning models to advance AD research is provided by Zhang et al. (2017)[87].

Multiple studies examined the progression from CU or MCI to dementia. However, only a few studies considered progression from CU to MCI or MCI prediction using patients EHR data. Often, MCI is not well recorded in EHR data because it is not a clinical diagnosis per se, and thus there are not enough datasets suitable to train MCI predictive models. In addition, discriminating CU patients from patients with MCI is a very challenging task as MCI is the stage between the expected cognitive decline of normal aging and the more serious decline of dementia[94]. In this thesis, the above-mentioned challenges are addressed and a potential of a deep learning model, coupled with natural language processing is demonstrated, to extract MCI risk factors and signals from unstructured EHRs and to predict onset of MCI. In addition, machine learning and deep learning techniques are utilized to visualize and cluster patients using EHR data and described a mechanism for EHR-based clustering.

3.2.2 Mayo Clinic Study of Aging Data

This study includes two main components: MCI prediction and patient clustering. In MCI prediction, a long short term memory technique architecture is trained to predict onset of MCI. In patient clustering, a denoising autoencoder is introduced to better represent the patient data. The outputs of this denoising autoencoder were visualized and clustered using t-SNE and K-means algorithms.

This study uses a longitudinal EHR data obtain from the Mayo Clinic Study on Aging (n=5,923; 1,376 MCI). The MCSA is a prospective population-based cohort study with comprehensive periodic cognitive assessment (at baseline and repeated every 15 months), initiated in 2004 to investigate the epidemiology of MCI and dementia. Eligible subjects from the Olmsted County, Minn., population, were randomly selected and evaluated comprehensively in person using the clinical dementia rating scale, a neurological evaluation and neuropsychological testing. A consensus committee used previously published criteria to diagnose the participants with normal cognition, MCI or dementia. MCSA participants have follow-up visits every 15 months and have accumulated more than 23,000 visits to date. In the current

analyses, only MCI patients who progressed from CU to MCI are included; i.e., the patients who were diagnosed with MCI in their first visit are excluded from the study to make sure the models are predicting initial MCI diagnosis. The MCSA cohort mainly encounters Mayo Clinic, Olmsted County Medical Center, and Mayo Clinic Health System for the regular healthcare. This study used clinical notes to automatically extract MCI risk factors and signals. To simplify the study, only patients who have any notes at Mayo Clinic—i.e. are included, excluded patients who do not have any clinical notes at Mayo Clinic during each visit interval. This reduced the size of cohort (n=3,265; 558 MCI). Further, the patient data after being diagnosed with MCI in MCSA were disregarded. Different types of data were used to predict MCI: demographic information, diseases/disorders, and neuropsychiatric symptoms, and activity of daily living (ADL). Table 3.9 contains a complete list of variables and their EHR sources that were used to train the models.

Total 783,090 clinical notes were used to extract diseases/disorders, neuropsychiatric symptoms, and other types of data (Table 3.9). To extract variables from clinical notes, the MedTaggerIE module in MedTagger[95] was used, which is the open-source clinical natural language pipeline developed by Mayo Clinic for pattern-based information extraction with a capability of assertion detection (i.e., negated, possible, hypothetical, associated with a patient). Note, only non-negated variables associated with patients were included.

3.2.3 MCI Patient Representation

The patients data were presented both in a temporal and static mode and used to train a temporal model (LSTM recurrent neural network) and a static model (random forest) for MCI prediction. To incorporate temporality, the data was converted into a visit-time format $X(V, T)$, where V is patients' visits and T is the visit dates, each visit v_i includes all variables for a given visit listed in Table 3.9, and each date t_i is the relevant visit date. All of the patients' visits for a period of 5 years before their first diagnosis dates for MCI patients and the latest visit for CU patients were used. A 15-month sliding window was used for the past 5 years of history of visits to make the temporal pattern asynchronous. Within each window an element-wise operation was used to combine the visits within the window. Further, the patients data was represented in a static mode to train the static model. In fact, instead of sliding a 15-month wide window, a 5-years window was used to cover the entire visit history of patients. As a result, the longitudinal data was converted to a matrix $Y(P, L)$, where

Table 3.9: List of variables and their resources.

Variable Category	Variable	Source
Demographics	Age, Sex, Education	MCSA
Diseases / disorders	Hypertension, Atrial fibrillation, Angina, Congestive heart failure, Coronary artery disease, Myocardial infarction, Coronary artery bypass graft, Diabetes	Clinical notes
Neuropsychiatric symptoms	Delusion, Hallucinations, Agitation, Depression, Anxiety, Euphoria, Apathy, Disinhibition, Irritability/lability, Motor behavior, Appetite/eating change	Clinical notes
ADL	Bathing, Dressing, Feeding, Housekeeping, Responsible for own medication, Transportation, Toileting, Transferring, preparing food	Patient provided information
Others	Slow gait, cognitive complaint, impaired judgment/orientation, memory concern, difficulty for concentrating, difficulty for finance	Clinical notes

P is the complete list of patients, each p_i is a vector including variables described in Table 3.9 and l_i is clinical diagnosis of MCI or CU of the patient p_i.

3.2.4 Denoising Autuencoders

Denoising autoencoders[96] have shown strong performance in efficiently representing participants data[97]. A four-layer denoising autoencoder was used in this study: an input layer, two hidden layers (encoder and decoder layers) and an output layer. Figure 3.9 shows the architecture of the network used in this study. The first hidden layer encodes the input x and then the decoder decodes the encoded vector (see Equation 3.2).

$$\begin{aligned} y &= \sigma(Wx + b) \\ z &= \sigma(W'y + b') \end{aligned} \tag{3.2}$$

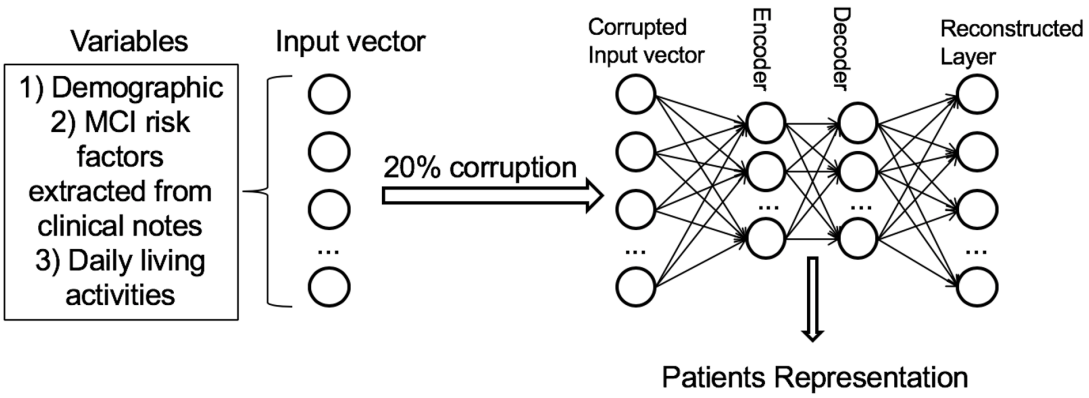


Figure 3.9: Architecture of the de-noising auto-encoder to represent patients using EHR .

20 percent of the patient's information was corrupted and the corrupted information was plugged as x to Equation 3.2. The loss function is a mean squared error of the output layer and the patient's information before the corruption. Adam optimizer at a learning rate of 0.01 was performed for 300 epochs to optimize the loss function. After training the autoencoder, all the samples were fed to the network without corruption and the outputs of the first hidden layer (y) were considered as a new representation for patient data. The new representation of data has lower dimension (60 nodes were used for both of the hidden layers) and is less sparse. The dimensionality of trained autoencoder's hidden nodes was further reduced using tSNE for visualization and clustering purpose. K-means was used for clustering with k=5 as a default value because there are 5 potential clinical subtypes: 1) CU patients, 2) MCI positive- amnestic, single domain, 3) MCI positive- amnestic, multiple domain, 4) MCI positive- non amnestic, single domain, 5) MCI positive- non amnestic, multiple domain.

3.2.5 Deep Learning prediction of MCI

The LSTM models were built under the Tensorflow platform[98] and were trained using two Tesla K80 GPUs. For systematic training and testing, a randomized cross validation method was used to find optimal parameters. First, the data was randomly

split into training (70 percent of data), validation (10 percent) and testing sets (20 percent). A stratified approach was used to split the data into train, validation and test sets; the patients were split based on their age at the study entry to make sure that different age groups are equally represented across training, validation and testing datasets. Each time a set of parameters was randomly selected in a predefined data pool; a LSTM model was trained on the training set and the model performance was measured on the validation set. This process was repeated 100 times and the best parameter set was selected based on the validation set performance. Then, the best model (based on their performance on the validation set) was used to evaluate prediction performance on the unseen test data.

Learning rates were randomly selected from $[10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 5^{-2}, 8^{-2}, 10^{-1}]$. The number of hidden neurons were selected from a wide range of integers which includes $[10, 20, 40, 60, 80, 100, 140, 180, 200, 300, 600]$. The batch size was set to be 2^n where n is an integer number in [5, 9]. The dropout probability was a random one decimal float number $(\%.1f)$ ranging from 0.3 to 0.7 and the number of iterations were selected from $[10^4, 10^5, 10^6, 2 \times 10^6, 5 \times 10^6, 10 \times 10^6, 15 \times 10^6]$.

Random forests served as a baseline. Parameters of the random forests model were tuned using 5-fold cross validation. To search the hypothesis space, a random selection of parameters in a predefined pool of possible parameters was utilized. Number of estimators was a random integer number $n \times 100$ where n is an even number in $[2, 20]$. The same unseen test data as the one used in LSTM model was used to evaluate the performance of the random forest model. It should be noted that random forest models are trained on the static data and LSTM models are trained on the longitudinal temporal data.

Performance of the LSTM and the random forest models are in Table 3.10. All the results reported are on the test set. In Table 3.10, RF and LSTM respectively indicate a random forest model and a LSTM model trained and tested on the original data. RF over-sampled and LSTM over-sampled are the same models but trained using an over-sampled training data and tested on the original test data; the population of MCI patients was increased using a sampling with replacement approach.

In Table 3.10, the LSTM over-sampled produced both the highest F1-Score (0.46) and ROC AUC (0.75). Both LSTM and LSTM over-sampled performed better than the baselines in terms of F1 score. Both RF models produced higher accuracy than LSTM models but their recall were much lower than LSTM models, showing that random forest was biased due to the imbalanced class distribution (i.e., higher numbers of CU than MCI). These results indicate that LSTM models with temporal

Table 3.10: Performance of MCI prediction using LSTM and random forest.

Model	Acc.	Prec.	Rec.	F1-score	AUC
RF	0.82	0.44	0.13	0.21	0.73
RF Over-sampled	0.79	0.33	0.25	0.28	0.69
LSTM	0.73	0.33	0.59	0.43	0.71
LSTM Over-sampled	0.71	0.33	0.76	0.46	0.75

Table 3.11: Top 5 MCI risk factors detected by random forest.

Risk Factor	Score
Age	0.16
Hypertension	0.08
Education	0.07
Depression	0.06
Anxiety	0.05

patterns of the data might have increased capability in predicting MCI compared to a traditional machine learning models using the static data.

In this study, the patients who were determined as MCI in their first visit to the MCSA were excluded. When including those MCI patients at the first visit (MCI=1,075), LSTM over-sampled produced higher performance than using the original data; i.e., precision, recall, F1-score and AUC were 0.53, 0.73, 0.61 and 0.74, respectively. This increased performance might be because there are more signals of patients who already had MCI before the first visit and/or the more data help create the more efficient deep learning models.

The random forests was also used to investigate important risk factors for MCI. Random forests have a capability to identify the important variables based on impurity using information gain; i.e., how much each variable contributes to decreasing the impurity. Table 3.11 shows the top 5 variables based on their effect in decreasing impurity. As can be seen from Table 3.11, age is the most important feature in discriminating between MCI and CU patients. Age, hypertension, education, depression and anxiety have also been reported as important risk factors in developing MCI in other literature[85, 99–101].

3.2.6 Clustering and Visualizing MCI Patients

A denoising autoencoder with four layers was used to efficiently represent the patient data addressing data sparsity and high dimensionality. Reduced dimensionality enables the patient data to map a more meaningful space for better clustering. The outputs of the first hidden layer were mapped to a $2D$ space for visualization and clustering using tSNE and K-means. Figure 3.10 visualizes the patient data. CU patients are indicated by blue dots and MCI patients are represented by red dots.

Figure 3.11 shows the distribution of MCI versus CU patients. Notably, ratio of MCI patients is much larger than the ratio of CU patients in cluster 2 but it is opposite in cluster 4. This may indicate the potential of a deep learning model for clustering distinct patient groups. Figure 3.12a shows the distribution of males versus females and 3.12b further zoomed into males versus females for both CU and MCI patients. As can be seen, cluster 0 mostly includes males; almost 50 percent of males are clustered in this group. On the other hand, cluster 1 mostly includes females; half of the female population is clustered within this group.

3.2.7 Deep Learning Prediction of MCI: Discussion

Despite the urgent need for early detection of MCI and dementia, there is a lack of automated tool available to healthcare providers using routine EHR data. Given that MCI is a precursor of dementia and could be a critical step in the prevention and control of AD and other dementias, it is crucial to find a more efficient way to determine MCI in its very early stages. This study addresses this issue by using routinely-collected EHR data as part of patient care for predicting onset of MCI.

The LSTM RNN demonstrated its potential for MCI prediction incorporating temporal patterns of patient data that were automatically extracted from EHRs. Although this study used limited set of MCI risk factors compared with the previous study that used manually annotated variables, our models still produced comparable (even slightly higher) performance. The LSTM RNN using longitudinal temporal data seems to be more efficient in predicting MCI compared to using the traditional machine learning models with static data. Machine learning techniques such as denoising autoencoders, K-means, and tSNE to visualize and cluster the patients EHR data also showed a good potential to identify certain patient groups. The predictive model and patient clustering could (have the potential to) assist in the clinic to support early identification of MCI patients as well as better characterization, determining more granular subgroups of MCI patients. This can enable tailored care plan

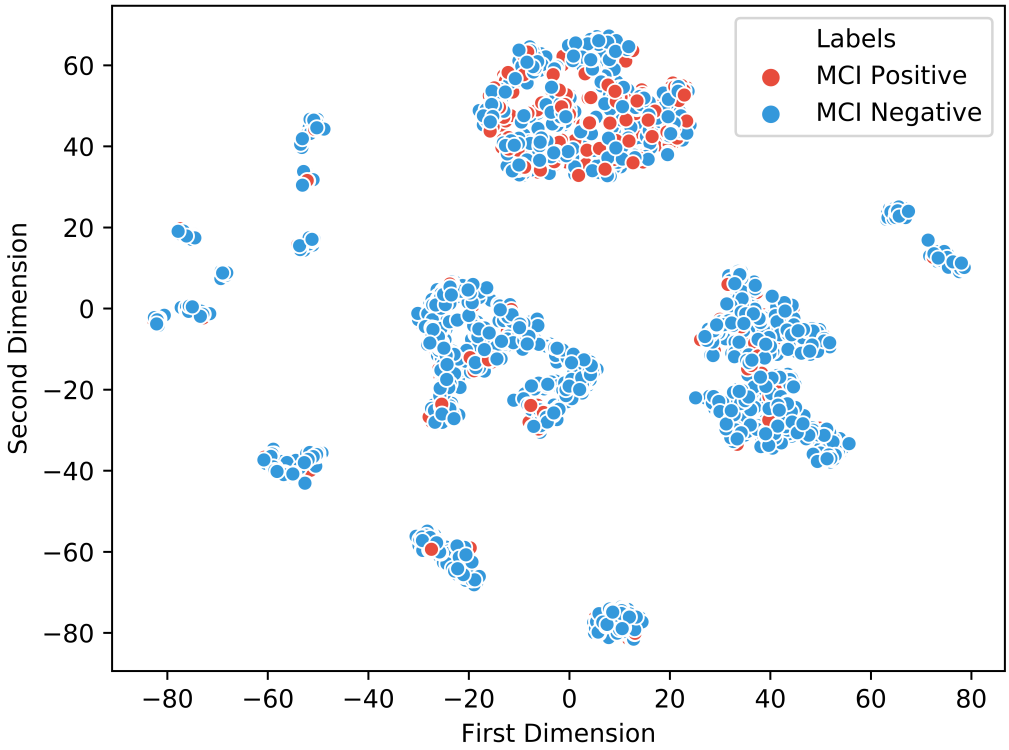


Figure 3.10: Visualization of CU patients (blue dots) and patients with MCI (red dots).

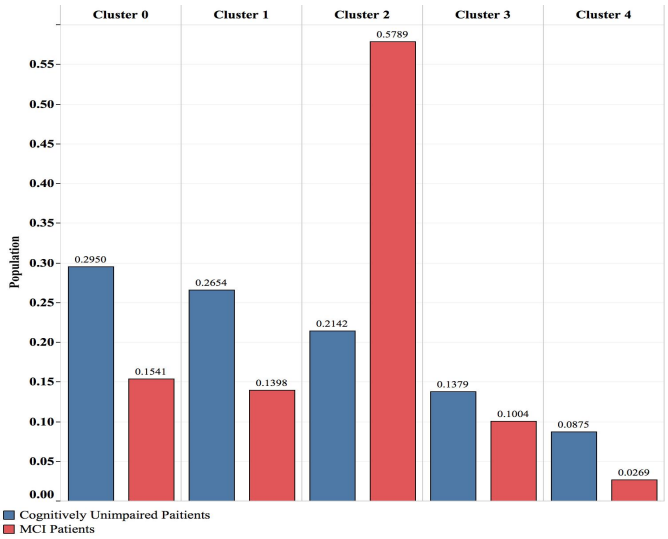
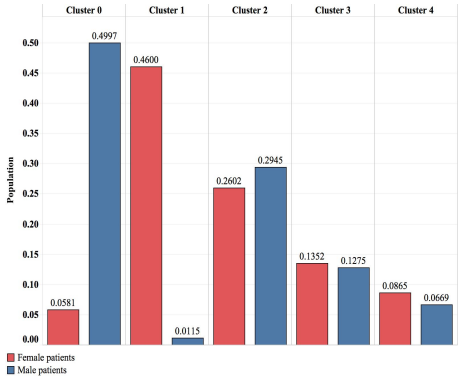
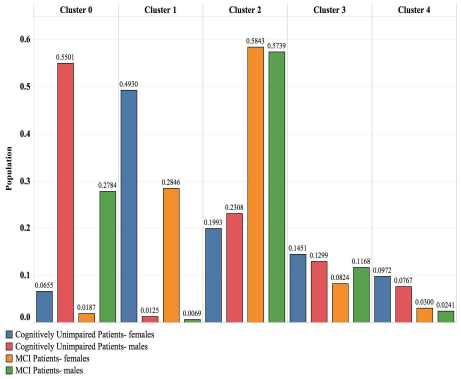


Figure 3.11: Distribution of MCI versus CU patients within each cluster. Y-axis shows the ratio of CU or MCI population within the relevant cluster.



(a) male and female within each cluster.



(b) CU and MCI patients with their sex.

Figure 3.12: Distribution of male and female patients within each cluster.

and open up new clinical practice opportunities using routine EHR data.

The limitation of this study includes the use of only Mayo clinical notes even though participants enrolled in the MCSA visit other healthcare institutions, and thus the patients might not be ideally represented by their comprehensive longitudinal data. Use of EHR data from other institutions requires significant effort.

3.3 Conclusion

This chapter, described two studies: SUD prediction using LSTM and MCI prediction using LSTM. Both studies use the same deep learning framework. This proposed model shows that LSTM models can capture the temporal patterns in big longitudinal data more precisely compared to traditional machine learning models. In the first study, the long-term impact of ADHD medication initiated during adolescence was systematically studied using the LSTM model. Long-term temporal medication application patterns appeared to be key factors that provide increased power to predict the development of subsequent SUD in adolescent ADHD enrollees using deep learning models. In the second study, LSTM RNN demonstrated a good potential incorporating temporal EHR patterns to predict the conversion from CU to MCI. When this model is combined with natural language processing to automatically extract MCI risk factors from EHR data, it could facilitate early detection of MCI addressing a current significant delay and thus improve treatment plans and health outcomes for patients.

Chapter 4 Multi-stream Transformer

In this chapter, limitations of the recent models to analyze sequential data is provided. Then, a novel model is proposed to address these limitations. The proposed model is designed to predict the onset of OUD using claims data. The motivations behind proposing this new model and different components of the proposed model are discussed to show how the proposed model can address previous limitations.

4.1 Introduction

Deep learning models are end to end algorithms that are constituted of nonlinear components that each transform the representation at one level into a representation at an abstract level[1,102]. They already demonstrated great performance and potential in analyzing sequential and longitudinal data[8,16,43,47]. Recurrent Neural Networks were the variant of deep learning models that showed promising performance in natural language processing and has been heavily used in sequence labeling and sequence to sequence modeling tasks. The sequential nature of RNNs and their ability to deal with vanishing gradient problem and to memorize long dependencies in input sequences, enables RNNs to tackle many issues in sequential and longitudinal data analysis.

However, RNNs suffer from some limitations that precludes them from being trained in a parallelized fashion and therefore performing well on more complex tasks and inputs with longer sequences. Attention mechanism and later the transformer models proposed to deal with the above mentioned problems of RNNs (refer to chapter 2). Given Transformers' demonstrated performance on various sequence modeling tasks, transformer paradigms introduce exciting new opportunities for processing sequential data. In this study, the technical and theoretical challenges in analyzing multi stream longitudinal data using transformer models are addressed and a new framework to analyze multi stream temporal big data is proposed.

There are several aspects of transformers that could be useful in analyzing multi stream longitudinal data, such as their high performance in capturing long dependencies in input sequences, end-to-end learning scheme with integrated feature learning, capability of parallelization within training examples. However, transformers are originally designed for NLP tasks and the input to these models is typically a sequence of words. This causes some limitations in evaluating and using transformers

in analyzing longitudinal data collected from multiple resources such as healthcare data.

First, in spite of NLP data which are commonly sequential data the longitudinal data is temporal. In fact, not only the order of events is important but also the time gaps between the events plays a crucial role while processing data. For example, earlier in this study is has been shown that temporal medication features, rather than stationary features, are the most important factors for predicting SUD in the IBM MarketScan data set. Therefore, instead of positional encoding of the order of inputs at various time steps, a novel encoding should be encoded to capture the temporal patterns of the longitudinal data and events rather than just the sequential order. Second, the transformer is designed to attend to different words in a sentence and find the relations between the words. In fact, input to the transformer is a single stream of words while in many applications such as healthcare data analysis there are multiple sources of inputs such as medications data, diagnoses data, and procedures data. A more sophisticated model is needed to find the associations between different input streams to capture hidden associations within and between the input streams. Third, the attention mechanism in transformers can be utilized wisely to make the deep learning models more transparent in applications such as biomedical applications where finding the association between events (medications, diagnoses and procedures in different visits) plays an important role and can make the process of decision making more transparent.

All in all, this study proposes a new transformer based deep learning model to tackle the above mentioned issues in processing longitudinal data with multiple input streams. Next section describes the proposed model and explains how it can tackle the above mentioned problems.

4.2 Opioid Use Disorder

Early identification and engagement of individuals at risk of developing an opioid use disorder (OUD) is a critical unmet need in healthcare[23,24]. Individuals with OUD often do not seek treatment or have internalized stigma about OUD that limits identification through traditional means, such as screening and clinical interview[25]. Significant disparities limit access to treatment for OUD resulting in less than 20% of all individuals with OUD receiving any form of treatment in the past year[26]. While there are currently tools developed to predict aberrant behavior when prescribing opioids[28] or to predict OUD from a general primary care population[29], there are only

a few clinical tools, such as the Opioid Risk Tool[30], developed for assessing the risk of OUD. Typical clinician workflow does not allow for comprehensive OUD screening, but available administrative and clinical data have the potential to help clinicians identify and screen higher risk patients providing an opportunity for primary care professionals to play a greater role in increasing OUD detection, treatment, and prevention.

Healthcare data are a growing source of information that can be harnessed together with machine learning to advance our understanding of factors that increase the propensity for developing OUDs as well as those that aid in the treatment of the disorders[31,32]. In healthcare data, patients' outcomes and treatments are collected at multiple follow-up times. Tools developed to analyze longitudinal healthcare data and to extract meaningful patterns from these ever growing data are critical in addressing real-world public health emergency including but not limited to OUD.

Analyzing real-world data is a complicated task with multiple computational challenges including high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity[1]. In particular, healthcare (and claim) data are typically collected from multiple sources, and the subsequent data analysis requires simultaneous analysis of the temporal correlation among multiple streams such as medications, diagnoses, and procedures. Deep learning models have demonstrated great potential in addressing some of these challenges and creating promising longitudinal healthcare data analysis tools. Among them, Doctor AI[8], RETAIN[9], and DeepCare[10] modeled multiple data streams including medications, diagnoses, and procedures using Recurrent Neural Network (RNN) models such as Long-Short Term Memory models (LSTMs)[2]. Doctor AI concatenated multi-hot input vectors to predict subsequent visit events[8]. RETAIN used two separated RNNs to generate attentions at the visit level and the variable level as well[9]. These applications demonstrate that RNNs are promising in longitudinal and sequential healthcare data analysis, since RNNs are capable of extracting contextual information from past time steps and pass this information forward; this helps to efficiently model long-term dependencies in input streams[3]. Nevertheless, the network architecture and design preclude RNNs from processing long streams in a reasonable amount of time[4]. Attention mechanism was introduced in RNNs to increase their capacity in capturing long range dependencies more efficiently[4–6]. Attention-based models bridge the gap between different states in RNNs using a context vector. Successful applications of multiple attention layers led to the transformer model[7], which removed recurrence in RNNs relying entirely on the attention mechanism.

The transformer is a type of attention-based deep learning models originally proposed for natural language processing (NLP) tasks such as machine translation[7]. Later, transformers have been applied on longitudinal EHR data[11] to predict patients' outcomes in the future. There are already several models that have been successfully applied on EHR data without significantly changing the network architecture or loss[12–14]. Of course, the typical transformer's structure can be altered to better fit the special needs of solving healthcare problems[11,15]. Choi et al. proposed a transformer model for healthcare data analysis by utilizing the conditional probabilities calculated from the encounter records to guide the self-attention mechanism in the transformer[11]. BEHRT[15] was developed based on BERT[16], a popular transformer model for NLP tasks, for analyzing EHR data. BEHRT considers the patients' existing diagnoses and demographic data to predict their future diagnoses. Similar to RNNs, transformers have been modified to model multiple data streams. Li et al developed a two-stream transformer to analyze both time-over-channel and channel-over-time features in human activity recognition tasks[17]. Two parallel, yet separate transformers were used to handle two input streams. Another multi-stream transformer has been developed to generate effective self-attentions for speech recognition[18]. They parallelized multiple self-attention encoders to process different input speech frames. Gomez et al. developed a multi-channel transformer for sign language translation using one self-attention encoder[19]. Their model finds the attentions across three different channels, i.e. hand shapes, mouthing, and upper body pose. A more recent work[20] showed that "transformer is all you need" by using multiple transformer encoders. The encoded outputs can be concatenated using a joint decoder that enables simultaneous model training. There are also works that analyze multi-stream data using transformer by simply stacking or parallelizing multiple transformer models[21,22].

Although the recently developed transformer models showed promising performance, especially on handling multiple data streams, the potential of applying transformers on healthcare data analysis has not yet been fully explored. One of the major limitations is the lack of capacity to model multiple data streams within the self-attention layer. The transformer was originally designed to process one data stream, which is mostly an order of words in a NLP task, at a time. The modified transformers either can only handle multiple streams at intra-stream level or they are not suitable to solve OUD identification problem as a real-time task where only previous clinical events can be used to make a decision at a specific time point. Here, OUD identification is a complex data analysis task that includes not only finding long term

effects of prescription opioids such as morphine and fentanyl, history of diagnoses such as mood disorders, but also the hidden associations between patient's prescriptions and diagnoses, since these input streams are highly correlated with each other. Identifying the relationships within and between input streams may reveal hidden patterns leading to an increased classification ability and interpretabilty for OUDs. Moreover, the medication application patterns and the interactions between medications across different visits as well as the patient's diagnoses patterns thorough his/her medical history may carry important information that should be extracted in order to develop precise and sensitive OUD identification tools.

This study proposes a novel transformer model called **Mu**lti-stream Transformer for **P**redicting **O**pioid Use **D**isorder (MUPOD) to analyze longitudinal healthcare data collected from multiple sources and predict the onset of OUD. First of all, MUPOD is capable of analyzing multiple data streams, such as medication and diagnosis, simultaneously and extracting associations within and between the streams. Second, MUPOD utilizes attention weights within and across data streams to interpret the classification results. In the experimenst, MUPOD successfully captured the complex associations within and between multiple streams including medications, diagnoses, and demographic information, and predicts the onset of OUD precisely.

4.2.1 Data Set

The large-scale administrative records in the IBM (formerly Truven Health Analytics) MarketScan Commercial Claims[103] database were used to train and test both baseline models and MUPOD. Data include person-specific clinical utilization, expenditures and enrollment across inpatient, outpatient, prescription drug and carve-out services. The database contains about 30 million enrollees, a nationally representative sample of the US population with respect to sex (50% female), regional distribution, and age.

Medications, diagnoses and demographic information of 682,402 patients who have at least one diagnosis of OUD (ICD-9: 304.0x, 305.5x and ICD-10: F11.xxx; where x can be any code) from 2009 to 2018 were extracted. The hyper-geometric[104] test was used to identify sub-cohorts of OUD with high statistical significance of whether a population consists the richest information of OUD. An OUD sub-cohort (p-value 0.00) with 229,214 patients who had at least one Clinical Classification Software code (CCS) of 205 (patients with Spondylosis; intervertebral disc disorders; other back problems) was identified. This sub-cohort was defined as the case cohort. Note that CCS 205 has already been shown to be a prevalent diagnosis in OUD patients in the

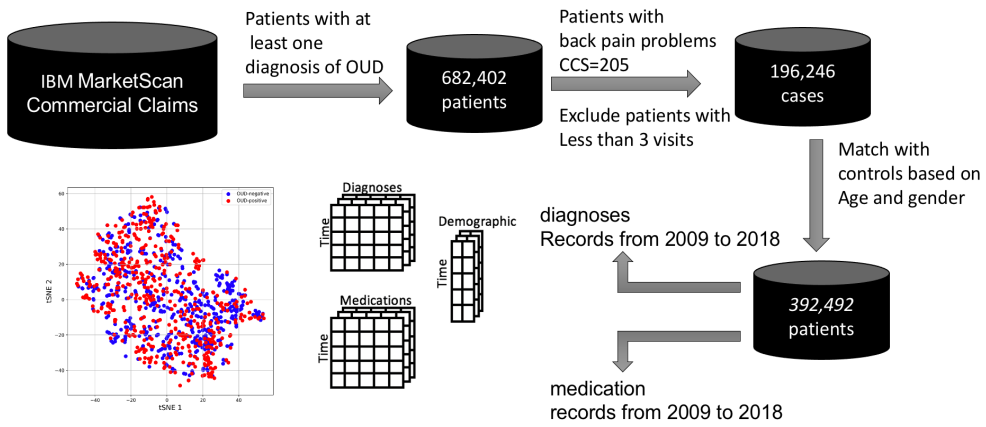


Figure 4.1: Data preprocesing schema for OUD prediction.

literature[105,106].

The case cohort (OUD positive and CCS 205) was matched with a subpopulation of OUD-negative patients called the control cohort. All the individuals in the control cohort have the same back pain diagnosis (CCS 205) but do not develop OUD. Cases and controls were first matched based on age and sex. The age of cases were divided into 10-year bins and selected controls so that their age distribution matches the age and sex distributions of cases. Second, they were matched based on the opioid medication use duration. Specifically, every opioid medication were grouped with a therapeutic class generic product identifier (TCGPI) of 65x as opioid medications. Buprenorphine and Methadone were excluded as they are often used as a treatment for opioid overdose. Next, OUD-negative patients who have the matched age and gender with the case were randomly sampled ensuring that the averaged opioid use ratio between case and control is almost equal. Figure 4.1 describes these preprocessing steps.

Note, it is important that in OUD prediction applications the case and controls are matched based on opioid medication use in addition to age and sex. Data visualization was used to show the importance of matching cases and controls based on opioid medication use. 1000 samples were randomly selected from each class and reduced the dimensionality to $2D$ for visualization purpose using tSNE. Figure 4.2 shows the distribution of OUD-positive (red dots) and OUD-negative (blue dots) in the data. In Figure 4.2 cases and controls in the data are finely overlapped and distributed in the

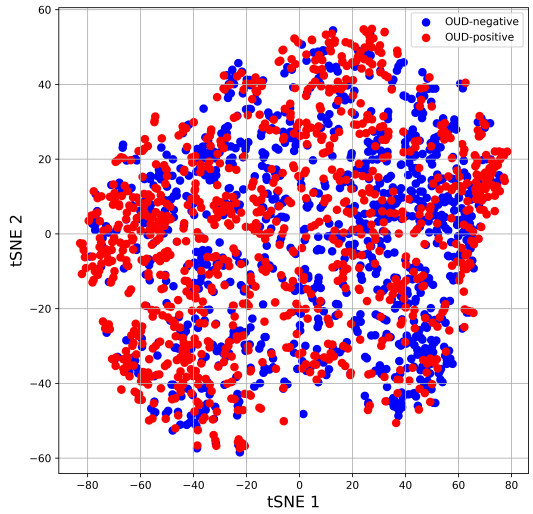


Figure 4.2: Distribution of OUD-positive and OUD-negative samples in the data. Cases and controls are matched based on age, sex and opioid medication use.

feature space. In fact, the samples are a good representation for the discriminative boundary between OUD-positive and OUD-negative patients. Therefore, the dataset is a valid data to challenge our proposed model and baselines.

Table 4.1 shows the characteristics of cases and controls regarding age, sex, top-10 most frequent medications and top-10 most frequent diagnoses. For age, average age were provided in each cohort as well as the standard deviation of the age within each cohort. The numbers within parenthesis are the standard deviations. For sex, the number of female patients is provided. The numbers within parenthesis in this case are the percentage of female patients in each cohort.

The diagnoses and the medications were classified using CCS codes and Generic Product Identifier codes (TCGPI) respectively. Opioid analgesics, anticonvulsants (neuromuscular agents), musculoskeletal therapy agents, and antianxiety agents were grouped based on the first two digits of their TCGPI codes as $65x$, $72x$, $75x$, and $57x$, respectively. The rest of medications were classified using the first 6 digits of their TCGPI codes (from left to right). The variables presented in Table 4.1 have already been reported as OUD risk factors in the literature[105,107]. Especially, diseases including "Other connective tissue disease", "Other nervous system disorders", "Essential hypertension", "Mood disorders", "Other non-traumatic joint disorders and Anxiety disorders" have been found to be more prevalent diagnoses among OUD patients than normal people[105]. Note, since the case and control cohorts were matched

based on age, sex and analgesics-opioid use, these three variables have similar statistical characteristics across both case and control cohorts. However, the distributions of other variables vary across the case and control cohorts and can be utilized by our deep learning models to discriminate OUD-positive patients from OUD-negative individuals.

4.2.2 Data Pre-processing

For each of the enrollees in the case and control cohort, his/her medications and diagnoses between Jan 2009 and Dec 2018 and demographic records were extracted. In total, 78,136,935 medication records and 143,275,864 diagnoses records were extracted. The original format of the prescription and professional service encounter claims in IBM MarketScan data is a table where each row is a visit and columns are enrollee ID, date of visit, and prescription/diagnoses. If an enrollee has multiple visits, each visit will occupy a row in the table.

To facilitate further study of the temporal patterns in the data, the data were converted into an enrollee-time matrix $X(P, T, F)$ where each $x_{i,j} \subseteq F$ is a set of medications or diagnoses (from feature space F) associated with enrollee $p_i \in P$ at time slot $t_j \in T$, where P is the enrollee set and T is the set of monthly slots between Jan 2009 and Dec 2018. Patients were excluded from $X(P, T, F)$ if the number of valid entries is less than 3. As a result, the final dataset includes 392,492 patients including equal number of OUD-positive and OUD-negative samples.

4.2.3 Patients Data Representation

Applying a temporal model on healthcare data is nontrivial due to technical challenges inherent in the data including high dimensionality, heterogeneity, temporal dependency, sparsity, and irregularity[1]. Therefore, patients data representation is critical to ensure high model performance.

The goal of data representation is to learn a function: $f_R : X \rightarrow \mathbb{R}^d$, where d is 10 in this work and it shows the dimension of the representation to which each input stream is mapped, $X \in \{M, D\}$, and M and D are medication and diagnosis, respectively. To train the function f_R, LSTM[43,45] was adopted. The outputs from all LSTM hidden states were used to represent both the OUD case and control cohorts. The general schema of the data pre-processing and representation is shown in Figure 4.3.

Table 4.1: Distributions of age, sex, medication, and diagnoses in case and control patients. Top 10 diagnoses and medications are provided. The numbers indicate the number of patients who had at least one such diagnosis or medication.

Variables	**Case**	**Control**	**Variables**	**Case**	**Control**
Demographics					
Age (SD)	45.62 (13.81)	52.35 (14.39)	Female (percentage)	109,121 (55.60%)	117,699 (59.98%)
Diagnoses (CCS Code)			**Medications (TCGPI Code)**		
Other connective tissue disease (211)	152,703 (77.81%)	165,112 (84.14%)	Analgesics - Opioid (65)	190,141 (96.89%)	196,246 (100%)
Other nervous system disorders (95)	138,866 (70.76%)	141,350 (72.03%)	Neuromuscular Agents Anticonvulsants (72)	105,508 (53.76%)	97,444 (49.65%)
Essential hypertension (98)	106,299 (54.17%)	132,049 (67.29)	Musculoskeletal Therapy Agents (75)	106,186 (54.11%)	102,888 (52.43%)
Mood disorders (657)	97,035 (49.45%)	81,306 (41.43%)	Antianxiety Agents (57)	76,830 (39.15%)	75,463 (38.45%)
Other aftercare (257)	127,131 (64.78%)	133,920 (68.24%)	Proton Pump Inhibitors (492700)	71,243 (36.30%)	86,561 (44.11%)
Residual codes; unclassified (259)	136,177 (69.39%)	152,748 (77.83%)	Serotonin-norepinephrine Reuptake Inhibitors (581800)	58,039 (29.57%)	48,323 (24.62%)
Other non-traumatic joint disorders (204)	134,042 (68.30%)	150,660 (76.77%)	Selective Serotonin Reuptake Inhibitors (581600)	69,665 (35.50%)	65,005 (33.12%)
Anxiety disorders (651)	91,736 (46.75%)	78,296 (39.90%)	Hmg Coa Reductase Inhibitors (394000)	53,806 (27.42%)	79,201 (40.36)
Disorders of lipid metabolism (53)	94,507 (48.16%)	122,322 (62.33%)	Non-barbiturate Hypnotics (602040)	46,965 (23.93%)	44,404 (22.63%)
Medical examination/evaluation (256)	129,224 (65.85%)	147,268 (75.04%)	Nonsteroidal Anti-inflammatory Agents (661000)	87,301 (44.49%)	98,639 (50.26%)

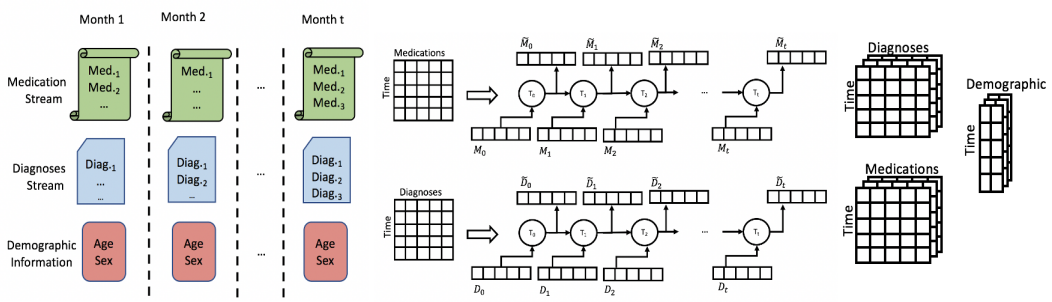


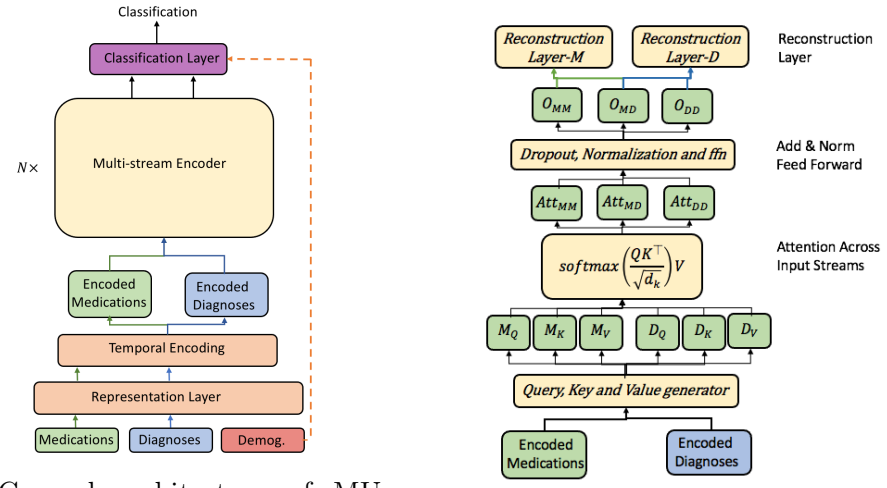
Figure 4.3: Data preprocessing and patient representation. EHR data are first converted to an enrollee-time matrix $X(P, T, F)$. Then, the data are fed to LSTM models to encode the medication and diagnosis streams separately.

4.3 MUPOD: Two-stream Transformer

MUPOD is a transformer-based deep learning model designed to analyze n highly correlated healthcare data streams simultaneously. To minimize ambiguity, the algorithm is described for a single patient and for $n = 3$. Each patient can be represented by $p = (S, y)$ in which S is a set of input streams and y is the target label. Herein, three input streams are considered: 1) medication tuples (T, M) in which t_i is the i^{th} time step and M is a list of medications that the patient is prescribed with at time t_i, 2), diagnoses tuples (T, D) where t_i is the i^{th} time step and D is a list of diagnoses assigned to the patient P at time t_i, 3) demographic tuples (T, G) in which t_i is the i^{th} time step and G is the demographic information of patient P at t_i.

This study uses the encoder part of transformer to identify the associations between medication and diagnosis across time and detect the onset of OUD. Medications M, diagnoses D, and demographics G are fed to the model in parallel. The first step is to incorporate the temporal patterns of the data stream into the encoder's inputs using positional encoding. The embedding layer in the transformer is replaced by the proposed LSTM based representation layer. This change has two computational advantages. Firstly, it deals with challenges in the input data such as variable dimension and data sparseness, which is common in longitudinal healthcare data. Secondly, it extracts hidden parameters and transforms the original input into a new feature space where cases and controls are better separated than in the original feature space.

The encoded input streams are plugged into the attention layer to generate Query, Key, and Value matrices for each input stream. For example, medications M are fed to a set of fully connected layers to generate M_Q, M_K, and M_V, representing the query, key, and value matrices for the encoded medication stream for patient P. Let $X, Y \in \{M, D\}$, the Query, Key, and Value matrices are used to find the attentions



(a) General architecture of MUPOD.

(b) Multi-stream encoder.

Figure 4.4: MUPOD architecture. $X_Q, X_K, and \ \ X_V$ represent query, key, and value matrices for the input stream X, where $X \in \{Medication, Diagnoses\}$. Att_{XY} represents the attention weights between different records across input streams X and Y, where $X, Y \in \{Medications, Diagnoses\}$. O_{XY} represents the outputs, which capture the associations between the input streams X and Y. The demographic information is plugged into the system before the last layer and in the classification layer.

across these three input streams:

$$Attention(X_Q, Y_K, Y_V) = softmax(\frac{X_Q Y_K^\top}{\sqrt{d_k}})Y_V \tag{4.1}$$

Note, the d_k is the same as the original transformer. Figure 4.4a describes how the data flows through the different layers of MUPOD. The raw medication and diagnose streams are first represented in the representation layer (the intermediate outputs of the LSTMS models in Figure 4.3). The temporal information is then encoded into the represented streams in the temporal encoding layer. The encoded streams are processed in the MUPOD's multi-stream encoder layer. This novel multi-attention layer is further described in more details in Figure 4.4b. In the figure, X_Q, X_K, and X_V represent query, key, and value matrices for stream X ($X \in \{M, D\}$). All possible combinations of the streams are used to determine the attention weights between different visits and across streams. Attentions are then passed through a set of dense layers to generate outputs. For example, given two data streams M and D, three combinations can be generated i.e. MM, MD, and DD.

The reconstruction layer receives the relevant outputs and maps them to appropriate format for the next layer as described in Equation 4.2. For example, only

the outputs relevant to the medications (M) including O_{MM} and O_{MD} are used to reconstruct the medication stream appropriate to be fed into the next encoder layer:

$$\begin{aligned} f : O_{XX}, O_{XY} &\longrightarrow \hat{X} \\ \hat{X} =& [Concat(O_{XX}, O_{XY})]W_x + b_x \end{aligned} \tag{4.2}$$

where $O_{XX}, O_{XY} \subset \{O_{MM}, O_{MD}, O_{DD}\}$, $\hat{X} \in \{\hat{M}, \hat{D}\}$, $X, Y \in \{M, D\}$, W_x and b_x are trainable reconstruction weight and bias matrices. The two reconstructed matrices generated by the last encoder layer are fed to classification layer to make the final decision for the current patient p as $Softmax([Concat(\hat{M}, \hat{D})]W + b)$.

4.3.1 Experimental Results

All the deep learning models in this work were deployed on the TensorFlow platform[65] and were trained using eight GeForce GTX 1080 GPUs. The original transformer model, LSTM models, Linear Regression (LR), Random Forest (RF)[67] and Support Vector Machine (SVM)[66] were compared with MUPOD as baselines. 314,504 samples were used for training, 38,776 samples for validation and 39,212 for testing the models. All results reported in this paper are on the test set. All models were optimized using a random search policy across hyper-parameters of each model. A grid of hyper-parameters values was set up and 10 random combinations of the hyper-parameters were selected to train the models.

The optimized SVM model uses a RBF kernel function and the optimum value for the parameter C is 0.0039 in this model. The optimized linear regression model uses the L2 norm with $C = 0.0625$ in penalization and the sag algorithm as it's solver method. The optimum number of trees in the random forest model is 1600 and the optimum value for maximum number of levels in a tree is 40. For the LSTMs, their learning rates were randomly set to 10^n where $n \in \{-2, -3, -4\}$. The batch size was randomly selected from $\{64, 256, 512\}$ and the number of iterations was randomly selected from $n \times 10^3$ where $n \in \{10, 50, 100, 200\}$. The regularization parameter for LSTM models was randomly selected from 10^n where $n \in \{-4, -5, -6\}$. The number of hidden neurons for the LSTMs in the representation layer was fixed to 10; because the outputs of these LSTM models were the inputs to MUPOD and the inputs to our model have to be of a fixed dimension (the dimension of our model in this paper is 20: 10 for medications and 10 for diagnoses stream). However, the number of hidden neurons for the other LSTM model used as a baseline (refer to Table 4.2) was randomly selected from 2^n where $n \in \{3, 4, 5, 6, 7, 8\}$.

Table 4.2 compares the classification performance of MUPOD with LR, RF, SVM, LSTM and the original Transformer model. The same train, validation and test data were used to train, validate and test all models in Table 4.2 except for the SVM model. Due to the hardware and time limitation this model was trained and tested using 10,000 randomly selected samples. Note, the LSTM model in Table 4.2 is trained using medication, diagnosis and demographic data. The vectors of medication, diagnosis and demographics were concatenated in each time step and formed a single vector which was fed to this LSTM model. The LSTM model were dynamically unrolled based on the input sequences' lengths and applied a fully connected layer and an argmax function on the last output of the unrolled LSTM model to make the final decisions. The hyper-parameter search space for this LSTM was the same as explained earlier in this section.

A randomized 5-fold cross validation was used to tune LR, RF and SVM models. The LR, RF and SVM were trained on the static data and the LSTM, transformer and MUPOD were trained on the longitudinal data. To create static data for LR, RF and SVM, the longitudinal data was converted to a new format $Y(P, L)$, where P is the complete list of patients, and L is a vector including aggregated values for all medication, diagnosis and demographic features across time steps (from Jan. 2009 to Dec. 2018). In fact, the frequencies for each medications and diagnoses were counted and concatenated with demographic information of the patients to create L.

Transformer is the original encoder block of the transformer model[7]. The vectors of medication, diagnosis and demographics were concatenated and fed to the original encoder block of the transformer model. Then, a fully connected layer and softmax function were used to perform the final classifications. In Table 4.2, MUPOD has the highest accuracy (0.775), precision (0.741), F1-score (0.790) and AUC (0.871). These results indicate that our proposed model captures important factors in the medication, diagnosis and demographic data and provides an increased power to detect the development of OUD, while LR, RF, SVM, LSTM and original Transformer appear to miss such factors.

In addition, the models' performances were tested on three imbalanced test data sets with the ratio of OUD-positive samples to OUD-negative samples set to 0.1, 0.2 and 0.5. OUD is an uncommon event and the ratio of OUD-positive to OUD-negative patients in patients who have used Opioid prescriptions at least 3 times is 3.2% in the data set. Therefore, the experiments in Table 4.3 were conducted to simulate the performance of the models on imbalanced datasets as well. Table 4.3 shows the model performances on imbalanced test sets. Table 4.3 shows that MUPOD maintains

Table 4.2: Performance of OUD classification using MUPOD compared to RF, SVM, LSTM and original transformer.

Model	**Acc.**	**Prec.**	**Rec.**	**F1-score**	**AUC**	**P@R=.8±0.001**
LR	0.638	0.641	0.625	0.633	0.689	0.463
RF	0.698	0.693	0.710	0.702	0.774	0.449
SVM	0.569	0.539	0.831	0.654	0.677	0.478
LSTM	0.693	0.784	0.533	0.635	0.790	0.666
Transformer	0.708	0.654	**0.880**	0.751	0.801	0.689
MUPOD	**0.775**	**0.741**	0.847	**0.790**	**0.871**	0.771

Table 4.3: OUD classification results for imbalanced test sets. The $.xN$ means the number of samples in the OUD-positive cohort are $0.x$ times smaller than the number of samples in the OUD-negative cohort.

Model	Precision			Recall			F1-score			AUC		
	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$
RF	.531	.313	.182	.710	.715	.701	.608	.436	.290	.773	.777	.770
LSTM	.539	.312	.189	.548	.532	.546	.544	.393	.281	.730	.723	.732
Transformer	.486	.276	.160	**.879**	**.885**	**.883**	.626	.420	.270	.799	.804	.796
MUPOD	**.588**	**.364**	**.221**	.845	.848	.843	**.693**	**.509**	**.351**	**.871**	**.870**	**.871**

higher performance on all imbalanced test sets compared to all baselines in terms of precision, F1-score and AUC. Note, the accuracy is not presented in Table 4.3, because this measure is not informative when assessing algorithms on imbalanced data.

The relationships between the medication and diagnosis streams were examined by aggregating the attention weights in the first layer of the model for all the records of each individual and visualized the results. While it is still unclear whether attentions can be used to explain deep learning models[108,109], attention weights have been used extensively to assess feature importance[15,110]. In particular, the aggregated attentions across all the records of the same patient may be useful to identify important relationships between his/her prescriptions and diagnoses. In the visualization, a rectangular node represents a medication type and an oval node represents a diagnosis code. The accumulated attention weights were divided to "moderate" and "strong" based on pre-defined thresholds (i.e. moderate: $0.3 \sim 0.6$, and strong: $\geqslant 0.6$) that were selected by visually inspecting the distribution of accumulated attention weights. The moderate and strong connections are represented using dashed

and solid lines respectively. The lines of an OUD-negative patient are colored black, while the lines of an OUD-positive patient are colored red.

Figure 4.5b shows the attention weights computed with MUPOD on one OUD-positive and one OUD-negative patient. The cosine similarities of the medication and diagnosis streams of the two patients are 0.85 and 0.27, respectively, indicating that they have different diagnoses but similar medication records. The connections belonging to the positive and negative patients are well separated. Besides, almost all the strong connections are from the OUD-positive patient, while all the moderate connections are from the OUD-negative patient.

Similarly, Figure 4.5c shows the attention weights on one OUD-positive and one OUD-negative patient. The cosine similarities of the medication and diagnosis streams of the two patients are 0.71 and 0.93, respectively, indicating that they have very similar diagnoses and medication records. Although they have similar records and similar connections between medication and diagnoses nodes, the strengths of attention are different for the OUD-positive patient versus the OUD-negative patient and MUPOD was able to correctly classify these two samples. Note that ONTJD and Opioid are collected with both the OUD-positive link (red) and the OUD-negative link (black), indicating the ONTJD-Opioid is often observed on both cases. Figure 4.5 shows that the attention weights in MUPOD can be used to: 1) discriminate OUD-positive from OUD-negative patients and 2) reveal the relationships between medications that the patient has been prescribed with and the diagnoses he/she has been diagnosed with. These attention weights can further be accumulated across all patients in the cohort to create more generalized conclusions and OUD risk factor identification.

4.4 MUPOD: Multi-stream Transformer

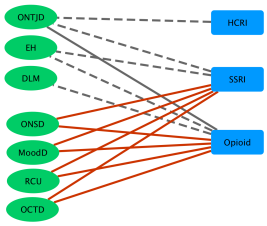
In this section, a multi-stream transformer is proposed and developed to predict OUD onset among patients in IBM MarketScan data. This proposed multi-stream transformer extends the two-stream transformer described in the previous section in multiple ways:

- The multi-stream transformer is capable of analyzing more than two streams of data. In fact, it is capable of simultaneously processing demographics, medications, diagnoses and procedures.

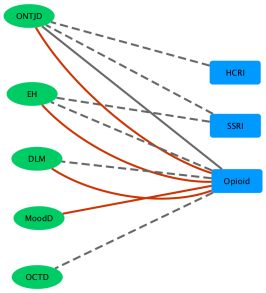
- The multi-stream transformer is trained using a much larger cohort. This model

Node	Name	Stream
AnxA	Antianxiety Agents	Medication
AnxD	Anxiety disorders	Diagnoses
DLM	Disorders of lipid metabolism	Diagnoses
EH	Essential hypertension	Diagnoses
HCRI	Hmg Coa Reductase Inhibitors	Medication
MTA	Musculoskeletal Therapy Agents	Medication
ME	Medical examination/evaluation	Diagnoses
MoodD	Mood disorders	Diagnoses
NH	Non-barbiturate Hypnotics	Medication
NAIA	Nonsteroidal Anti-inflammatory Agents	Medication
NAA	Neuromuscular Agents Anticonvulsants	Medication
OA	Other aftercare	Diagnoses
OCTD	Other connective tissue disease	Diagnoses
ONSD	Other nervous system disorders	Diagnoses
Opioid	Analgesics - Opioid	Medication
ONTJD	Other non-traumatic joint disorders	Diagnoses
PPI	Proton Pump Inhibitor	Medication
RCU	Residual codes; unclassified	Diagnoses
SSRI	Selective Serotonin Reuptake Inhibitors	Medication
SNRI	Serotonin-norepinephrine Reuptake Inhibitors	Medication

(a) Medication and diagnoses abbreviations and full names.



(b) A pair of OUD-positive and OUD-negative samples that have different diagnoses but similar medication records.



(c) A pair of OUD-positive and OUD-negative samples that have very similar diagnoses and medication records.

Figure 4.5: Attention weights. Rectangular nodes represent medications and oval nodes represent diagnoses. Solid, dashed and dotted edges respectively mean strong, moderate and weak connections. Abbreviations were used for medications and diagnoses, and the full names are provided in (a).

is trained and tested using all patients in IBM MarketScan data who have been prescribed with a certain amount of opioid.

- Further, this model uses all available medications, diagnoses and procedures instead of the top-20 features which was used to train two-stream transformer. In fact, the model is trained with 50 high level medications, 138 CCS diagnoses codes, 80 procedure CCS codes and 2 demographics variables. These features are selected after performing a sparse feature filtering originally performed on more than 600 features.

- The prediction window size increased from 1 month to 6 months.

- An open source tool is released to efficiently process big claim data sets and is available for other researchers to use the multi-stream transformer.

4.4.1 Cohort Selection and pre-processing

The cohort selection for multi-stream transformer is more comprehensive compared to the cohort selected to train and test the two-stream transformer. Here, the patients who have been prescribed with Buprenorphine or Methadone are also considered as cases as the prescriptions of these two medications is a strong indication of OUD. The cohort selection to choose case samples in this study can be described as:

- At least one OUD diagnoses or one prescription for Buprenorphine or Methadone.

- At least 3 opioid (other than Buprenorphine and Methadone) prescriptions 180 days prior to OUD diagnoses date.
- At least 12 month of data availability.

The cohort selection for controls can be described as:

- Never been diagnose with OUD or been prescribed with Buprenorphine or Methadone.
- At least 3 opioid (other than Buprenorphine and Methadone) prescriptions 180 days prior to their last record in the data.
- At least 12 month of data availability.

This process resulted in 7,910,707 OUD-negative and 257,084 OUD-positive patients. These cases and controls were matched based on multiple criteria: 1) age, 2) gender, 3) the opioid use duration, and 4) the data availability. However, due to the large volume of the data (more than 1M patients data over 10 years and with more than 600 variables at each time step) the matching can not be (efficiently) directly applied. Therefore, an anchor-based method was used to perform the matching. The OUD-positive cohort was first divided into tow sub-cohorts based on the patients' genders. Each sub-cohort is then clustered using K-means algorithm. Elbow method was used to define the number of cluster for each sub-cohort. Figures 4.6 and 4.7 shows the results of the elbow method applied to find the optimum number of clusters on male patients and female patients, respectively.

Then, cluster centers as well as one percent of the data points in each cluster were used as anchors and cases and control samples were matched based on their distance to these anchors. Figure 4.8 shows the distribution of a sample of cases and controls after matching was performed. Red dots represent OUD-positive cases and blue dots represent OUD-negative controls. The cases and controls are well matched as their distance is close to each other in the $2D$ space created by the tSNE algorithm.

Here, the medications are grouped using the first two digits of the TCGPI codes, the diagnoses variables are grouped using CCS code, and the procedures variables are grouped using procedure CCS codes. As a result, each record includes 94 medications, 284 diagnoses and 243 procedures. Including the demographic features (age and sex) the total number of features in the data is 623 features. Figures 4.9, 4.10 and 4.11 respectively describe the frequencies for medication, diagnosis and procedure features in the cohort. In these figures, the x axis shows the features like f_i and the y axis

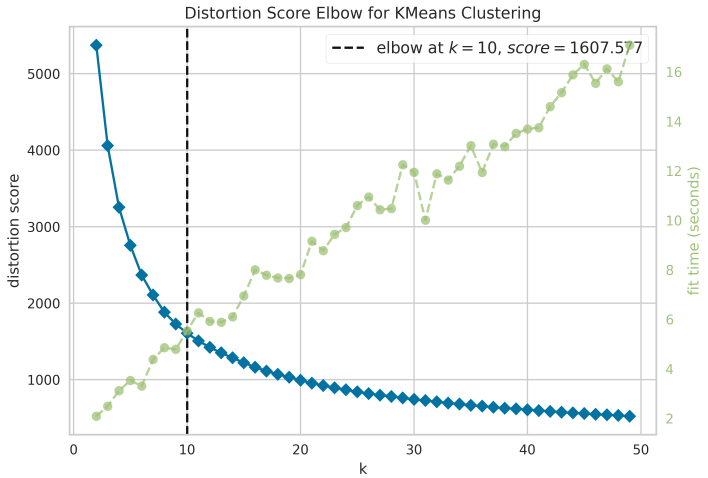


Figure 4.6: Elbow diagram results for OUD-positive male patients. The optimum number of clusters is 10.

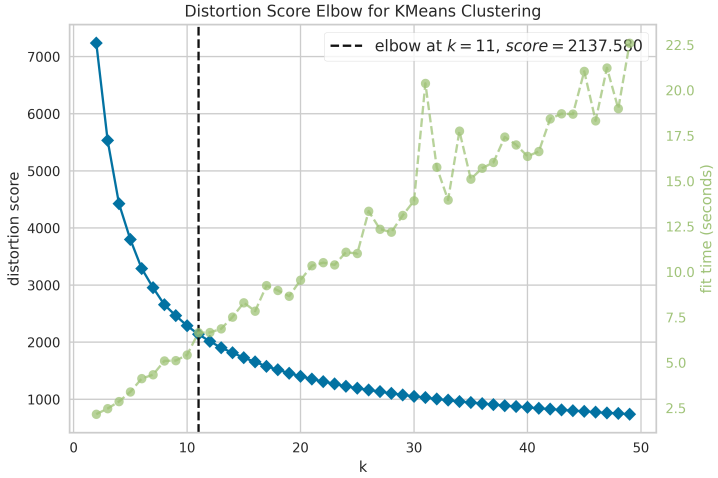


Figure 4.7: Elbow diagram results for OUD-positive female patients. The optimum number of clusters is 11.

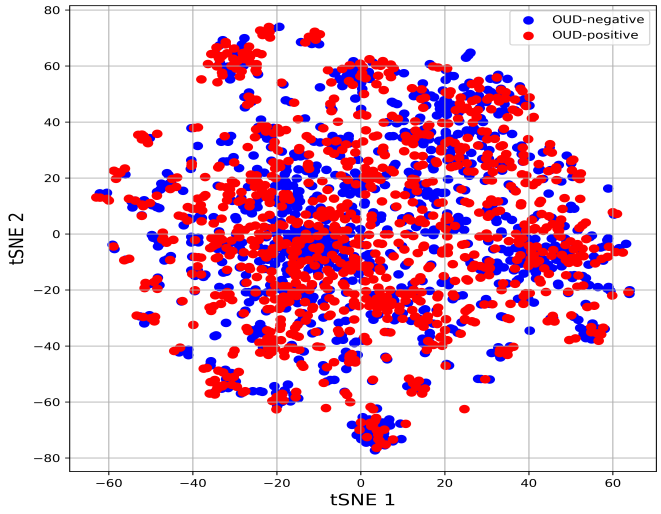


Figure 4.8: tSNE visualization of OUD-positives and OUD-negatives.

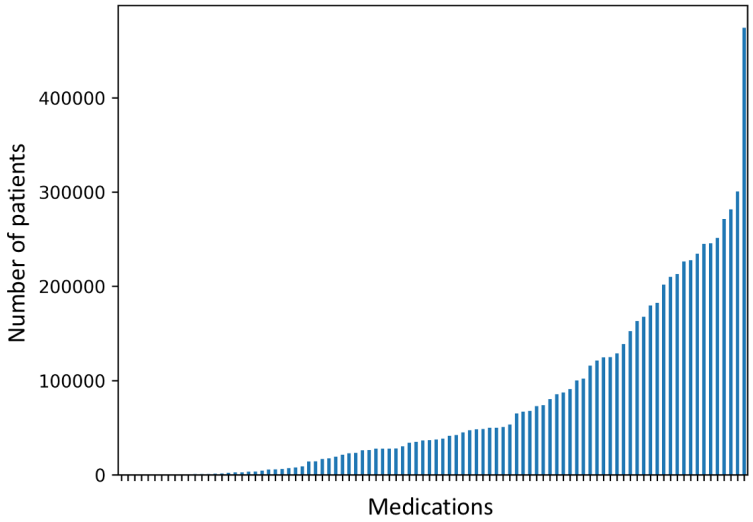


Figure 4.9: Distribution of medication features frequencies.

shows the number of patients for whom the frequency of occurring f_i is larger than zero.

To reduce the sparsity of data and due to hardware limitations, the features outside of the mean plus two standard deviations of the distributions were excluded from the feature space. This reduced the number of features from 623 to 270 features (50 medications, 138 diagnoses, 80 rocedures and 2 demographics). Similar to the process described in Figure 4.3, the LSTM models were used to represent patients' data. However, since this study included procedures in addition to medications and diag-

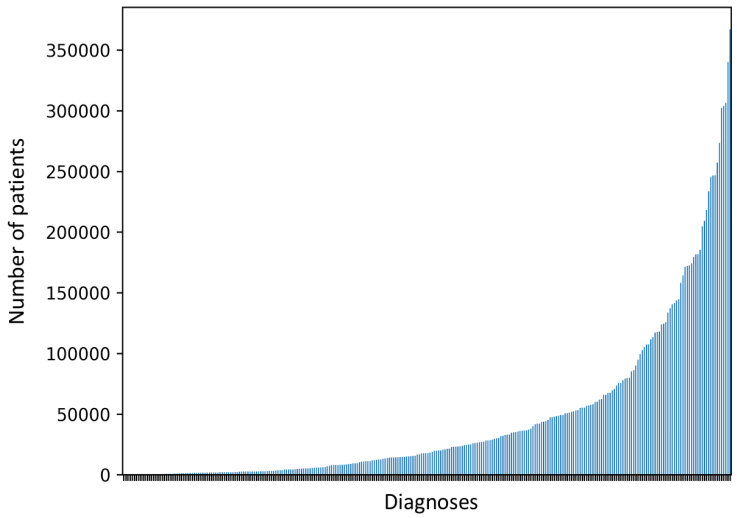


Figure 4.10: Distribution of diagnosis features frequencies.

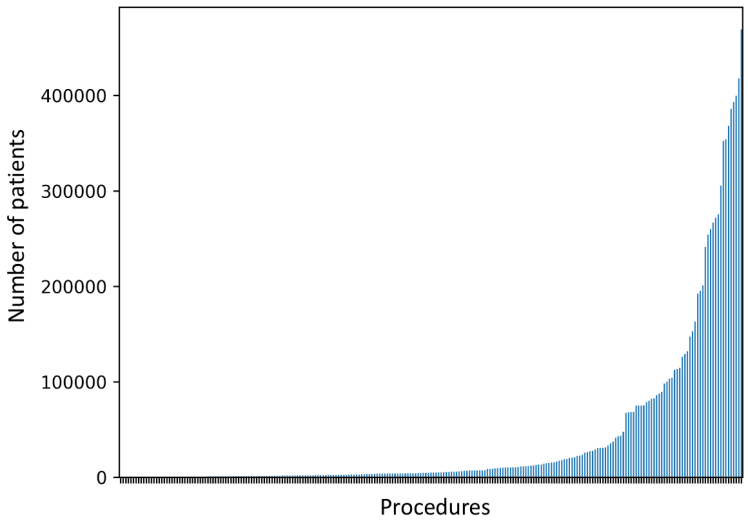


Figure 4.11: Distribution of procedure features frequencies.

noses, an extra LSTM model was trained for procedures stream as well. Note, the LSTM for representing medications and diagnoses were re-trained too as the cohort and features are changed in this study. The represented medications, diagnoses and procedures were then used to train a multi-stream transformer model described in the following section.

4.4.2 Multi-stream Transformer

The model proposed in this study is a transformer based model to analyze temporal claims data. Each patient can be represented by $P = (S, y)$ in which S is a set of input sequences and y is the target label. Here in, three input sequences are considered as a common approach in healthcare data analysis: 1) medications tuples (t_i, M_i) in which t_i is the i^{th} time step and M_i is a list of medications that the patient used (is prescribed with) at time i^{th}, 2) diagnoses tuples (t_i, D_i) where t_i is the i^{th} time step and D_i is a list of diagnoses that have been assigned to the patient P at time t_i, 3) procedures tuple (t_i, P_i) in which t_i is the i^{th} time step and P_i is a list of procedures that patient P has gone through at time t_i. The opioid use disorder is considered as an example application of the proposed method in this study and therefore the label y_i defines if the patient P_i is diagnosed with OUD or not. Figure 4.12 shows the high-level overview of the proposed model.

Medications M, diagnoses D, and procedures P are fed to the model at the same time and in parallel. The first step is to incorporate the temporal pattern of the visits into raw inputs. The encoded input sequences are plugged into the Attention layer which is the most important component of the proposed model. This layer first uses fully connected network to generate query, key and value matrixes for each of the input streams. For example, medications M are fed to the fully connected layer to generate M_Q, M_K, and M_V which are query, key, and value matrixes for the medications stream for the patient p. These query, key, and value matrixes are used to find the attention weights across different input sequences using 4.3.

$$Attention(X_Q, Y_K, Y_V) = softmax(\frac{X_Q Y_K^\top}{\sqrt{d_k}})Y_V \tag{4.3}$$

$$Where \quad X, Y \in \{M, D, P\} \tag{4.4}$$

Figure 4.13 describes how the attention weights are calculated across different input streams and converted to new outputs for the next encoder layer. In figure 4.13, X_Q, X_K, and X_V represent query, key, and value matrixes for input stream

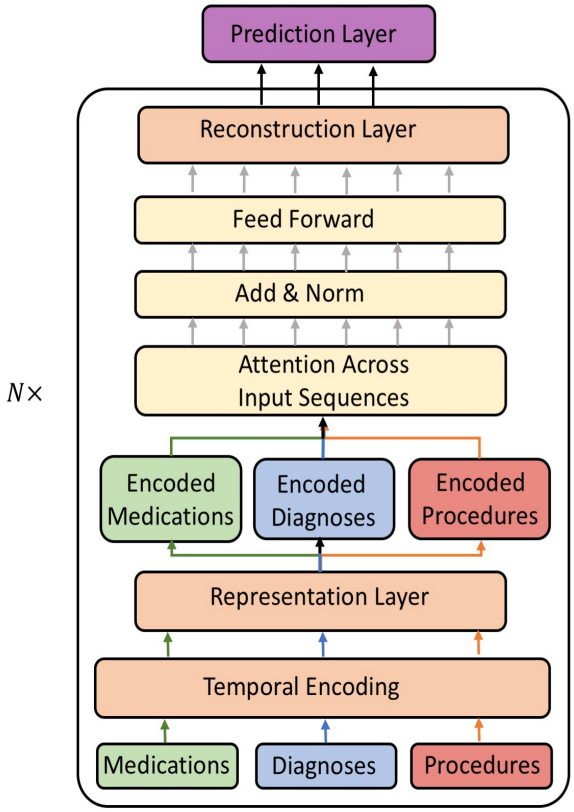


Figure 4.12: General architecture of the multi-stream transformer model.

X, where $X \in \{M, D, P\}$. All possible permutations of the input streams is used to find the attention weights between different visits and across different input sequences; medications, diagnoses, and procedures. Attentions are then passed through a dense layer to generate outputs. In this application, there are three input streams (medications, diagnoses, and procedures), and this leads to six permutations (MM, MD, MP, DD, DP, PP, where M, D, and P represent medications, diagnoses, and procedures, respectively).

Reconstruction layer gets relevant output and map them to appropriate format for the next layer as described in 4.5. For example, only the outputs relevant to the medications (M) including O_{MM}, O_{MD}, and O_{MP} are used to reconstruct the medication stream appropriate to be fed into the next encoder layer. The three reconstructed matrixes generated by the last encoder layer are fed to a prediction layer to make the final prediction for the current patient p. The prediction layer is implemented in 4.6.

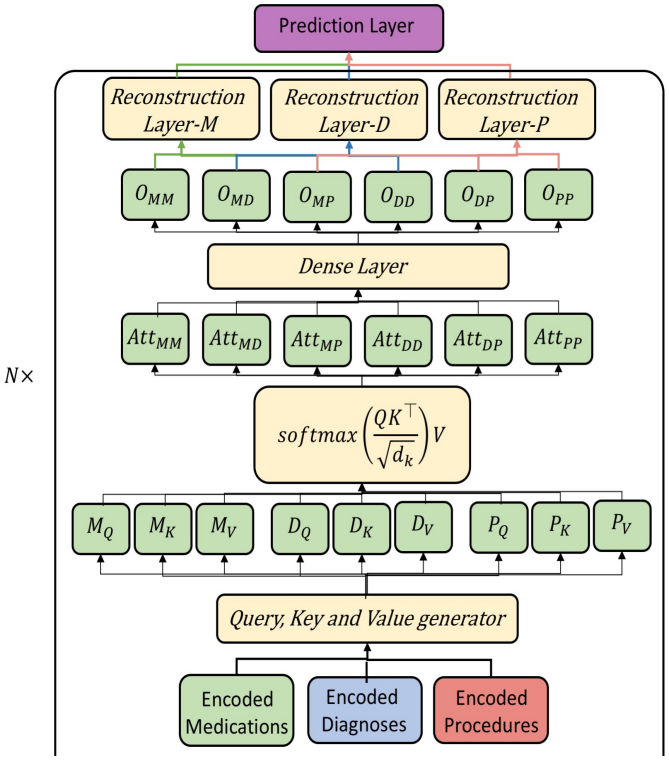


Figure 4.13: Self attention layer of the multi-stream transformer. Proposed model-output calculation across medications, diagnoses, and procedure. $X_Q, X_K, and \quad X_V$ represent query, key, and value matrixes for input stream X, where $X \in \{Medication, Diagnoses, Procedures\}$. Att_{XY} represents the attention weights between different visits of patient P and across input streams X and Y, where $X, Y \in \{Medications, Diagnoses, Procedures\}$. O_{XY} represents the outputs which captures the association between input streams X and Y, where $X, Y \in \{Medications, Diagnoses, Procedures\}$.

$$\hat{X} = f(O_{XX}, O_{XY}, O_{XZ})$$
$$f : O_{XX}, O_{XY}, O_{XZ} \subset \{O_{MM}, O_{MD}, O_{MP}, O_{DD}, O_{DP}, O_{PP}\} \longrightarrow \hat{X}$$
$$where$$
$$\hat{X} \in \{\hat{M}, \hat{D}, \hat{P}\} \quad and \quad X, Y, Z \in \{M, D, P\} \tag{4.5}$$

$$Prediction = g(\hat{M}, \hat{D}, \hat{P})$$

$$g : \hat{M}, \hat{D}, \hat{P} \longrightarrow Prediction$$

$$where$$

$$g(\hat{M}, \hat{D}, \hat{P}) = [Concat(\hat{M}, \hat{D}, \hat{P})]W + b \tag{4.6}$$

In which, *average* is an element-wise averaging function, W and b are trainable weight and bias matrixes to generate the last predictions.

4.4.3 Multi-stream Transformer: Results

The proposed multi-stream transformer was trained and tested using 474,208 patients' data over 12 years (2009-2020). The feature set at each time step include 270 medications, diagnoses , procedures and demographics features. Logistic regression, random forest, LSTM and the encoder part of the transformer model with a fully connected layer on top of it were used as baselines. 90 percent of the data was used for training and validation and all models were tested on the same unseen test set which included 10 percent of the data selected randomly. The classical machine learning models (random forest and logistic regression) were fine tuned using a randomized cross validation. However, deep learning models were only trained once due to high time and hardware complexity and their performance on the test set are reported.

Table 4.4 compares the performance of the deep learning models including the multi-stream transformer model with the classical machine learning models. Deep learning models show a better performance in predicting OUD 6 months before the onset of this disorder. Among the deep learning models, multi-stream transformer predicts OUD with higher recall, F1-score and AUC. This results show that the proposed multi-stream transformer can generally be more accurate and efficient in predicting OUD 6 months prior to the onset of OUD and using 270 features.

Further, the performance of the models were tested on three imbalanced test sets with OUD-positive population size to OUD-negative population size ratio being 0.5, 0.2 and 0.1. Table 4.4 shows the performance of logistic regression, random forest, LSTM, single-stream transformer and multi-stream transformer on these imbalanced test sets. Note, the models are trained on the balanced train set. The prediction performance of the models drops as the test set becomes more imbalance. However, the multi-stream transformer still has the highest AUC and recall compared to the

Table 4.4: Performance of OUD classification using multi-stream transformer compared to RF, SVM, LSTM and original transformer.

Model	**Acc.**	**Prec.**	**Rec.**	**F1-score**	**AUC**
LR	0.609	0.630	0.519	0.569	0.651
RF	0.631	0.624	0.650	0.637	0.679
LSTM	0.668	0.648	0.738	0.690	0.732
Single-stream transformer	0.651	0.603	0.874	0.713	0.725
Multi-stream transformer	0.652	0.602	0.890	0.718	0.742

Table 4.5: OUD classification results for imbalanced test sets. The $.xN$ means the number of samples in the OUD-positive cohort are $0.x$ times smaller than the number of samples in the OUD-negative cohort.

Model	Precision			Recall			F1-score			AUC		
	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$	$.5N$	$.2N$	$.1N$
LR	**.460**	**.254**	.146	.518	.517	.522	.487	.340	.229	.650	.651	.655
RF	.454	.247	.145	.651	.644	.663	.535	.357	.237	.679	.675	.686
LSTM	.457	.250	**.147**	.718	.712	.733	.559	**.371**	**.245**	.706	.704	.716
Transformer	.431	.232	.132	.872	.870	.872	.577	.367	.229	.724	.721	.724
MUPOD	.430	.230	.132	**.888**	**.883**	**.897**	**.579**	.366	.230	**.741**	**.737**	**.746**

other models. LSTM and logistic regression also showed good performance in some cases in terms of precision and F1-score.

Chapter 5 Discussion

In this thesis, multiple deep learning based models were proposed to analyze longitudinal claims data collected from multiple sources. The first proposed model was a LSTM based framework to analyze claims and EHR data, and create disease prediction tools. This model addresses the challenges in analyzing temporal claims data such as sparsity, class imbalance problem, and temporal dependency. This framework has three main components: 1) data pre-processing which convert the data set format into an appropriate input format for recurrent neural networks training, 2) disease prediction using LSTM models, 3) hypothesis exploration by varying model and its inputs. The proposed framework was used to perform disease prediction on substance use disorder and mild cognitive impairment. In the first case, IBM MarketScan claims data was used and in the latter study, the Mayo Clinic Study of Aging Data was used. LSTM models in the proposed framework performed better in predicting these diseases compared to classical machine learning in both cases.

However, recent advances in deep learning has shown the inefficiency of recurrent neural networks in terms of computational performance and accuracy in processing long sequences. This problem is originated in the sequential nature of recurrent neural networks. Factorization techniques[111], attention based models[5,6], and transformers[7] were proposed to improve computational efficiency and the RNN' performance in analyzing longer sequences. The transformer based models was proposed entirely based on attention mechanism and removed the sequential nature of RNNs. These models outperformed state of the art deep learning models in almost all of sequential data analysis tasks. However, transformers were generally developed for natural language processing tasks and originally designed to analyze a single sequence of words.

Here, the predictive modeling capabilities of the transformer was compared with LSTMs and traditional machine learning in OUD prediction. Further, a novel transformer based model was proposed to efficiently apply them on longitudinal claim data. The transformer was traditionally designed to analyze a single sequence of words. The proposed transformer in this thesis offers a more efficient model to analyze multiple sequences. In fact, the proposed model extracts the associations within and across different sources of patient information including medications, diagnoses, and procedures.

The proposed transformer is designed to simultaneously analyze multiple types

of healthcare data streams, such as medications and diagnoses, by attending to segments within and across these data streams. This model along with a single stream transformer and classical machine learning models such as random forest and logistic regression were trained on more than 470K patients' data including 270 medications, diagnoses, procedures and demographics in one experiment, and on 390K patinets' data including 22 medications, diagnoses and demographics features in another experiment. In both experiments the deep learning models and especially the proposed multi-stream transformer performed better in predicting opioid use disorder. The experimental results conducted in this thesis shows that the proposed multi-stream transformer is an efficient model that is capable of predicting opioid use disorder from 1-6 months before the onset of this disorder significantly better than the traditional models and recently developed deep learning models. Further, the explainability of the proposed model was investigated and it showed that the attention weights computed by the multi-stream transformer can be utilized to provide some explanations on the predictions provided by the model.

5.1 Limitations

There are some limitations in this thesis that need to be considered before using the proposed models. Although the proposed multi-stream transformer uses a thorough history of medications, diagnoses, procedures and demographics information of patients, there are other aspects of patient information that can help boost the prediction performances. For example, the opioid dosage which is typically measured by Morphine Milligram Equivalent (MME), can be utilized to increase the performance of the models. The MME can be incorporated as a parallel signal along with medications, diagnoses and procedures or it can be added to the system in the last layer when the model is performing the final predictions. Second, the explainability of the proposed model was explored using a few representative samples. These samples could show how the multi-stream transformer can discriminate OUD-positive samples from OUD-negatives and provided intuitions in how medications and diagnoses are related for those patients. However, these attentions cannot be used to infer causal associations.

Another limitation of this thesis is that the models are trained and tested on the same data bases. In fact, the multi-stream transformer was trained and tested only on IBM MarketScan data. Note, the test set was still randomly selected from IBM MarketScan and was unseen during the training process for all models. However,

testing the proposed models on data sets other than IBM MarketScan is needed to further test the generalizability of the models.

Bibliography

[1] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges.," *Briefings in Bioinformatics,*, pp. 1–11,, 2017.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[3] A. Graves, "Generating Sequences With Recurrent Neural Networks," *ArXiv e-prints*, Aug. 2013.

[4] F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1903–1911, 2017.

[5] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 1412–1421, Association for Computational Linguistics, 2015.

[6] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 5998–6008, Curran Associates, Inc., 2017.

[8] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," in *Machine Learning for Healthcare Conference*, pp. 301–318, 2016.

[9] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," *arXiv preprint arXiv:1608.05745*, 2016.

[10] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia conference on knowledge discovery and data mining*, pp. 30–41, Springer, 2016.

[11] E. Choi, Z. Xu, Y. Li, M. Dusenberry, G. Flores, E. Xue, and A. Dai, "Learning the graphical structure of electronic health records with graph convolutional transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 606–613, 2020.

[12] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[13] Y. Wang, X. Xu, T. Jin, X. Li, G. Xie, and J. Wang, "Inpatient2vec: Medical representation learning for inpatients," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1113–1117, IEEE, 2019.

[14] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," *arXiv preprint arXiv:1906.00346*, 2019.

[15] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "Behrt: transformer for electronic health records," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] W. W. L. Z. Z. C. Bing Li, Wei Cui and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[18] K. J. Han, R. Prieto, and T. Ma, "State-of-the-art speech recognition using multi-stream self-attention with dilated 1d convolutions," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 54–61, IEEE, 2019.

[19] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," in *European Conference on Computer Vision*, pp. 301–319, Springer, 2020.

[20] R. Hu and A. Singh, "Transformer is all you need: Multimodal multitask learning with a unified transformer," *arXiv preprint arXiv:2102.10772*, 2021.

[21] J. Libovickỳ, J. Helcl, and D. Mareček, "Input combination strategies for multi-source transformer decoder," *arXiv preprint arXiv:1811.04716*, 2018.

[22] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[23] L. O. Gostin, J. G. Hodge, and S. A. Noe, "Reframing the opioid epidemic as a national emergency," *Jama*, vol. 318, no. 16, pp. 1539–1540, 2017.

[24] A. L. Pitt, K. Humphreys, and M. L. Brandeau, "Modeling health benefits and harms of public policy responses to the us opioid epidemic," *American journal of public health*, vol. 108, no. 10, pp. 1394–1400, 2018.

[25] Y. Olsen and J. M. Sharfstein, "Confronting the stigma of opioid use disorder—and its treatment," *Jama*, vol. 311, no. 14, pp. 1393–1394, 2014.

[26] L.-T. Wu, H. Zhu, and M. S. Swartz, "Treatment utilization among persons with opioid use disorder in the united states," *Drug and alcohol dependence*, vol. 169, pp. 117–127, 2016.

[27] "National ambulatory medical care survey: 2016 national summary tables.," tech. rep., Centers for Disease Control and Prevention.

[28] "Risk assessment: Safe opioid prescribing tools." `https://www.practicalpainmanagement.com/resource-centers/opioid-prescribing-monitoring/risk-assessment-safe-opioid-prescribing-tools`. Accessed: July 06, 2021.

[29] W. Gao, C. Leighton, Y. Chen, J. Jones, and P. Mistry, "Predicting opioid use disorder and associated risk factors in a medicaid managed care population.," *The American Journal of Managed Care*, vol. 27, no. 4, pp. 148–154, 2021.

[30] L. R. Webster and R. M. Webster, "Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the opioid risk tool," *Pain medicine*, vol. 6, no. 6, pp. 432–442, 2005.

[31] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential.," *health information science and systems,*, vol. 2, no. 3, 2014.

[32] Z. Segal, K. Radinsky, G. Elad, G. Marom, M. Beladev, M. Lewis, B. Ehrenberg, P. Gillis, L. Korn, and G. Koren, "Development of a machine learning algorithm for early detection of opioid use disorder," *Pharmacology Research & Perspectives*, vol. 8, no. 6, p. e00669, 2020.

[33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[34] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, no. 2, pp. 197–227, 2016.

[35] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[37] G. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *NIPS*, vol. 15, pp. 833–840, Citeseer, 2002.

[38] D. M. Chan, R. Rao, F. Huang, and J. F. Canny, "t-sne-cuda: Gpu-accelerated t-sne and its applications to modern data," in *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*, pp. 330–338, IEEE, 2018.

[39] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649, Ieee, 2013.

[40] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to Diagnose with LSTM Recurrent Neural Networks," *ArXiv e-prints*, Nov. 2015.

[41] Z. C. Lipton, D. C. Kale, and R. C. Wetzel, "Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks," *ArXiv e-prints*, Oct. 2015.

[42] A. N. Jagannatha and H. Yu, "Structured prediction models for rnn based sequence labeling in clinical text.," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 2016, pp. 856–865, NIH Public Access, 2016.

[43] S. Fouladvand, E. R. Hankosky, H. Bush, J. Chen, L. P. Dwoskin, P. R. Freeman, D. W. Henderson, K. Kantak, J. Talbert, S. Tao, and G.-Q. Zhang, "Predicting substance use disorder using long-term attention deficit hyperactivity disorder medication records in truven," *Health Informatics Journal*. PMID: 31106686.

[44] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult.," *IEEE transactions on neural networks,*, vol. 5, no. 2, pp. 157–166, 1994.

[45] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," *ArXiv e-prints*, Sept. 2014.

[46] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[47] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *arXiv preprint arXiv:1906.08237*, 2019.

[48] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[49] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[50] M. L. Danielson, R. H. Bitsko, R. M. Ghandour, J. R. Holbrook, M. D. Kogan, and S. J. Blumberg, "Prevalence of parent-reported adhd diagnosis and associated treatment among u.s. children and adolescents.," *Journal of Clinical Child and Adolescent Psychology,*, vol. 47, no. 2, pp. 199–212, 2016.

[51] "Centers for disease control and prevention data and statistics, 2011 and national survey of children's health database 2011/2012.," 2012.

[52] R. C. Harvey, S. Sen, A. Deaciuc, L. P. Dwoskin, and K. M. Kantak, "Methylphenidate treatment in adolescent rats with an attention deficit/hyperactivity disorder phenotype: Cocaine addiction vulnerability and dopamine

transporter function.," *Neuropsychopharmacology,*, vol. 36, no. 4, pp. 837–847, 2011.

[53] S. S. Somkuwar, C. J. Jordan, K. M. Kantak, and L. P. Dwoskin, "Adolescent atomoxetine treatment in a rodent model of adhd: Effects on cocaine self-administration and dopamine transporters in frontostriatal regions.," *Neuropsychopharmacology,*, vol. 38, no. 13, pp. 2588–2597, 2013.

[54] C. J. Jordan, R. C. Harvey, B. B. Baskin, L. P. Dwoskin, and K. M. Kantak, "Cocaine-seeking behavior in a genetic model of attention-deficit/hyperactivity disorder following adolescent methylphenidate or atomoxetine treatments.," *Drug Alcohol Depend,*, vol. 140, pp. 25–32, 2014.

[55] B. M. Baskin, L. P. Dwoskin, and K. M. Kantak, "Methylphenidate treatment beyond adolescence maintains increased cocaine self-administration in the spontaneously hypertensive rat model of attention deficit/hyperactivity disorder.," *Pharmacology Biochemistry and Behavior,*, vol. 131, pp. 51–56, 2015.

[56] C. J. Jordan, C. Lemay, L. P. Dwoskin, and K. M. Kantak, "Adolescent d-amphetamine treatment in a rodent model of attention deficit/hyperactivity disorder: Impact on cocaine abuse vulnerability in adulthood.," *Psychopharmacology,*, vol. 233, pp. 3891–3903, 2016.

[57] C. J. Jordan, D. M. Taylor, L. P. Dwoskin, and K. M. Kantak, "Adolescent d-amphetamine treatment in a rodent model of adhd: Pro-cognitive effects in adolescence without an impact on cocaine cue reactivity in adulthood.," *Behavioural Brain Research,*, vol. 297, pp. 165–179, 2016.

[58] B. M. Baskin, B. A. N. Dhonnchadha, L. P. Dwoskin, and K. M. Kantak, "Blockade of α2-adrenergic receptors in prelimbic cortex: impact on cocaine self-administration in adult spontaneously hypertensive rats following adolescent atomoxetine treatment.," *Psychopharmacology,*, vol. 234, pp. 2897–2909, 2017.

[59] E. van den Ban, P. C. Souverein, H. Swaab, H. van Engeland, T. C. G. Egberts, and E. R. Heerdink, "Less discontinuation of adhd drug use since the availability of long-acting adhd medication in children, adolescents and adults under the age of 45 years in the netherlands.," *ADHD Attention Deficit and Hyperactivity Disorders,*, vol. 2, no. 4, pp. 213–220, 2010.

[60] J. L. Lund, D. B. Richardson, and T. Stürmer, "The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application.," *Current Epidemiology Reports,*, vol. 2, no. 4, pp. 221–228, 2015.

[61] M. A. Mazurowski, P. A. Habasa, J. M. Zuradaa, J. Y. Lob, J. A. Bakerb, and G. D. Tourassib, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance.," *Neural Networks,*, vol. 21, pp. 427–436, 2008.

[62] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *ArXiv e-prints*, Oct. 2017.

[63] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study.," *Intelligent Data Analysis,*, vol. 6, pp. 429–449, 2002.

[64] M. Steinbach, G. Karypis, V. Kumar, *et al.*, "A comparison of document clustering techniques.," in *KDD workshop on text mining*, vol. 400, pp. 525–526, Boston, 2000.

[65] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: a system for large-scale machine learning.," in *12th (USENIX) Symposium on Operating Systems Design and Implementation (OSDI 16)*, vol. 16, (Savannah, GA), pp. 265–283, 2016.

[66] C. Cortes and V. Vapnik, "Support-vector networks.," *Machine Leaming,*, vol. 20, pp. 273–297, 1995.

[67] L. Breiman, "Random forests.," *Machine Learning,*, vol. 45, no. 1, pp. 5–32, 2001.

[68] E. Alpaydin, *Introduction to Machine Learning.* MIT Press, 2004.

[69] "Dementia statistics." `https://www.alzint.org/about/dementia-facts-figures/dementia-statistics/`. Accessed: September 28, 2021.

[70] S. Goudarzvand, J. S. Sauver, M. M. Mielke, P. Y. Takahashi, and S. Sohn, "Analyzing early signals of older adult cognitive impairment in electronic health records," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1636–1640, IEEE, 2018.

[71] S. Goudarzvand, J. St Sauver, M. M. Mielke, P. Y. Takahashi, Y. Lee, and S. Sohn, "Early temporal characteristics of elderly patient cognitive impairment in electronic health records," *BMC medical informatics and decision making*, vol. 19, no. 4, pp. 1–14, 2019.

[72] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1589–1604, 2017.

[73] V. S. Pankratz, R. O. Roberts, M. M. Mielke, D. S. Knopman, C. R. Jack, Y. E. Geda, W. A. Rocca, and R. C. Petersen, "Predicting the risk of mild cognitive impairment in the mayo clinic study of aging," *Neurology*, vol. 84, no. 14, pp. 1433–1442, 2015.

[74] M. M. Mielke, H. J. Wiste, S. D. Weigand, D. S. Knopman, V. J. Lowe, R. O. Roberts, Y. E. Geda, D. M. Swenson-Dravis, B. F. Boeve, M. L. Senjem, *et al.*, "Indicators of amyloid burden in a population-based study of cognitively normal elderly," *Neurology*, vol. 79, no. 15, pp. 1570–1577, 2012.

[75] M. M. Mielke, R. O. Roberts, R. Savica, R. Cha, D. I. Drubach, T. Christianson, V. S. Pankratz, Y. E. Geda, M. M. Machulda, R. J. Ivnik, *et al.*, "Assessing the temporal relationship between cognition and gait: slow gait predicts cognitive decline in the mayo clinic study of aging," *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences*, vol. 68, no. 8, pp. 929–937, 2013.

[76] Y. E. Geda, R. O. Roberts, D. S. Knopman, R. C. Petersen, T. J. Christianson, V. S. Pankratz, G. E. Smith, B. F. Boeve, R. J. Ivnik, E. G. Tangalos, *et al.*, "Prevalence of neuropsychiatric symptoms in mild cognitive impairment and normal cognitive aging: population-based study," *Archives of general psychiatry*, vol. 65, no. 10, pp. 1193–1198, 2008.

[77] R. O. Roberts, D. S. Knopman, Y. E. Geda, R. H. Cha, V. S. Pankratz, L. Baertlein, B. F. Boeve, E. G. Tangalos, R. J. Ivnik, M. M. Mielke, *et al.*, "Association of diabetes with amnestic and nonamnestic mild cognitive impairment," *Alzheimer's & Dementia*, vol. 10, no. 1, pp. 18–26, 2014.

[78] R. C. Petersen, R. O. Roberts, D. S. Knopman, Y. E. Geda, R. H. Cha, V. Pankratz, B. Boeve, E. Tangalos, R. Ivnik, and W. Rocca, "Prevalence of

mild cognitive impairment is higher in men: The mayo clinic study of aging," *Neurology*, vol. 75, no. 10, pp. 889–897, 2010.

[79] R. O. Roberts, Y. E. Geda, D. S. Knopman, R. H. Cha, V. S. Pankratz, B. F. Boeve, R. J. Ivnik, E. G. Tangalos, R. C. Petersen, and W. A. Rocca, "The mayo clinic study of aging: design and sampling, participation, baseline measures and sample characteristics," *Neuroepidemiology*, vol. 30, no. 1, pp. 58–69, 2008.

[80] M. Albert, Y. Zhu, A. Moghekar, S. Mori, M. I. Miller, A. Soldan, C. Pettigrew, O. Selnes, S. Li, and M.-C. Wang, "Predicting progression from normal cognition to mild cognitive impairment for individuals at 5 years," *Brain*, vol. 141, no. 3, pp. 877–887, 2018.

[81] B. Stephan and C. Brayne, "Risk factors and screening methods for detecting dementia: a narrative review," *Journal of Alzheimer's Disease*, vol. 42, no. s4, pp. S329–S338, 2014.

[82] A. Pozueta, E. Rodríguez-Rodríguez, J. L. Vazquez-Higuera, I. Mateo, P. Sánchez-Juan, S. González-Perez, J. Berciano, and O. Combarros, "Detection of early alzheimer's disease in mci patients by the combination of mmse and an episodic memory test," *BMC neurology*, vol. 11, no. 1, pp. 1–5, 2011.

[83] J. J. Gomar, C. Conejero-Goldberg, P. Davies, T. E. Goldberg, A. D. N. Initiative, *et al.*, "Extension and refinement of the predictive value of different classes of markers in adni: four-year follow-up data," *Alzheimer's & Dementia*, vol. 10, no. 6, pp. 704–712, 2014.

[84] S.-S. Poil, W. De Haan, W. M. van der Flier, H. D. Mansvelder, P. Scheltens, and K. Linkenkaer-Hansen, "Integrative eeg biomarkers predict progression to alzheimer's disease at the mci stage," *Frontiers in aging neuroscience*, vol. 5, p. 58, 2013.

[85] E. Ford, N. Greenslade, P. Paudyal, S. Bremner, H. E. Smith, S. Banerjee, S. Sadhwani, P. Rooney, S. Oliver, and J. Cassell, "Predicting dementia from primary care records: A systematic review and meta-analysis," *PLoS One*, vol. 13, no. 3, p. e0194735, 2018.

[86] I. H. Ramakers, P. J. Visser, P. Aalten, J. H. Boesten, J. F. Metsemakers, J. Jolles, and F. R. Verhey, "Symptoms of preclinical dementia in general prac-

tice up to five years before dementia diagnosis," *Dementia and geriatric cognitive disorders*, vol. 24, no. 4, pp. 300–306, 2007.

[87] R. Zhang, G. Simon, and F. Yu, "Advancing alzheimer's research: A review of big data promises," *International journal of medical informatics*, vol. 106, pp. 48–56, 2017.

[88] M. Van Gils, J. Koikkalainen, J. Mattila, S. Herukka, J. Lötjönen, and H. Soininen, "Discovery and use of efficient biomarkers for objective disease state assessment in alzheimer's disease," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 2886–2889, IEEE, 2010.

[89] M. Li, K. Oishi, X. He, Y. Qin, F. Gao, S. Mori, and A. D. N. Initiative, "An efficient approach for differentiating alzheimer's disease from normal elderly based on multicenter mri using gray-level invariant features," *PloS one*, vol. 9, no. 8, p. e105563, 2014.

[90] O. Kohannim, X. Hua, D. P. Hibar, S. Lee, Y.-Y. Chou, A. W. Toga, C. R. Jack Jr, M. W. Weiner, P. M. Thompson, A. D. N. Initiative, *et al.*, "Boosting power for clinical trials using classifiers based on multiple biomarkers," *Neurobiology of aging*, vol. 31, no. 8, pp. 1429–1442, 2010.

[91] S. Lovestone, P. Francis, and K. Strandgaard, "Biomarkers for disease modification trials-the innovative medicines initiative and addneuromed," *The journal of nutrition, health & aging*, vol. 11, no. 4, p. 359, 2007.

[92] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 305–312, Springer, 2014.

[93] Y. Shmulev, M. Belyaev, A. D. N. Initiative, *et al.*, "Predicting conversion of mild cognitive impairments to alzheimer's disease and exploring impact of neuroimaging," in *Graphs in biomedical image analysis and integrating medical imaging and non-imaging modalities*, pp. 83–91, Springer, 2018.

[94] R. C. Petersen, "Mild cognitive impairment as a diagnostic entity," *Journal of internal medicine*, vol. 256, no. 3, pp. 183–194, 2004.

[95] H. Liu, S. J. Bielinski, S. Sohn, S. Murphy, K. B. Wagholikar, S. R. Jonnalagadda, K. Ravikumar, S. T. Wu, I. J. Kullo, and C. G. Chute, "An information

extraction framework for cohort identification using electronic health records," *AMIA Summits on Translational Science Proceedings*, vol. 2013, p. 149, 2013.

[96] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, 2008.

[97] B. K. Beaulieu-Jones, C. S. Greene, *et al.*, "Semi-supervised learning of the electronic health record for phenotype stratification," *Journal of biomedical informatics*, vol. 64, pp. 168–178, 2016.

[98] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

[99] P. Gracia-García, C. De-La-Cámara, J. Santabárbara, R. Lopez-Anton, M. A. Quintanilla, T. Ventura, G. Marcos, A. Campayo, P. Saz, C. Lyketsos, *et al.*, "Depression and incident alzheimer disease: the impact of disease severity," *The American Journal of Geriatric Psychiatry*, vol. 23, no. 2, pp. 119–129, 2015.

[100] R. C. Petersen, E. S. Lundt, T. M. Therneau, S. D. Weigand, D. S. Knopman, M. M. Mielke, R. O. Roberts, V. J. Lowe, M. M. Machulda, W. K. Kremers, *et al.*, "Predicting progression to mild cognitive impairment," *Annals of neurology*, vol. 85, no. 1, pp. 155–160, 2019.

[101] R. Roberts and D. S. Knopman, "Classification and epidemiology of mci," *Clinics in geriatric medicine*, vol. 29, no. 4, pp. 753–772, 2013.

[102] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[103] "IBM MarketScan Research Databases." `https://www.ibm.com/products/marketscan-research-databases`. Accessed: March 10, 2021.

[104] J. A. Rice, *Mathematical statistics and data analysis*. Nelson Education, 2006.

[105] T. L. Mark, J. Dilonardo, R. Vandivort, and K. Miller, "Psychiatric and medical comorbidities, associated pain, and health care utilization of patients prescribed buprenorphine," *Journal of substance abuse treatment*, vol. 44, no. 5, pp. 481–487, 2013.

[106] P. G. Barnett, "Comparison of costs and utilization among buprenorphine and methadone patients," *Addiction*, vol. 104, no. 6, pp. 982–992, 2009.

[107] H. Clarke, N. Soneji, D. T. Ko, L. Yun, and D. N. Wijeysundera, "Rates and risk factors for prolonged opioid use after major surgery: population based cohort study," *Bmj*, vol. 348, 2014.

[108] S. Serrano and N. A. Smith, "Is attention interpretable?," *arXiv preprint arXiv:1906.03731*, 2019.

[109] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[110] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *arXiv preprint arXiv:1908.04626*, 2019.

[111] O. Kuchaiev and B. Ginsburg, "Factorization tricks for lstm networks," *ArXiv*, vol. abs/1703.10722, 2017.

Vita

Sajjad Fouladvand

Education

- Shahid Chamran University of Ahvaz, Khouzestan, Iran — Sep. 2013
  MSc. in Computer Engineering.
- University of Zanjan, Zanjan, Iran — Aug. 2011
  B.S. in Computer Software Technology Engineering.

Professional Positions

- Graduate Research Assistant, University of Kentucky — Aug. 2017–Oct. 2021
- Summer Intern, Mayo Clinic — May 2019–Jul. 2019

Leadership and Awards

- PI: Sajjad Fouladvand. An Application for Predicting Coronary Heart Disease using Artificial Intelligence Methods and a Case-control Study, Sponsor: Lorestan University of Medical Science, Lorestan, Iran, Duration: 2015-2017.
- Mentored a masters computer science student on a project focused on development of deep learning models intended to analyze big longitudinal biomedical data using LSTM and Transformer models.

Selected Publications

- **S. Fouladvand**, J. Talbert, L. P. Dwoskin, H. Bush, A. L. Meadows, L. E. Peterson, R. Kavuluru, and J. Chen, "Predicting opioid use disorder from longitudinal healthcare data using multi-stream transformer," in American Medical Informatics Association Annual Symposium (AMIA 2021), (San Diego, California, USA), 2021 (in press).
- **S. Fouladvand**, E. R. Hankosky, H. Bush, J. Chen, L. P. Dwoskin, P. R. Freeman, D. W. Henderson, K. Kantak, J. Talbert, S. Tao, and G. Q. Zhang, "Predicting substance use disorder using long-term ADHD medication records in Truven," Health Informatics Journal, vol. 26, pp. 787–802, 2020.

- **S. Fouladvand**, M. M. Mielke, M. Vassilaki, J. St Sauver, R. C. Petersen, and S. Sohn, "Deep learning prediction of mild cognitive impairment using electronic health records," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2019), (San Diego, California, USA), IEEE, 2019.

- **S. Fouladvand**, A. Osareh, B. Shadgar, M. Pavone, and S. Sharafi, "DENSA: An effective negative selection algorithm with flexible boundaries for self-space and dynamic number of detectors," Engineering Applications of Artificial Intelligence, vol. 62, pp. 359–372, 2017.

Selected Presentations

- **S. Fouladvand**, E. R. Hankosky, D. W. Henderson, H. Bush, J. Chen, L. P. Dwoskin, P. R. Freeman, K. Kantak, J. Talbert, S. Tao, and G. Q. Zhang, "Predicting substance use disorder in ADHD patients using long-short term memory model," in 2018 IEEE International Conference on Healthcare Informatics Workshop (IEEE ICHI-W), (New York City, NY, USA), Jun., 2018.

- **S. Fouladvand**, E. R. Hankosky, H. Bush, J. Chen, L. P. Dwoskin, P. R. Freeman, D. W. Henderson, K. Kantak, J. Talbert, S. Tao, and G. Q. Zhang, "Predicting substance use disorder using long-term attention deficit hyperactivity disorder medication records in Truven," Mayo Clinic, Rochester, MN, 2019.

Professional Service

- Member of IEEE computational intelligence society–Task Force on Artificial Immune Systems (IEEE CIS ECTC) Program Committee Member of: 2019 IEEE Symposium on Immune Computation (IEEE IComputation).

- Reviewer for: AMIA 2021, Health Informatics Journal, Engineering Applications of Artificial Intelligence, Plant Physiology, Swarm and Evolutionary Computation, Neurocomputing, IEEE International Conference on Healthcare Informatics (ICHI), IEEE Symposium Series on Computational Intelligence (SSCI), IEEE Symposium on Immune Computation (IComputation), IEEE Congress on Evolutionary Computation (CEC2020).